

Statistical Consequences of Data Selection

Y. H. DOBYNS

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

Abstract—Data selection can result from unconscious biases or preferences on the part of experimenters, or from deliberate efforts to skew the apparent character of an experimental database. In either case the same formalism can be applied to compute the statistical signature of the selection process. Since the result of a suitably chosen selection process can be arbitrarily close to any desired distribution of experimental outcomes, it also is necessary to take into account the fraction of data that would have had to be discarded. When the selection formalism is applied to the Princeton Engineering Anomalies Research (PEAR) benchmark random event generator (REG) database, it is found that no selection model examined is consistent with the data. An unusual subset of these data, produced by a single operator, which has in the past been the target of suspicion, is likewise inconsistent with any selection hypothesis, even under a worst-case scenario of deliberate fraud.

Keywords: Statistical Methods — Meta-Analysis — REG — Human-Machine Interaction — Consciousness-Related Anomalies — PEAR

Introduction

"Data selection" is a common term for the biased selective reporting of data in scientific research. It is frequently invoked as a dismissive explanation for peculiar or anomalous results. While it usually implies a deliberate attempt to deceive, data selection also can result from unconscious biases of experimenters or inadequate controls in the recording and reporting of data (Gould, 1996). The work presented here explores the possibility that the anomalous effects seen in the benchmark random event generator (REG) database generated at the Princeton Engineering Anomalies Research (PEAR) laboratory (Jahn *et al.*, 1997) might be due to such a selection process. None of the selection models examined are consistent with the data, and some general properties of selection models suggest that no possible model of this class can be constructed to be consistent with the data.

Premises and Definitions

The formalism developed here assumes that the data under consideration accrete in the form of samples from a standard normal distribution, each

individual sample being the final outcome of a single human action to generate data (e.g., pressing a button to start the apparatus). While the only experimental data considered here come from the PEAR program's REG experiments, the normal distribution is ubiquitous enough that it may be hoped the formalism and general arguments have a broader application. In the case of the PEAR REG data, the human action is the initiation of data generation for a single "run," and the resulting standard normal deviate is simply the mean score for that run, as normalized by the expected mean and standard deviation. That is, $x = (m - \mu) / \sigma$, where x is the normalized outcome, m the observed mean, μ the theoretical mean, and σ the theoretical standard deviation.

REG run scores are in fact binomially rather than normally distributed, but since a single REG run involves a minimum of 10,000 bits (and may involve as many as 200,000), the deviations from normality are inconsequential. For analytical purposes, all REG data will be treated here at this level of normalized run deviations. The observed anomalous effect in the PEAR REG experiments, as reported elsewhere (Jahn *et al.*, 1997), is a shift in these mean run scores, correlated with the operator's pre-stated intention.

These experiments involved a tripolar protocol, in which approximately equal amounts of data were generated under three intentional conditions, high, low, and baseline. In the high intention, the operator's goal was to increase the mean value of the data; in the low intention, to decrease it. (The baseline was a passive intentional condition in which the operator was not directed to make any effort.) The only distinction between the two intentional conditions is the direction of effort; any formal analysis whatsoever applies equally to both intentions, up to a sign change. The following discussion, therefore, is written as applying only to the high intention, with "positive" outcomes or shifts being in the direction of intention, and "negative" outcomes or results contrary to the intention. The reader should bear in mind that exactly the same formalism, up to a sign reversal, applies to the low intention.

Because of this theoretical symmetry, all subsequent comparisons between theoretical predictions and data will pool the results from both intentional conditions. A single distribution of intentional outcomes is computed by inverting the sign of all low-intention deviations and combining these inverted results with the high-intention deviations to construct a single population of deviations in the direction of intention.

We will consider two genres of explanatory model for the observed anomalies. The hypothesis that these data reflect an actual change in the machine's operation is referred to as a "mean shift model." In contrast, a "selection model" presumes that certain runs are selectively discarded, depending on their values, so as to bias the distribution statistics of the retained runs and create the spurious appearance of a mean shift. This altered distribution of runs will be called the selected distribution. The undistorted distribution existing before selection will be referred to as the source distribution. It is assumed throughout that, since the selection model is an alternative to an

anomalous change of the distribution, the source distribution is the undisturbed, standard-normal-distributed output of the apparatus. In general, the selection process is probabilistic (reject a fraction of runs of a given value); it can be made deterministic by setting the selection probability for a given value at 0 or 1.

The symbol $p(x)$ will be used to refer to the selected distribution, where x is the run value. The functional notation $f(x)$ will be used to denote the standard normal probability distribution, *i.e.* $f(x) \equiv \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. $F(x)$ refers to the antiderivative of $f(x)$, the cumulative normal probability function: $F(x) = \int_{-\infty}^x f(t) dt$.

By hypothesis, the selection process produces $p(x)$ from a source distribution $f(x)$. The probability that a run of value x will be retained is given by a selection function $s(x)$. Since it is a probability, $0 \leq s(x) \leq 1$ for any x . In general the total integral $S = \int_{-\infty}^{\infty} s(x)f(x)dx$ will be less than 1 (the only exception being the trivial "selection" function $s(x)=1$). Since $p(x)$ should be a properly normalized probability distribution, its value must be $p(x) = s(x)f(x)/S$.

