Empirical Evidence Against Decision Augmentation Theory

Y. H. DOBYNS AND R. D. NELSON

Princeton Engineering Anomalies Research,
Princeton University, Princeton, New Jersey 08544

Abstract — A reference on the Decision Augmentation Theory (May *et al.,* 1995) includes a claim that certain data from the Princeton Engineering Anomalies Research program support the DAT model while refuting bitwise influence models at the 8.6σ level. We present here an analysis of the entire PEAR database published in Jahn *et al.* (1996), which shows that the database as a whole is consistent with bitwise influence, and rejects DAT at well over the 5σ level in a linear regression test. The subset of the data used by May *et al.* is examined in detail, and it is shown that the 8.6σ figure results from an erroneous analysis procedure. Both the data set available to May, and the overall database, are strongly inconsistent with DAT predictions while remaining consistent with a bitwise influence model.

Keywords: human/machine interactions — Decision Augmentation Theory

1. Introduction

Recently, a considerable amount of work has been devoted to a variety of consciousness-related anomalies. Much of that work has investigated what seems superficially to be two distinct classes of phenomena: the anomalous influence of human intention on physical systems (Jahn *et al.,* 1987; Jahn *et al.,* 1996; Nelson *et al.,* 1991; Radin and Nelson, 1989), and the anomalous acquisition of information by human consciousness (Bem and Honorton, 1994; Dunne *et al.,* 1989; Jahn *et al.,* 1987; May, 1996; Nelson *et al.,* 1996; Puthoff, 1996; Targ, 1996; Utts, 1996). A model called Decision Augmentation Theory (May *et al.,* 1995) constitutes a theoretical effort to unify these phenomena as two manifestations of the same underlying effect. In essence, the DAT model argues that the apparent influence experiments actually do not perturb the behavior of the experimental apparatus in any way. Instead, they reflect the ability of the human participant to choose propitious times to initiate data collection in light of anomalous knowledge about the future outcome of the experiment. In their paper, May *et al.* analyze a number of experiments and purport to show that these data are consistent with the quantitative predictions of DAT and inconsistent with influence models.

1.1 Distinguishing DAT and Influence

The distinction between any two theoretical models has empirical content only if they make different predictions for the outcomes of experiments. In most cases, we know too little about the underlying phenomena to make

quantitative predictions that would distinguish DAT from influence models within a single experiment. However, the models make different predictions for the way in which the effects scale between different experiments, which can be used to test the relative merits of the two hypotheses.

An influence model posits that the intervention of human consciousness changes the probability distribution of experimental outcomes. This is essentially a minimal description: the experimental evidence for an anomaly in these experiments consists in an observed distribution of probabilistic outcomes which differs from that predicted by theory. The influence "model," therefore, in its most general form is simply the presumption that this change in the observed probability is in fact a result of a change in the probability distribution describing the device's performance. DAT, on the other hand, is a biased sampling model which assumes that the mechanism of the anomaly is derived from the operator's choice of times to initiate data collection. For an experiment where the output data are discrete, and therefore in principle can be resolved as a composite of multiple binary decisions, the simplest influence model assumes a change in the probability of the elementary binary events; it thus predicts a constant statistical yield per bit (*i.e.,* the final Z score attained should scale as the square root of the number of bits processed), while the DAT model predicts a constant statistical yield per data-initiation event or "buttonpush." Comparison of experiments where different amounts of data are generated for a given buttonpush then should allow direct comparison of the two models.

It is possible, of course, to propound more complicated versions of each model. For example, psychological effects or fatigue might induce a length-dependence in the strength of influence which cause it to mimic DAT, or the availability of extra perceptual information might cause DAT to improve in efficiency on long datasets so as to mimic influence. While human psychology is a sufficiently rich field that some such elaboration might even be justified at some point, given our current level of understanding it seems appropriate to consider only the simplest form of each explanatory theory.

2. Nomenclature

The usual terminology of PEAR's reports differs from that customarily employed by May and other DAT analysts. To minimize linguistic confusion, we will define here the meaning of terms used in this paper. This may differ somewhat from PEAR's standard nomenclature, since most of our publications have not hitherto been concerned with DAT.

- **REG.** For historical reasons, the microelectronic noise generator used at PEAR is typically referred to as a Random Event Generator or REG, rather than the more common term Random Number Generator or RNG. REG may be considered in every way synonymous with RNG.

- **Trial.** This refers to the most elementary unit of data presented to the operator, or available for subsequent analysis. A trial is a single number generated by the REG. It may be the sum of 20, 200, or 2000 individual binary events.
- **Run.** This refers to the elementary unit of data archiving; in various experimental protocols, a run comprises 50, 100, or 1000 trials.
- **Series.** This refers to the minimal unit of a completed experiment. A series is defined by a set number of trials, which in various experiments ranges from 1000 to 5000 per intentional condition. (All of the experiments discussed here are tripolar, in that the operator generates data under two oppositely directed intentions, as well as a third "baseline" condition in which no attempt to influence the data is made.) An operator's formal database always consists of an integral number of completed series.
- **DAT Unit.** The elementary unit of the DAT model is the outcome of a single activation of the experimental apparatus by the operator; *i.e.,* the data generated by one push of a metaphorical "Start" button. (In the case of the REG this is, in fact, a literal button.) For the majority of the PEAR data this corresponds to a run, but there are some datasets which were generated using the "manual" mode of data acquisition. Although these preserve exactly the same run and series structure as the automatic operation, each individual trial requires a separate activation event by the operator. Thus, the manual data have a *de facto* runlength of 1 trial, despite the fact that they are collected in "runs" of 50, 100, or 1000 trials by the data acquisition software. Since no usual PEAR nomenclature corresponds to this quantity, which is essential for discussion of DAT predictions, we adopt the term "DAT unit" for the purpose.
- **Sequence Length.** To correspond to the usage of May *et al.,* we use the phrase "sequence length" to refer to the number of bits in a DAT unit.
- **Effect Size.** This refers to the magnitude of an observed anomalous effect, or the size of the deviation from expected behavior. The natural units for measuring effect size, in a particular model, are the data elements viewed as fundamental by that model. Thus, the natural unit for the influence model is the effect per bit, ε_b, while the natural unit of the DAT model is the effect per DAT unit, ε_D. The simplest form of each model predicts that the effect should be constant across experimental regimes when measured according to that model's natural units. It is worth noting that, since the authors of the DAT theory consider the DAT unit the fundamental element for analysis, the quantity here called ε_D is identical to that referred to by May *et al.* as the "average Z score." The ε_D notation was chosen for the current work since it points up the fact that ε_b and ε_D are each regarded as fundamental measures by different models.

3. Procedures for Testing the Models

To compare the two models against the available empirical data, we require a standard measure for experimental yield.

3.1 Measures of Effect

May *et al.* choose to perform all of their analyses in terms of Z^2 (except for a case where they estimate an effect size from Z before comparing it with a Z^2 value from another experiment; see below). The rationale for using Z^2 is un-clear. It might be because Z^2 scales with sequence length n when an influence model is assumed, which allows them to perform linear regression tests on Z^2 *vs.* n. However, linear regression can be applied just as readily to the slope of Z against $\sqrt{n}$. There is, moreover, no particular reason to adhere to models and representations amenable to linear regression. For *any* model that predicts some sensitivity of statistical yield to sequence length, one may extract the se-quence-length-dependent scaling factor, whatever it may be, and construct a χ^2 test for the consistency of the resulting population of measurements.

On the other hand, it may be that Z^2 is the parameter of choice for May *et al.* because it can accomodate the so-called "*psi*-missing" results described in the parapsychological literature. This, however, should be considered a strate-gy of last resort useful primarily for extreme cases (that is, cases where the rate of *psi*-missing is extremely high). If an operator is producing a mean shift of magnitude μ, whether by influencing the performance of the apparatus or by biasing the selection of active data, the direct evaluation of a mean shift Z is proportional to μ, while the change in Z^2 from its expectation is proportional to μ^2. Almost all such experiments have operated in a regime of extremely small μ; therefore $\mu^2 \ll \mu$. The Z test, on the other hand, is affected by the *psi*-missing rate: if a fraction x of the data contain an effect of magnitude μ but directed oppositely to intention, $Z \propto (1 - 2x)\mu$. For a Z^2 test to be more sensitive than a Z test, therefore, requires that (assuming the other relevant constants of proportionality are equal) $\mu > (1 - 2x)$; for the small values of μ that are typical, x must be very close to 50% for this condition to hold.

In their Table 1, May *et al.* identify the expected value of Z^2 in a DAT model to be $\mu_z^2 + \sigma_z^2$, where μ and σ are the parameters of the DAT biased selection process. Since the foregoing discussion considers only the mean shift, it might be argued that a Z^2 test can pick up changes in the variance of the distribution of run outcomes that are not predicted by a simple mean shift model. While this is technically correct, it is not necessarily helpful. There is nothing in the DAT model that constrains σ_z^2 to increase when a DAT process is employed. In fact, when one explores specific phenomenological models by which DAT might take place, one finds that most of them predict a diminution of σ_z^2. A previous publication (Dobyns, 1996) discusses the issue at length, and presents an argument that such variance reduction can be expected as a very general feature of DAT-like mechanisms (although the term is not used

there) given only the empirical fact that the observed effect sizes are quite small. For our current purposes, we will illustrate the point with an admittedly oversimplified and extreme example. Consider an operator with a precognitive faculty that informs him with perfect efficiency about the sign of a prospective run, but does not provide any other information. This person can, by employing this ability to avoid unwanted outcomes, generate nothing but positive output; if the natural distribution of the REG device is standard normal, this operator's output will be the $z > 0$ half of the standard normal distribution. For this biased selection process, $\mu = \sqrt{2/\pi} \approx .7979$, and $\sigma = \sqrt{1 - 2/\pi} \approx .6028$. A mean-shift Z test thus will ascertain quickly that the operator is producing a tremendous anomalous yield. Yet, despite the fact that we have defined his ability in terms of a DAT mechanism, $\mu_z^2 + \sigma_z^2 \equiv 1$, exactly the expected chance value, for this operator.

Since the PEAR experiments have been designed around a directed hypothesis specifying a mean shift in the direction of intent and show a strong anomalous yield in accord with that prediction, the remainder of the analyses in this paper will not use Z^2 as a measure. Instead, we will treat every data subset as a measurement of an underlying effect size parameter, and apply several statistical tests as discussed in the next section. An influence model predicts that the shift in elementary probabilities, or effect size per bit, should be constant. DAT, in contrast, predicts that the statistical yield per DAT unit should be constant. This is formally equivalent to May *et al.*'s formulation that the slope of Z^2 vs. n is predicted to be zero under DAT, since either statement can be derived from the other (presuming that the only variable under consideration is the change in sequence length). These effect sizes will be calculated directly from the Z scores and the number of data elements in each dataset.

The effect size in an experiment is, in fact, usually *defined* as the overall Z divided by the square root of the number of fundamental data units, $Z / \sqrt{N}$. If the average deviation from theory is the same in each data unit, the Z score will tend to grow as $\sqrt{N}$; therefore an effect size calculated according to this formula is independent of the amount of data collected. We need, in addition to this measure of the observed effect size, a measure of the accuracy with which the effect size is known. Now, a Z score is by construction normalized to the measurement error of the quantity under examination: to say that an observation has a Z of 5, for example, is equivalent to saying that the observation is 5σ away from the theoretical expectation. Therefore, if we divide Z by $\sqrt{N}$ to construct an effect size estimate, the measurement uncertainty of this estimate is necessarily 1 divided by $\sqrt{N}$. This gives us our explicit formulae for calculating effect sizes under the two models:

$$\varepsilon_b = \frac{Z}{\sqrt{N_b}} \pm \frac{1}{\sqrt{N_b}} \qquad (1)$$

$$\varepsilon_D = \frac{Z}{\sqrt{N_D}} \pm \frac{1}{\sqrt{N_D}}$$

where N_b is the number of bits in a given dataset, and N_D the number of DAT units. The sequence length n does not enter explicitly into this calculation, but

obviously $N_D = N_b/n$. As mentioned above, an influence model predicts that ε_b is a constant, regardless of n, while DAT predicts that ε_D is the constant. Each model predicts functional n-dependence for the other measure, as summarized in the following table:

TABLE 1
Functional Dependence of ε_b and ε_D on n Under the Competing Models

Model	ε_b	ε_D
Influence	Constant	$\varepsilon_D \propto \sqrt{n}$
DAT	$\varepsilon_b \propto 1/\sqrt{n}$	Constant

Note that the expressions in Table 1 follow from a fundamental relationship: since a DAT unit, by definition, comprises n bits, it is necessarily the case that regardless of model.

$$\varepsilon_D = \varepsilon_b \sqrt{n}, \tag{2}$$

3.2 Statistical Tools

The relations summarized in Table 1 allow several different approaches to evaluating the relative merits of the two models. First, we note that each model predicts that one measure should be a constant across experiments, while the other should show some n-dependence. We may therefore examine the constancy of ε_b and ε_D by any standard test for homogeneity, such as a χ^2 evaluation. For either model (momentarily omitting b and D subscripts for clarity), Eq. (1) allows us to calculate ε_i and σ_i for the i-th experiment under consideration. Across experiments we may calculate the mean effect size

$$\mu = \sum_i \frac{\varepsilon_i / \sigma_i^2}{1 / \sigma_i^2} \tag{3}$$

and thence $\chi^2 = \sum_i (\varepsilon_i - \mu)^2 / \sigma_i^2$. The number of degrees of freedom will be one less than the number of databases in question. Having calculated χ_D^2 and χ_b^2 for DAT and influence models respectively, we note from Table 1 that the influence model predicts χ_b^2 should follow the standard χ^2 distribution while χ_D^2 is inflated by the extra variation coming from the n-dependence of ε_D; the DAT model predicts the opposite relationship. Comparing the two χ^2 values will allow us to evaluate the relative merits of each model.

A slightly simpler evaluation is possible when only two data subsets need to be considered. We then have (again, for the moment, considering only one model) just two observations ε_1, ε_2 each with its known, normally distributed measurement uncertainty σ_1, σ_2. The variance of the difference $\varepsilon_1 - \varepsilon_2$ is the sum of the individual variances: $\sigma_1^2 + \sigma_2^2$. Under a hypothesis that predicts constant effect size, the difference has expectation 0, and the difference divided by its own standard deviation has unit variance: thus, by construction, the quantity $Z = (\varepsilon_1 - \varepsilon_2) / \sqrt{\sigma_1^2 + \sigma_2^2}$ is a standard normal deviate, or Z score, if the hypothesis in question is correct.

An alternative test of the hypotheses, much more similar to the approach used by May *et al.,* can be constructed by considering only one of the effect size measures at a time, and looking for the explicit functional dependence required by the two models. For example, looking at Table 1, we note that ε_D is predicted to be constant by DAT, and to vary as $\sqrt{n}$ by the influence model. We may therefore plot the various observed values of ε_D against $\sqrt{n}$, and do a linear regression to determine whether the resulting linear fit has a significant slope. If DAT is correct, we expect this line to have zero slope; if the influence model is correct, we expect the slope to differ strongly from zero. Again referring to Table 1 we can perform an exactly analogous calculation for ε_b against $1 / \sqrt{n}$; in this case, we expect a sloped line if DAT is correct, and a horizontal line if influence is correct.

With these tools in hand, we can evaluate the merits of the DAT model for the overall PEAR database discussed in Jahn *et al.,* (1996).

4. Evaluation of DAT in Overall PEAR Data

Jahn *et al.* (1996) summarizes all PEAR data obtained to its date in a subclass of human/machine experiments: namely, all *completed* experiments with *discrete* output. Ongoing studies using the same sources were not included, nor were experiments with analog output. Although these experiments cover a variety of sequence lengths, the data as presented in Jahn *et al.* cannot be used for a DAT analysis, because datasets that are quite distinct from the standpoint of DAT are presented as a single category. Table 2 represents these data with all subsets run at different DAT sequence lengths separated appropriately, to permit analytical comparisons between the DAT and influence models.

4.1 Explanation of Experiment Names

- **Benchmark.** This is the original REG source run in its standard mode of 200 bits per trial.
- **Remotes.** This is the same REG source run in a remote protocol where the operator was not physically present.
- **P-Random.** This was an early attempt to utilize a pseudorandom source; the source was found after the fact to contain a nondeterministic component.