The description of a selection process entails that a certain fraction of the original data has been discarded. For purposes of treating this "filedrawer" quantitatively, we will define the filedrawer quotient Q as the ratio of the amount of discarded data to the amount of data retained. Since the quantity S , defined above, gives the total integral of the selected distribution relative to the original, $Q = (1 - S)/S$.

Accusations of data selection frequently invoke an intuitive but generally incorrect rule of thumb: since the mean shift is a very small fraction of the observed mean value, one supposes that it could be attained by discarding a comparably small fraction of negative data, with the implication that bias or carelessness easily could cause a few critical runs quietly to disappear. The commonest form of this criticism asserts that, since PEAR's effect amounts to 1 part in 10^4 , discarding 1 part in 10^4 of contrary data would be adequate to produce it. Nevertheless, the explicit calculation of filedrawer values will show that, despite the small absolute scale of the effect, the amount of discarded data required to produce it usually would be quite substantial. In some extreme cases, the discarded data would need to be comparable in quantity to the entire reported database. Even in cases where the fraction of discarded data is modest, it is orders of magnitude larger than the naive criticism would suggest, and entails a remarkably large aggregate of actual instances of discarded data due to the large sizes of the observed databases.

General and Limitative Results

While an arbitrarily chosen selection function $s(x)$ is likely to produce a distorted distribution, it is natural to speculate whether this is necessarily so for all such functions. The answer is no: the output of a selection process on a normal source distribution can produce a shifted distribution that is also normal.

Let $f(x, \sigma) = (1/\sqrt{2\pi}\sigma) e^{-x^2/2\sigma^2}$ be the generalization of $f(x)$ to non-unit variance. $f(x, a)$ is a proper probability distribution with total integral 1, for any

nonzero a . Suppose that a selection process applied to standard normal input with distribution $f(x)$ produces output distributed according to $f(x - \mu, \sigma)$ for some positive μ (since the topic of interest is the production of spurious positive mean shifts via selection) and some a not necessarily equal to 1. This implies that

$$Sf(x - \mu, \sigma) = s(x)f(x),$$

since the total integral of $s(x)f(x)$ is S by definition and the integral of $f(x - \mu, \sigma)$ is 1 for any μ, σ . It then follows that $s(x) = Sf(x - \mu, \sigma)/f(x)$, which by substitution of the explicit forms for f can be seen to be

$$s(x) = \frac{S}{\sigma} \exp\left(\frac{1}{2\sigma^2}[(\sigma^2 - 1)x^2 + 2\mu x - \mu^2]\right).$$

Three distinct cases can be distinguished in this formula according to the value of a .

- If $a^2 > 1$, the leading dependence on x is as e^{cx^2} with positive coefficient c . This grows without limit for large x . However, $s(x)$ cannot exceed 1 for any x . Therefore this is not a possible case. Selection cannot produce a normal distribution with greater variance than its source; any selected distribution with increased variance must show some departures from normality.
- If $\sigma^2 = 1$, the leading dependence on x is $e^{\mu x}$. This also grows without limit for increasing positive x , and is therefore again an impossible case. The selected distribution, if normality is preserved, cannot have the same variance as its source.
- If $\sigma^2 < 1$, the polynomial in the exponent, and hence the ratio itself, has a maximum at $x = \mu/(1 - a^2)$. This is a possible situation; thus, a selection process can produce an undistorted normal function as its output, provided the variance of the output normal distribution is less than that of its source.

This result can also be used to derive a relation among the variables S, μ , and a , on the assumption that $s(x) = 1$ at its maximum. This gives the largest possible value of S , and hence the smallest possible value of $Q = (1 - S)/S$, the filedrawer quotient. Solving the equation $Sf(x - \mu, \sigma) = f(x)$ at $x = \mu/(1 - a^2)$ gives, after some algebra, the relation

$$S = \sigma \exp\left[-\frac{1}{2}\left(\frac{\mu^2}{1 - \sigma^2}\right)\right].$$

It is obvious that S vanishes at $a = 1$, as the argument of the exponential goes to $-\infty$. It is equally obvious that S vanishes at $a = 0$. The derivatives of S in the region between are

$$\frac{\partial S}{\partial \mu} = -\frac{\mu}{1 - \sigma^2} S; \quad \frac{\partial S}{\partial \sigma} = S \left[\frac{1}{\sigma} - \frac{\sigma \mu^2}{(1 - \sigma^2)^2} \right].$$

Since μ is positive by assumption, and S is nonnegative, it is obvious that $\partial S/\partial \mu < 0$ anywhere S is nonzero. This accords with intuition: the larger the shift in the distribution mean, all else being equal, the more stringent the selection must be. Setting $\partial S/\partial \sigma = 0$ and applying the quadratic formula shows that S has exactly one maximum in a in the allowed range $0 < a < 1$, at

$$\sigma_{\max}^2 = 1 + \frac{\mu}{2} \left(\mu - \sqrt{4 + \mu^2} \right).$$

This again accords with intuition: if a is too small, much of the source distribution must be discarded in order to make the selected distribution narrow enough. If a is too large, much of the source distribution must be discarded to match the limit imposed by available data in the upper tail.