TABLE 2
PEAR Database for DAT Analysis

Experiment	n	N_b	N_D	Z
		Random Experiments		
Benchmark	200	2.767×10^7	138350	2.333
Benchmark	10^4	1.1082×10^8	11082	2.536
Benchmark	2×10^4	6.68×10^7	3340	2.325
Benchmark	2×10^5	1.30×10^8	650	1.032
Remotes	2×10^5	1.832×10^8	916	2.214
P-Random	200	5×10^6	25000	0.220
P-Random	10^4	2.4×10^7	2400	2.978
P-Random	2×10^5	2.04×10^7	102	0.964
Co-Op	10^4	1.7×10^7	1700	1.106
Co-Op	2×10^4	3.6×10^6	180	0.164
Co-Op	2×10^5	1.56×10^7	78	1.257
REG2000	2000	4.86×10^7	24300	1.070
REG2000	10^5	2.04×10^8	2040	2.446
REG2000	2×10^5	7.2×10^7	360	0.775
REG20	20	3.2×10^5	16000	0.251
REG20	1000	1.2×10^6	1200	−1.063
REG20	2000	8×10^4	40	−0.636
REG20	2×10^4	4×10^4	2	−0.110
MC-Local	3.78×10^5	8.550×10^8	2262	3.891
MC-Remote	3.78×10^5	1.96×10^8	518	2.139
MC-Co-Op	3.78×10^5	9.07×10^7	240	−0.040
		Deterministic Experiments		
P-Determ	200	4×10^5	2000	−0.313
P-Determ	2×10^4	1.6×10^6	80	−2.229
P-Determ	2×10^5	7.2×10^6	36	−0.447
Algorithm	200	8.8×10^6	44000	−0.881
Algorithm	10^4	2.24×10^7	2240	-1.755
Algorithm	2×10^4	5.36×10^7	2680	0.662
Algorithm	2×10^5	7.36×10^7	368	-0.239
Alg-Remote	2×10^5	3.44×10^7	172	0.335
Alg-Co-Op	10^4	2×10^6	200	0.491
Alg-Co-Op	2×10^5	1.2×10^6	6	0.064

- **Co-Op.** All experiments labeled Co-Op involved two operators attempting to influence the device simultaneously. Those labeled only as Co-Op without further qualifiers were run on the benchmark REG (Dunne, 1991; Jahn *et al.,* 1987).
- **REG2000.** This was the same source as the Benchmark, but generating 2000 bits per trial rather than 200.
- **REG20.** This was the same source as the Benchmark, but generating 20 bits per trial rather than 200.
- **MC.** The Mechanical Cascade experiments used a device in which a single buttonpress by the operator released 9000 plastic balls sequentially into an array of pegs which distributed them among a set of counting bins. Analysis of the final cascade distribution indicates that each ball's

trajectory amounts on average to the information equivalent of 42 successive binary decisions, hence the sequence length calculated for MC of 3.78×10^5 bits per DAT unit. The MC experiments are subdivided into local, remote, and co-operator datasets in the same fashion as the REG experiments.

- **P-Determ.** This was the result of a second, successful attempt to construct a deterministic hardware noise source.
- **Algorithm.** These experiments used a software pseudorandom number generation technique. Like the REG and MC experiments, they were used in remote and co-operator protocols as well as in local experiments at various sequence lengths.

4.2 Reduced Form

We note that many of the lines in Table 2 are redundant. Although segregation of datasets for DAT required splitting many of the experiments reported in Jahn *et al.* into subsets, the commonality between many experimental protocols results in multiple datasets at each of several sequence lengths. Neither DAT nor the basic influence model have anything to say about differences between experiments at the same sequence length, so there is no reason not to combine all experiments run at the same sequence length into a single observation for purposes of analysis.

There is, however, one caveat that should be observed in making such a combination. Jahn *et al.* note that there is a considerable difference between the random-source and deterministic-source experiments, in that there is no aggregate evidence for an anomaly in the latter. Since the purpose of the current analysis is to identify the structure of an observed anomaly, combining datasets which have been seen not to contain an anomaly with datasets which do contain one will serve no purpose. Therefore, Table 3 reports the combined datasets by sequence length, but still segregates the random-source from the deterministic-source data.

All of the statistical tools discussed in the foregoing section need effect size estimates rather than Z scores for their application. Table 4 presents the effect sizes ε_b and ε_D, together with their respective error estimates σ_b and σ_D, as calculated from the data in Table 3. As noted in the header of the table, the effect sizes have been multiplied by constant factors for clarity of presentation.

4.3 Analysis Results

The χ^2 test for homogeneity of effect size is quite clear on the random-source data. The set of nine observations on ε_b produces $\chi^2 = 8.83$, with 8 degrees of freedom ($p = 0.357$). In contrast, the nine observations of ε_D produce $\chi^2 = 38.72$, again with 8 d.f. ($p = 5.53 \times 10^{-6}$).

TABLE 3
Combined Data by Sequence Length

Sequence Length (n)	N_b	N_D	Z
	Random Sources		
20	3.20×10^5	16000	0.251
200	3.27×10^7	163350	2.233
1000	1.20×10^6	1200	−1.063
2000	4.87×10^7	24340	1.043
10^4	1.52×10^8	15182	3.720
2×10^4	7.04×10^7	3522	2.299
10^5	2.04×10^8	2040	2.466
2×10^5	4.21×10^8	2106	2.808
3.78×10^5	1.14×10^9	3020	4.242
	Deterministic Sources		
200	9.20×10^6	46000	−0.926
10^4	2.44×10^7	2440	−1.541
2×10^4	5.52×10^7	2760	0.273
2×10^5	1.16×10^8	582	−0.113

TABLE 4
Influence and DAT Effect Sizes

n	$\varepsilon_b \times 10^4$	$\sigma_b \times 10^4$	$\varepsilon_D \times 10^3$	$\sigma_D \times 10^3$
	Random Sources			
20	4.438	17.678	1.985	7.906
200	3.908	1.750	5.526	2.474
1000	−9.700	9.129	−30.674	28.868
2000	1.495	1.433	6.684	6.410
10^4	3.019	0.812	30.194	8.116
2×10^4	2.739	1.191	38.736	16.850
10^5	1.727	0.700	54.598	22.140
2×10^5	1.368	0.487	61.194	21.791
3.78×10^5	1.256	0.296	77.192	18.197
	Deterministic Sources			
200	−3.054	3.297	−4.320	4.663
10^4	−3.121	2.024	−31.205	20.244
2×10^4	0.367	1.346	5.191	19.035
2×10^5	−0.104	0.927	−4.672	41.451

Figure 1 shows the results of plotting ε_D against $\sqrt{n}$. The regression line shown has a slope of $(1.379 \pm 0.241) \times 10^{-4}$.[1] The model predictions for this slope are 0 for DAT and 1.546×10^{-4} for influence: thus the observed slope

---

[1]The linear regression formula used here and in all such graphs is weighted, that is, it takes the observational uncertainty of each fitted point into account.
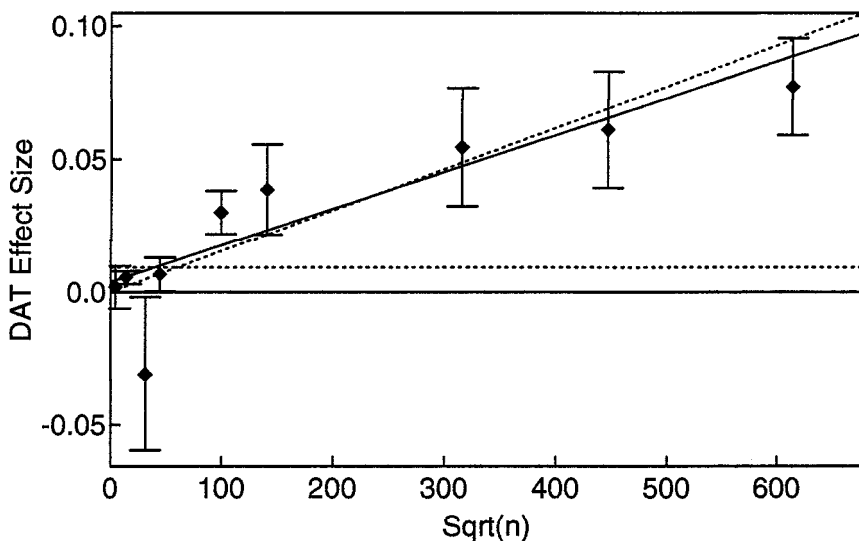
Fig. 1. ε_D plotted against $\sqrt{n}$. Data are from the random-source experiments, plotted as points with 1σ error bars. The horizontal dotted line is the prediction from the DAT model. The sloped dotted line is the prediction from the influence model. The solid sloped line is the linear regression fit to the observed data.

has Z = 5.723 against DAT (p = 1.04×10^{-8}, two-tailed), but only Z = -0.695 against the influence model.

Figure 2 shows the results for ε_b plotted against $1/\sqrt{n}$. In this presentation, the influence model predicts slope 0, while the DAT model predicts a slope of 9.224×10^{-3}. The linear fit has a slope of $(3.865 \pm 2.283) \times 10^{-3}$, leading to a Z score of 1.694 against influence (p = 0.090, two-tailed) and -2.347 against DAT (p = 0.019). We note that this result, while qualitatively in agreement with the other two, is quantitatively much weaker. A possible reason for this is that the regression line is poorly identified due to the structure of the data: comparison of Figures 1 and 2 will show that, although the data points in Figure 1 are well spaced across the abscissa range, the points of Figure 2 are much more heavily clustered, and the greatest contribution to the linear regression slope comes from isolated points with large error bars.

The data from deterministic sources, in contrast, show no ability to discriminate between the two models. This should not be surprising, since there is no anomaly to account for in the deterministic database; when the effect-size parameters for each model are indistinguishable from zero, the models are likewise indistinguishable from each other. The χ^2 test for homogeneity of effect, on 3 degrees of freedom, produces values of 2.87 for influence and 1.983 for DAT. The linear regression on ε_D against $\sqrt{n}$ produces Z = -0.139 against
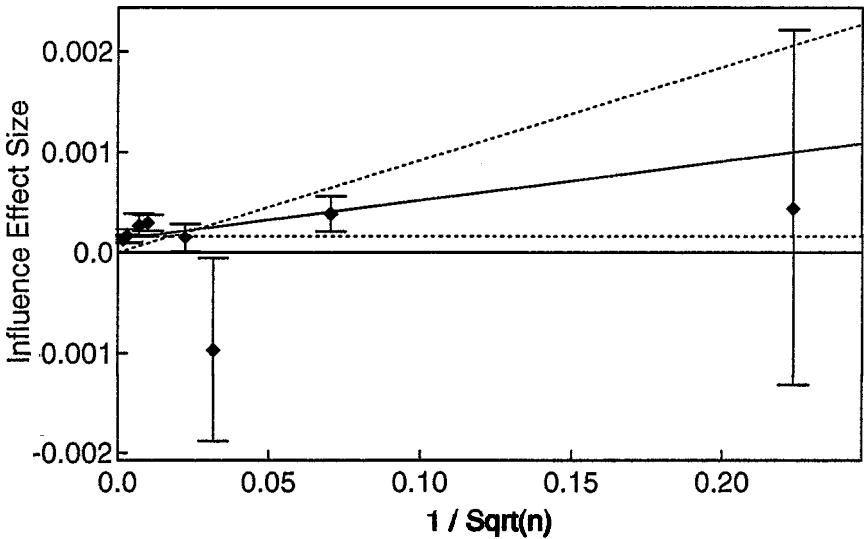
Fig. 2. ε_b plotted against $1/\sqrt{n}$, for the same data as Figure 1. Here the horizontal dotted line is the influence model prediction, while the sloped dotted line is the DAT prediction. Again, the sloped solid line is the linear regression fit.

DAT, $Z = .451$ against influence. Linear regression of ε_b against $1/\sqrt{n}$ returns $Z = .0662$ for DAT, $Z = -0.952$ for influence. Figure 3 illustrates ε_D against $\sqrt{n}$ for these data, to provide a graphic demonstration of the inconclusiveness of the model comparison when there is no anomalous effect.

It should be noted that the lack of effect in the deterministic experiments is itself evidence against the validity of the DAT hypothesis; since the experiments were run in the same protocols as REG experiments, with only the source differing, it is difficult to see why a DAT mechanism should discriminate between them. Leaving aside this issue, the strong effect in the random-source experiments allows us to reach a quite unambiguous conclusion that the effect shows the statistical signature of an influence process, and does not show the statistical patterns expected of a DAT phenomenon.

5. Possible Objections

Although the results in the previous section seem quite clear and striking, there are various objections to the analyses that might be and in some cases actually have been raised. This section will deal with such counterarguments.
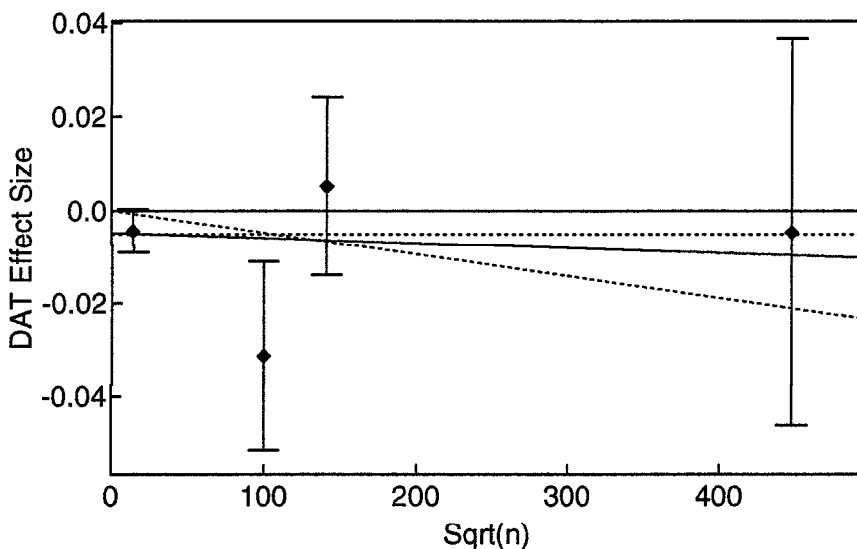
Fig. 3. ε_D plotted against $\sqrt{n}$ for the four deterministic data sets. Note that each dataset is close to the chance expectation line at 0, relative to its own error bars, and that both the fits and the regression line are likewise close to 0. This illustrates that both models account about equally well for a null effect.

5.1 Is It Legitimate to Include the MC Data?

E. C. May, a coauthor of the original DAT paper, has presented an argument against the inclusion of the mechanical cascade data in this analysis (May, 1997). The essence of this argument is that the mechanical cascade experiment represents a totally different physical system, and a very different psychological experience.

One can counter the physical argument quite well on theoretical grounds. The DAT model, in particular, is indifferent to the physical nature of the apparatus used in an experiment of this class. Since it asserts that the human operator *does not affect* the functioning of the device, but merely predicts favorable outcomes and starts data collection accordingly, the details of the physical device should be entirely irrelevant. If the inclusion of a different physical device causes a problem, it is that it predisposes the test to favor DAT; despite this predisposition, it is the DAT model that fails to fit the actual results.

The psychological difference between the running conditions of the MC and REG experiments may also be dismissed in principle. We may note that any difference in effect resulting from such psychological differences will be present for both model analyses (since the same data are being analyzed in two different ways), and therefore cannot predispose the test to favor one hypothesis or the other. Furthermore, the actual outcome of the test performed above requires, if this objection be valid, a rather startling numerical coincidence,

that the psychological effects altering the MC effect size should do so in *precisely* the correct way to make it fit an influence-model prediction derived from the REG experiments. In any case, if these considerations are not fully convincing, we may simply re-do the analysis on the REG data alone. Table 5 summarizes the results of such a recalculation.

As in the previous discussion, all p values quoted for Z scores are two-tailed. While there is some weakening of the result, not surprising with the loss of statistical resolution from the removal of so large a database, the conclusions are entirely unchanged. Figure 4 illustrates the effect of removing the MC dataset, in the ε_D *vs.* $\sqrt{n}$ representation.

5.2 Is the Segregation of Deterministic Data Legitimate?

The random data have an aggregate $\varepsilon_b = (1.55 \pm 0.22) \times 10^{-4}$, resulting in a Z score of 7.0 against the null hypothesis that there is no anomaly. In contrast, the deterministic data have a combined $\varepsilon_b = (-4.69 \pm 6.98) \times 10^{-5}$, indistin-

TABLE 5
Results of Model Evaluations with RMC Removed

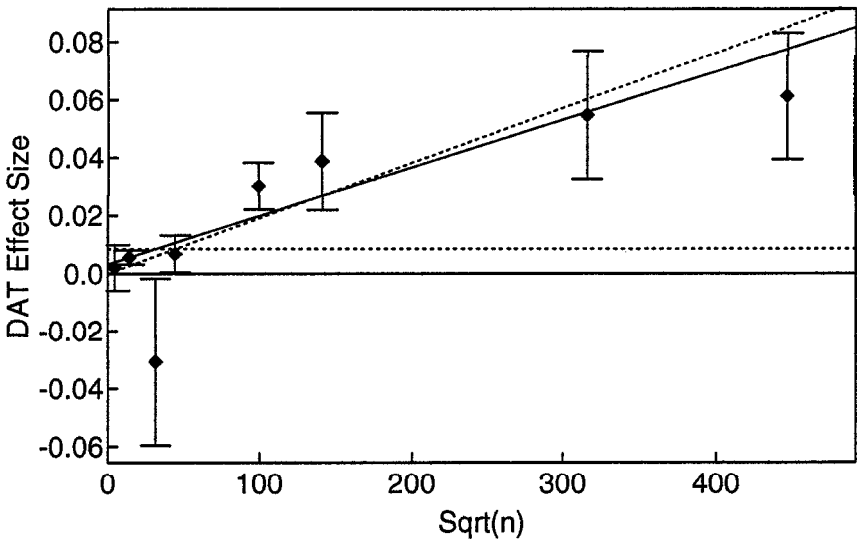| Test	DAT (p)	Influence (p)
Homogeneity (χ^2, 7 d. f.)	24.58 (9.0×10^{-4})	6.68 (0.46)
Regression Z, ε_D *vs.* $\sqrt{n}$	4.43 (9.5×10^{-6})	−0.636 (0.52)
Regression Z, ε_b *vs.* $1/\sqrt{n}$	−2.16 (0.03)	1.30 (0.19)

Fig. 4. ε_b plotted against $\sqrt{n}$ for REG data only. Compare with Figure 1, which includes the MC data as well, to confirm how little this changes the interpretation of the data.

guishable from zero. The Z score between these two values is 2.75. A corresponding calculation on ε_D arrives at 2.91 for the difference. It might seem that these figures bear out a legitimate difference between the databases, but as a precaution we can also re-do the analysis with the deterministic datasets added into the total databases at the appropriate sequence lengths. The results of doing so are shown graphically in Figure 5 (in the ε_D vs. $\sqrt{n}$ plot), and summarized in Table 6.

We see that, as with removing the MC data, adding the deterministic datasets makes no appreciable change to the conclusions: DAT is strongly refuted, while influence is consistent with the results.
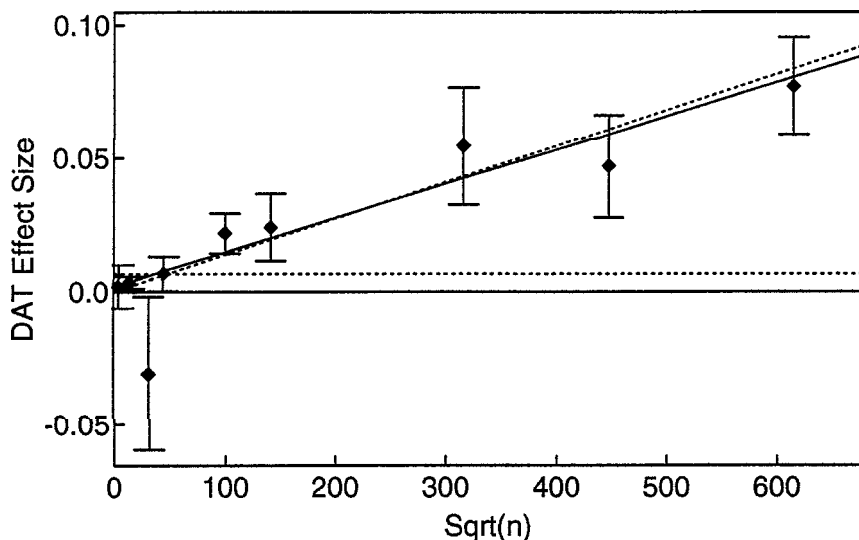
Fig. 5. ε_D plotted against $\sqrt{n}$ for all data, deterministic combined with random. Compare with Figure 1 and Figure 4.

TABLE 6
Model Evaluations on Combined Random and Deterministic Data

Test	DAT (p) (p)	Influence
Homogeneity $(\chi^2,$ 7 d. f.)	$34.24\ (3.7 \times 10^{-5})$	$4.15\ (0.84)$
Regression Z ε_D vs. $\sqrt{n}$	$5.56\ (2.7 \times 10^{-8})$	$-0.369\ (0.71)$
Regression Z ε_b vs. $1/\sqrt{n}$	$-2.30\ (0.02)$	$0.890\ (0.37)$

5.3 Is the χ^2 Test Reliable?

In the same presentation cited above (May, 1997), it was pointed out that the χ^2 test is subject to possible confounds from extraneous sources of variance. If differences in the operator population or psychological response between experiments produced model-independent changes in the effect, the net effect would be an increase in the variance between observations, leading to an increased value of the χ^2.

Several responses to this criticism are possible. First, the fact that the χ^2 test consistently returns reasonable values for the influence model is *prima facie* evidence that the observations do not in fact contain confounds of the sort discussed by May, at least not to an extent sufficient to be detectable in this test. Second, concerns with the χ^2 test are to some extent moot, since the conclusions from the χ^2 test are consistently confirmed by linear regression tests of the sort recommended by May *et al.* (although we are, in their notation, regressing on Z against $\sqrt{n}$ rather than Z^2 against n). Finally, we may observe that, because the same data are being analyzed in two different ways, any spurious variance contributions would inflate both χ^2 values equally, and would not change the relative preference of the two models.

This last point can be verified directly by employing a Monte Carlo procedure. We took, as a starting point, the actual amount of data present in each of the nine random-source databases in Table 3. Synthetic data were then generated under each hypothesis using the average effect size observed for that hypothesis. The method of synthesis was to calculate the expected Z score for the experiment, given the average effect size, and to add a normal random deviate to represent the observation error. This was done with both $\sigma = 1$ random deviates, to model the case where spurious variance contributions are absent or inconsequential, and with $\sigma = 2$ random variates, to model the most extreme case of spurious contribution discussed by May (1997). The process was then iterated 100 times, for each model and each variance. The average χ^2 values, each with 8 d.f., for each model analysis on each of these four synthetic datasets are summarized in Table 7.

As we expected *a priori,* we find that the extra contribution from the variance inflation is present equally in both model evaluations. At worst it inflates both model evaluations to significant rejection, while retaining the feature that

TABLE 7
Monte Carlo Verification of χ^2 Test

Model Used for Data Generation	σ	Influence χ^2	DAT χ^2
Influence	1	8.58	48.73
Influence	2	31.71	70.71
DAT	1	22.30	8.41
DAT	2	49.20	34.08

the incorrect model is rejected much more strongly than the correct one. We recall that the actual data show χ^2 values close to expectation under the bitwise influence calculation, and very large values under the DAT calculation. Table 7 demonstrates that this condition only arises in the case when the synthetic data were generated under an influence model and the variance was not inflated: thus, the observed characteristics of the data refute both the DAT model and May's objection.

5.4 Bayesian Evaluation

Analyses of this sort are frequently criticized for their reliance on p-values. Bayesian analysts, in particular, object (correctly) that a p-value has no direct interpretation as a posterior probability on the truth of some hypothesis. Rather than reprise the ongoing argument between Bayesian and frequentist approaches, we find it easier simply to report the empirical odds adjustment between the two hypotheses under consideration.

As a brief reminder, this form of hypothesis test proceeds from the assumption that the prior probabilities on two competing hypotheses have been stated in the form of odds, a ratio p_1/p_2 expressing (for example) the relative prior probability of the hypotheses under consideration. The advantages of this representation are twofold: First, the empirical factor obtained from the data, expressing the relative support given by the data to one hypothesis as compared to the other, can be expressed as a fairly simple ratio of the probabilities of the observed data under each of the hypotheses. Second, because the empirical factor refers to the ratio of probabilities, it can be applied by any analyst regardless of the specific choice of priors. Whether an analyst considers DAT to be likely or unlikely *a priori,* the same factor can be used to arrive at the posterior estimate of the odds. Table 8 presents odds adjustments in favor of influence (that is, the probability of an influence explanation is taken as the numerator of the odds ratio), for the three tests and three datasets considered above.

We note that both the homogeneity test and the linear regression test on ε_D enormously increase the posterior odds in favor of an influence explanation. The discussion of the linear regression on ε_b in Section 4 pointed out that this test will have poor statistical resolution on the current database, due to the uneven placement of the observational data in $1/\sqrt{n}$ space; the relatively small

TABLE 8
Odds Adjustment Favoring Influence

Test	Random sources	REG Only	All Sources
Homogeneity χ^2	3.7×10^4	300	6100
Linear ε_D on $\sqrt{n}$	1.0×10^7	1.5×10^4	4.7×10^6
Linear ε_b on $1/\sqrt{n}$	3.7	4.4	9.5

factors appearing in the last line of Table 8 are the result of this low resolving power, rather than any ambiguity in the data or disagreement between the tests.

6. Integration with Previous Results

We note that the analysis described above is in substantial disagreement with the analysis reported by May *et al.,* where an examination of a subset of PEAR's REG data produced a Z of 8.6 against "micro-AP" while remaining consistent with DAT. This disagreement can be traced to two errors in the earlier analysis. One error involves the use of an ill-conceived statistical test; the other is their ignorance of a confounding factor present in the data.

6.1 Analysis Error

May *et al.,* in their 1995 paper, analyze a subset of PEAR REG data generated by a single operator. For convenience, we are reproducing their relevant paragraph verbatim, noting that n is the symbol used for sequence length, and AP the authors' acronym (Anomalous Perturbation) for what herein is called an influence model.

At $n = 200$, 5918 trials yielded $Z = 0.044 \pm 1.030$ and $Z^2 = 1.063 \pm 0.019$. We compute a proposed AP effect size $Z/200^{1/2} = 3.10 \times 10^{-3}$. With this effect size, we computed what would be expected under the micro-AP model at $n = 100,000$. Using the theoretical expressions in Table 1, we computed $Z^2 = 1.961 \pm 0.099$. The 1σ error is derived from the theoretical variance divided by the actual number of trials (597) at $n = 100,000$. The *observed* values were $Z = 0.100 \pm 0.997$ and $Z^2 = 1.002 \pm 0.050$. A t-test between the observed and expected values of Z^2 gives $t = 8.643$, $df = 1192$. Considering this t as equivalent to a Z, the data at $n = 100,000$ fails to meet what would be expected under the influence model by 8.6σ. Suppose, however, that the effect size observed at $n = 100,000$ (3.18×10^{-4}) better represents the AP effect size. We computed the predicted value of $Z^2 = 1.00002 \pm 0.018$ for $n = 200$. Using a t-test for the difference between the observed value and this predicted one gives $t = 2.398$, $df = 11,834$. The micro-AP model fails in this direction by more than 2.3σ. DAT predicts that Z^2 would be statistically equivalent at the two sequence lengths, and we find that to be the case ($t = 1.14$, $df = 6513$, $p = 0.127$).

The relevant formula "from Table 1" referred to is that for the expectation value of Z^2; in the notation of this paper, it would be $\langle Z^2 \rangle = 1 + \varepsilon_b^2 n$.

Let us examine the analysis described above in detail. An effect size is first calculated from the observed Z at $n = 200$, and then used to calculate a predicted value for Z^2 at $n = 100,000$. The error estimate on this prediction "is derived from the theoretical variance divided by the actual number of trials (597) at $n = 100,000$."

It is not clear what relevance the number of trials at $n = 100,000$ has to the accuracy with which we know a number computed from observations at $n = 200$. In point of fact, the effect size computed from the $n = 200$ data is an

empirical value known with limited precision: it should rather have been written as $(3.10 \pm 0.95) \times 10^{-3}$. This uncertainty in the value of ε_b propagates into any quantity subsequently calculated from ε_b.

This leads to a second flaw, namely the operation of comparing an effect size estimate obtained from Z at one sequence length with an observation of Z^2 at the other sequence length. The observed value of Z has a normally distributed error by construction, and the observed value of Z^2, although intrinsically χ^2 distributed, has so many degrees of freedom that it may be considered effectively normal. The calculated value of ε_b from the Z observation inherits the normally distributed uncertainty of its source. However, to make the comparison with Z^2, we must compute ε_b^2; and *this* quantity, the square of a normally distributed variable, has the grossly non-normal distribution of a χ^2 with only one degree of freedom. The 1σ range of the empirical value for ε_b at $n = 200$ (*i.e.,* from 2.15×10^{-3} to 4.05×10^{-3}) translates to a range of predicted Z^2 values at $n = 100,000$ from 1.452 to 2.640. In contrast, the value 1.961 ± 0.099 presented in the quoted paragraph would produce a 1σ range from 1.862 to 2.060. Thus, the procedure used by May *et al.* grossly underestimates the statistical uncertainty of one of their parameters and therefore hugely inflates the significance of their results.

This need not be left at the level of a theoretical objection, but can be demonstrated directly by applying the procedure of May *et al.* to synthesized data in a Monte Carlo procedure. For convenience, this was done using round numbers of observations closely similar to those reported above (6000 instead of 5918 observations at $n = 200$, 600 instead of 597 at $n = 100,000$). First, a set of 6000 trials at $n = 200$ was constructed, using exactly the mean shift quoted in the published analysis. Then, two sets of 600 trials for $n = 100,000$ were constructed, one using the DAT model prediction and the other, the influence model prediction. Thus, by comparing the single shared set of $n = 200$ data with the two sets of $n = 100,000$ data, we can evaluate the performance of any test procedure in circumstances where either DAT or influence is known to be true by construction.

Three tests were used: the "forward" procedure of May *et al.,* where $n = 200$ data are used to obtain an effect size for extrapolation to $n = 100,000$; the "reverse" procedure described later in the quoted paragraph, where the $n = 100,000$ observation is used to predict an effect size at $n = 200$; and finally, a Z score between the two observed values of ε_b, as described in Section 3 above. One hundred iterations of the process were performed, to evaluate the mean and standard deviation of the test statistic in these circumstances. While the score emerging from either procedure used by May *et al.* is conceptually a T score, the number of degrees of freedom is so large that all three of these tests can effectively be treated as Z scores. Since they are tests of the validity of a bitwise-effect hypothesis, they should display mean zero and standard deviation 1 on the constructed influence data, and dis-

TABLE 9
Monte Carlo Evaluation of Test Statistic

Test	Data Constructed By	Mean	Std. Dev.
May *et al.*, forward	Influence	−0.830	7.712
May *et al.*, forward	DAT	13.124	6.480
May *et al.*, reverse	Influence	0.014	1.030
May *et al.*, reverse	DAT	−0.091	1.030
Z on ε_b	Influence	−0.152	1.002
Z on ε_b	DAT	3.051	0.976

play a reasonably consistent measure of the disparity on the constructed DAT data. Table 9 summarizes the actual performance of these three measures.

Thus, the theoretical objections to the test of May *et al.* are abundantly justified in practice. Their alleged 8.6σ result derives in fact not from a distribution with standard deviation 1, as would be required for the interpretation May *et al.* put on it, but instead from a distribution with standard deviation ≈ 7.7. Its real significance, then, is roughly that of a Z score of 1.1; scarcely powerful evidence. In datasets constructed according to the DAT model, the average value of this test statistic is about 13, considerably larger than the value obtained on the actual data.

The reverse test of May *et al.* also fails in that it retains a $\mu = 0$, $\sigma = 1$ distribution under either model, displaying no ability to discriminate between the two. The modest result (2.3) reported by May *et al.* for this measure must, insofar as it represents a real phenomenon, be the consequence of attributes in the data unrelated to their status relative to the DAT or influence models, since the test is insensitive to this distinction.

Only the Z comparison between the observed values of ε_b displays the proper behavior for a test statistic, following a standard normal distribution when the influence model is true and showing a consistently shifted value when the DAT model is true. From the figures used by May *et al.* for the values of ε_b observed at the two sequence lengths, when the observational uncertainty is properly taken into account, this test produces a result $Z = 2.91$. It might therefore seem that the above dataset does in fact support DAT, though not nearly as strongly as May *et al.* claim; but there remains one more element to take into account.

6.2 The Data Are Confounded

The PEAR data originally transmitted to May constituted subsets of the database of one particular operator, provided in response to a request for data from a single successful individual. Specifically, they included the operator's full accumulation through the date they were provided (1987), in the original REG experiment and in a variant experiment run at 2000 samples per trial rather than 200. This provided four values of n: 200, 2000, 10,000, and

100,000, depending on whether the DAT unit comprised one or 50 trials. Data from several other variant experiments, such as the remote protocol, were not included in the transmission.

One of the most noteworthy features of this individual's database is a tremendous change in anomalous yield that appeared concurrent with a change in the general experimental approach. In the earliest experiments, before the scale of the databases that would eventually be produced was appreciated, operators were allowed to choose secondary parameters (run length, sampling rate, instructed or volitional intention assignment) *ad libitum* from session to session within a given series.[2] It was not until later that the decision was made, for ease and consistency of analysis, that all such secondary parameters would be held constant throughout an entire series. The old data, from multi-parameter series, were flagged in the database management system as the "X protocol," as a signal that special handling was required for any analysis of secondary parameters involving these series.

Whether it was the added restriction of consistency of secondary parameters, or some other factor, we do not know, but as this operator's database continued to grow it became evident that the X protocol data had a much larger statistical yield than the subsequent data (Nelson *et al.,* 1991). (It should be noted that the protocol transition was purely a matter of convenience in data handling, rather than a tightening of controls, so the possibility of a spurious earlier result eliminated by increase in rigor need not be considered.) As it happens, *all* of this operator's sequence length 200 data were run under the multiparameter X protocol, while *none* of the $n = 100,000$ data came from that protocol. This immediately suggests a confounding effect in the existing comparison of $n = 200$ with $n = 100,000$ data. Since the X protocol data are known to have a much larger (bitwise) effect size than the subsequent data, some or all of the difference between the data at the two sequence lengths might be due to the difference in protocols. On the other hand, DAT might actually be the explanation for the interprotocol difference: since DAT predicts a much larger effect per bit in the $n = 200$ data than for longer sequences, the difference in performance between protocols might simply be a result of the normal operation of DAT in light of the fact that all the $n = 200$ data were generated under the X protocol.

The only way to determine whether the X protocol difference confounds the DAT analysis, or whether conversely the effect of DAT causes the X protocol difference, is to separate out the datasets by both sequence length and protocol. Ideally, one would use ANOVA or general regression analysis to isolate the respective explanatory power of sequence length and protocol difference. However, there would be several empty cells in such a breakdown, making

---

[2]In this phase of the program the series as defined required several hours of effort by the operator, and a series was routinely completed over several experimental sessions of length convenient to the operator, the only constraint being that roughly equal amounts of data in each intention should be generated in any given session.

TABLE 10
Data by Sequence Length and Protocol

Protocol	Seq. Length	ZScore	N_b	N_D
X	200	3.28255	1.25×10^6	6250*
X	10,000	4.72477	3.78×10^6	378
Standard	2000	0.72744	3.06×10^7	15300
Standard	10,000	1.86267	1.70×10^7	1700
Standard	100,000	2.44773	6.00×10^7	600*

ANOVA impossible and even regression analysis problematic. There are, on the other hand, enough data that we can quite readily compare datasets run in the same protocol at different sequence lengths, and at the same sequence length in different protocols. Table 10 presents the database transmitted to May, broken down by sequence length and protocol, along with the overall Z scores, number of bits, and number of DAT units.

The two datasets flagged with asterisks are those reported by May *et al.,* and it should be noted that the amount of data presented here is not precisely what was reported in that reference. Somehow, 332 trials at sequence length 200 and 3 runs at sequence length 100,000 have been omitted from their presentation. Since not enough data to comprise a complete series are missing in either case, the source of the loss is obscure; we cannot identify whether it occurred in our preparation or their handling of the transmitted data. In any event, the missing data do not appreciably change the yield calculations for those two datasets (May *et al*. report 5918 trials, with Z = 3.37, to our 6250 with Z = 3.28, in the first case, and 597 *vs*. 600 DAT units in the second, with a Z of 2.45 in both sets).

Table 11 uses the data from Table 10 to calculate ε_b and ε_D in the manner discussed in preceding sections.

The second and fourth lines of Table 11 show the two datasets run under different protocols at sequence length n = 10,000. Under either model, comparison of the two effect size figures yields a Z of 3.479, confirming the presence of a model-independent difference between the two protocols. This establishes that in order to find the functional dependence on sequence length, we must

TABLE 11
Effect Size by Sequence Length and Protocol

Protocol	Seq. Length	ε_b	ε_D
X	200	$(2.9360 \pm 0.8944) \times 10^{-3}$	$(4.1521 \pm 1.2649) \times 10^{-2}$
X	10,000	$(2.4302 \pm 0.5143) \times 10^{-3}$	0.24302 ± 0.05143
Standard	2000	$(1.315 \pm 1.808) \times 10^{-4}$	$(5.8810 \pm 8.0845) \times 10^{-3}$
Standard	10,000	$(4.518 \pm 2.425) \times 10^{-4}$	$(4.5176 \pm 2.4254) \times 10^{-2}$
Standard	100,000	$(3.160 \pm 1.291) \times 10^{-4}$	$(9.9928 \pm 4.0825) \times 10^{-2}$

compare data generated in the same protocol, rather than between different protocols.

Figure 6 illustrates the data shown in Table 11. In Figure 6a, only the two datasets presented by May *et al.* are shown. Figure 6b shows the full dataset transmitted to May in 1987. Figure 6c, corresponding to the ε_D column of Table 11, displays the five datasets that result when the four sequence lengths and the two disparate protocols are disentangled. Note the uniform progression to larger effect sizes shown by each protocol independently; since this is a plot of ε_D, we expect the effects to be constant if DAT is true, and increasing with n if influence is true.

Figure 7 shows the results of linear fits to the two protocols individually, with 7a displaying the X protocol data and 7b the standard protocol data. The quantitative results of this analysis are summarized in Table 12. As in previous cases, the "Linear Z" is the Z score obtained by comparing the slope of the linear regression fit with the predicted slopes from the two models, in each representation.

We see that without exception each test shows a preference for the influence model as a better fit to the data, and in most cases the preference is also highly significant.

We have noted that this dataset comprises two distinct experimental databases with different effect sizes. Nevertheless, it is possible to obtain a single, quantitative bottom line for the comparison of DAT with influence in this dataset by using the χ^2 test of homogeneity. Independent χ^2 values obey a summation rule; the sum of a χ^2 with N degrees of freedom and a χ^2 with M d.f. is itself a χ^2 with $N + M$ degrees of freedom. Summing the values from the two tests displayed in Table 12, we find χ^2 values with 3 d.f.: 1.491 for the influence model, 21.541 for DAT. This latter result has $p = 8.1 \times 10^{-5}$.

6.3 Integration with Primary Database

In the previous sections we found that the overall database strongly refutes the DAT hypothesis. In the current section, however, we find a subset where a mixing of two experimental conditions with identifiably distinct effect sizes produces a spurious appearance of support for DAT when only sequence length is taken into account (Section 6.1). These data are part of the overall database analyzed in Section 4, where only sequence length was taken into account. It follows that the analysis in Section IV is inappropriately lenient to the DAT hypothesis, since at least part of the data used there contain a DAT-mimicking confound.

A more accurate estimate of the extent to which DAT is inconsistent with the observed data can be achieved by re-doing the analysis on those data remaining after the dataset transmitted to May is removed. The χ^2 homogeneity measure on this dataset then constitutes a completely independent test of the hypothesis, and, as per the summation rule described above, can be combined with the result of the same test on the dataset transmitted to May. Removing
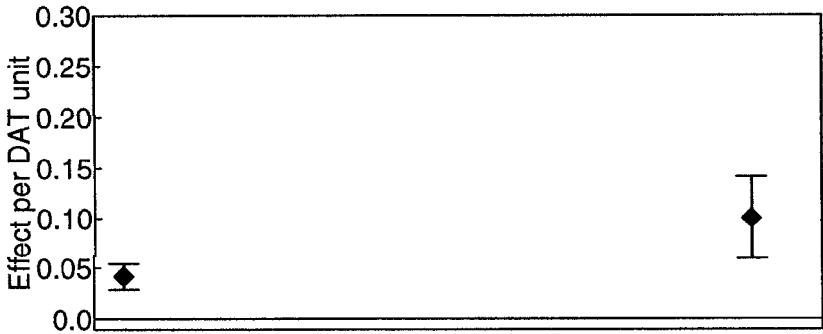
(6a)
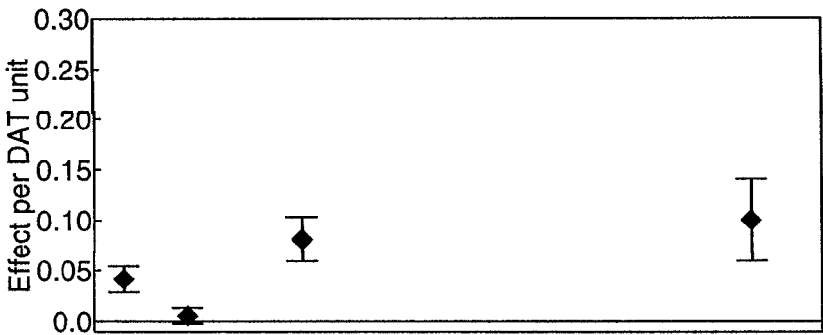
(6b)
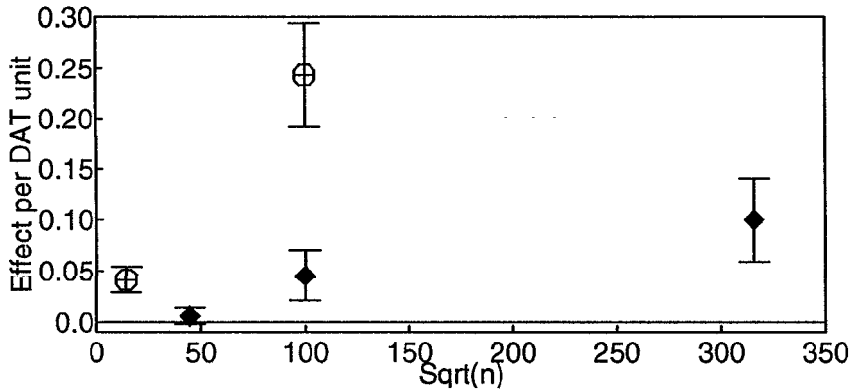
(6c)

Fig. 6.   Dataset provided to E. C. May. (a) Data as presented in May *et al.,* 1995, using only two
sequence lengths. (b) Full dataset transmitted to May. (c) Dataset with the two differently
yielding protocols separated, and marked with distinct icons. Note that the point at
$\sqrt{n} = 100$ has split into two distinct observations. Circles show X protocol data; filled dia-
monds show standard protocol data. Note progressive trend to larger effects with increas-
ing n, independently visible in both protocols.
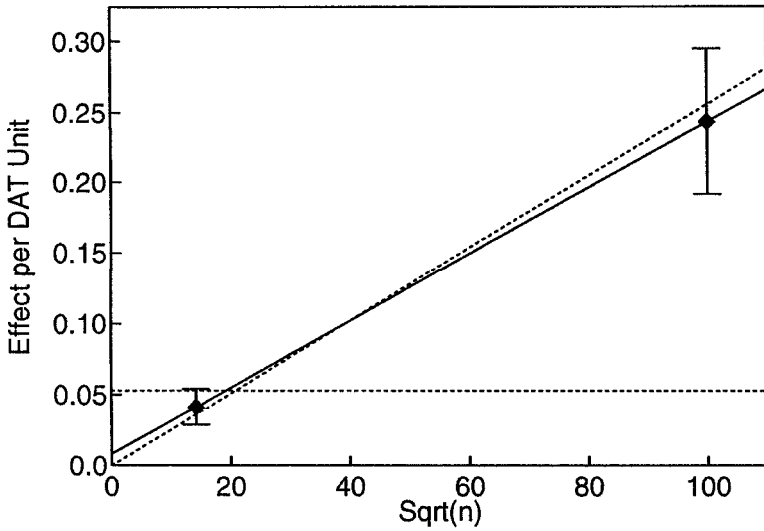
(7a)
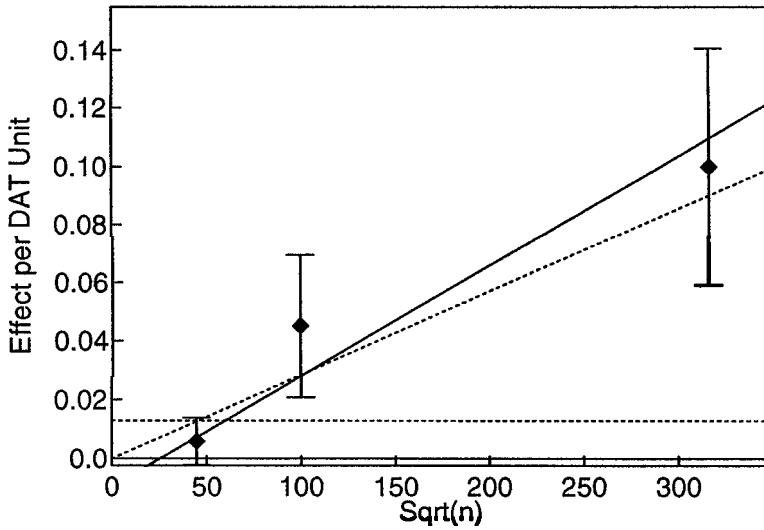
(7b)

Fig. 7. Results of linear fit to observations, in each protocol separately. In each graph, the horizontal dotted line is the DAT prediction, the sloped dotted line is the influence model prediction, and the sloped solid line is the linear fit to the data. (a) Fit to X protocol. Although with only two observations the linear fit automatically passes through both plotted points, the known error estimates on each point permit determination of the uncertainty of the line's slope. (b) Fit to standard protocol.

TABLE 12
Statistical Evaluations on Data Transmitted to May

Protocol	Test Used	Influence (p)	DAT (p)
X	Homogeneity χ^2 (1 d.f.)	0.240 (0.624)	14.472 (0.00014)
X	Linear Z, ε_D on $\sqrt{n}$	−0.339 (0.734)	3.804 (0.00014)
X	Linear Z, ε_b on $1/\sqrt{n}$	0.490 (0.624)	−2.629 (0.0086)
Standard	Homogeneity χ^2 (2 d.f.)	1.251 (0.535)	7.069 (0.029)
Standard	Linear Z, ε_D on $\sqrt{n}$	0.621 (0.534)	2.547 (0.011)
Standard	Linear Z, ε_b on $1/\sqrt{n}$	−0.819 (0.412)	−1.932 (0.053)

TABLE 13
Modifications to Table 3

Sequence Length (n)	N_b	N_D	Z
200	3.14×10^7	157100	1.623
2000	1.81×10^7	9040	0.765
10^4	1.31×10^8	13104	2.531
10^5	1.44×10^8	1440	1.355

the data analyzed in the previous subsection from the overall accumulation of random data changes four lines in Table 3 **(page 10)**, those dealing with sequence lengths 200, 2000, 10^4, and 10^5. The modified lines are shown in Table 13.

When effect sizes are calculated for these modified data values, and the overall dataset tested for homogeneity of effect size, the results are once again unambiguous. Under the influence-model assumption that ε_b is constant, we obtain $\chi^2 = 4.709$ on 8 degrees of freedom, $p = 0.788$. Under the DAT assumption that ε_D is constant, $\chi^2 = 32.180$, $p = 8.65 \times 10^{-5}$.

When these results are added to the consistency measures found in the previous subsection, the overall homogeneity rating for the influence hypothesis is $\chi^2 = 6.200$, 11 d.f., $p = 0.860$. In other words, the data are entirely consistent with the hypothesis of a bitwise alteration of probability. The same measure for DAT returns a value of 53.721, $p = 1.327 \times 10^{-7}$.

Conclusions

The overall analysis of PEAR's human-machine database shows a strong refutation of the DAT model. The anomalous effects observed are consistent with an actual change in the operation of the experimental devices, taking place at the level of elementary binary decisions; they are inconsistent with the process of biased sampling required by the DAT model. A previous analysis where a subset of PEAR data was claimed to support DAT has been shown to result from a combination of invalid analytical technique and confounded data; the subset in question actually refutes DAT at a very high level of statistical significance. Alternative analyses, including the use of linear regression to

identify functional dependence as recommended by May *et al.,* have been found to confirm the results of simple tests for homogeneity of effect size as predicted by the two hypotheses. In short, the anomalous human-machine interactions observed at PEAR are not the results of a DAT process.

Acknowledgments

References

Bem, D. J., & Honorton, C. (1994). Does *psi* exist? Replicable evidence for an anomalous process of information transfer. *Psychological Bulletin,* 155, 1, 4.

Dobyns, Y. H. (1996). Selection versus influence revisited: New method and conclusions. *Journal of Scientific Exploration,* 10, 2, 253.

Dunne, B. J. (1991). Co-Operator Experiments with an REG Device. PEAR Technical Note 91005.

Dunne, B. J., Dobyns, Y. H., and Intner, S. M. (1989). Precognitive Remote Perception III: Complete Binary Data Base with Analytical Refinements. PEAR Technical Note 89002.

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

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

Jahn, R. G., Dunne, B. J., Nelson, R. D., Dobyns, Y. H., and Bradish, G. J. (1996). Correlations of random binary sequences with pre-stated operator intentions. PEAR Technical Report 96003. Also available in *Journal of Scientific Exploration,* 1997, 11, 3, 345.

May, E. C., Utts, J. M., and Spottiswoode, S. J. P (1995). Decision augmentation theory: applications to the random number generator database. *Journal of Scientific Exploration,* 9, 4, 453.

May, E. C. (1996). The American Institutes for Research Review of the Department of Defense's STAR GATE Program: A commentary. *Journal of Scientific Exploration,* 10, 1, 89.

May, E. C. (1997). Presentation to the 16th Annual Meeting of the SSE, 7 June 1997.

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. PEAR Technical Note 91004.

Nelson, R. D., Dunne, B. J., Dobyns, Y. H., and Jahn, R. G. (1996). Precognitive remote perception: replication of remote viewing. *Journal of Scientific Exploration,* 10, 1, 109.

Radin, D. I., & Nelson, R. D. (1989). Evidence for consciousness-related anomalies in random physical systems. *Foundations of Physics,* 19, 12, 1499.

Puthoff, H. E. (1996). CIA-initiated remote viewing program at Stanford Research Institute. *Journal of Scientific Exploration,* 10, 1, 63.

Targ, R. (1996). Remote viewing at Stanford Research Institute in the 1970's: A memoir. *Journal of Scientific Exploration,* 10, 1, 77.

Utts, J. (1996). Evaluation of a program on anomalous mental phenomena. *Journal of Scientific Exploration,* 10, 1, 3.