For a general consideration of what is possible with selection, the important points to note from these formulae are:

1. For any given value of a in the selected distribution, increasing μ requires a decrease in S .
2. For any particular value of μ in the selected distribution, there is an optimal value of a that maximizes S . As μ increases, the optimal value of a decreases, as does S at that value of a .
3. Any attempt to increase a beyond the value that maximizes S leads to a rapid decrease in S , since $S = 0$ at $\sigma = 1$ is required.

Point 3 in particular means that there is an unavoidable tradeoff between the degree of distortion from the original distribution, and the amount of data discarded to achieve it; the more nearly the selected distribution approaches an otherwise undistorted, mean-shifted version of the source, the smaller the fraction of the source is being retained.

The discussion above proves that, while a selection process can produce a shifted but still normal selected distribution from a normal source distribution, the output distribution necessarily has smaller variance than the source. Moreover, the larger the mean shift, and/or the more closely the selected distribution approaches the source distribution's variance, the more data must be discarded in the selection process.

A more general selection process that is not constrained by producing normal output need not obey this variance rule: for example, if all data smaller than a certain absolute value are discarded, the selected distribution will have greatly increased variance. It will also be bimodal and hence grossly non-normal. Given the above result for normal output, in which all properties of the selected distribution except the variance are fixed (since the target mean shift is chosen in advance), it seems reasonable that more general selection processes must be subject to a similar tradeoff between the intensity of selection and the degree of departure from normality in the output. This conjecture will be revisited after some development and examination of specific selection models in the following sections.

Some Plausible Selection Models

The selection function $s(x)$ can vary arbitrarily at every x , subject only to the constraint $0 \leq s(x) \leq 1$ for all x . However, a real selection process is not likely to employ an arbitrarily complicated selection function. The selection models discussed below attempt to sample a reasonably broad range of the space of practical and plausible models.

Each heading below describes a one-parameter family of models, where the mean and all other distribution statistics are determined by a single free parameter. The reason for examining one-parameter families is that once the observed mean is fit by choosing an appropriate value of the free parameter, the remaining distribution statistics have been fixed; this places each model on a footing comparable to the one-parameter mean shift model. We adopt the convention of calling the selection parameter a in all cases. The functional forms of the various models are illustrated graphically in Figure 1.

- **Simple Cutoff.** The simplest possible data selection model is simply to reject all runs with a value less than some fixed cutoff a ; $s(x) = 0$ for $x < a$, 1 otherwise. This model bears some theoretical importance in that it is provably the model with the smallest filedrawer quotient for any given mean shift. (The proof is almost trivial: it rejects all those and only those data with the greatest negative contribution to the mean. Changing the rule in any way will therefore reduce the mean shift generated per discarded data point, and thus require that more data be discarded to achieve the same mean shift.) The fact that the cutoff model also produces the greatest departures from normal distribution statistics of any model examined is therefore strong support for the general "tradeoff" thesis of the preceding section.
- **Fractional Rejection.** It seems overly simplistic to expect that the simple cutoff model ever would appear in actual data. Only the most naïve of frauds could imagine that the total disappearance of runs below some set value could be invisible; even the most biased of researchers would need phenomenal powers of self-delusion to convince themselves that *all* runs, and *only* those runs, lying below a particular value were methodologically invalid. The fractional rejection model supposes that a constant fraction a of all unsuccessful ($x < 0$) runs are discarded:

$$s(x) = 1 - a \text{ for } x < 0, 1 \text{ otherwise.}$$

- **Short Left Tail.** For a further increase in psychological plausibility, we can suppose that an increasing rate of exclusion is employed as the value of the run goes more negative. For a simple representation of this possibility, a short left tail selected distribution compresses the variance for all negative values of x . In other words, for some compression factor $a < 1$,

$$p(x) = \frac{2}{1+a} \times \begin{cases} f(x), & \text{for } x > 0; \\ f(x/a), & \text{for } x < 0. \end{cases}$$

where $p(x)$ has been renormalized to a proper probability distribution. The selection function for negative x is

$$s(x) = \frac{f(x/a)}{f(x)} = \exp\left\{\frac{x^2}{2} \left(1 - \left(\frac{1}{a^2}\right)\right)\right\}; \text{ as before } s(x) = 1 \text{ for } x > 0.$$

- **Increasing Bias.** Rather than presuming that only negative runs will be rejected, we may suppose that an experimenter preference for positive results may manifest itself as a graduated probability of rejection. For this model we presume that runs that exceed the positive one-tailed $p = 0.05$ significance criterion $x = 1.645$ are always retained, while runs with $x < -1.645$ are rejected with probability a . In the region $-1.645 < x < 1.645$, the rejection probability drops linearly from a to 0 as x increases.

Thus

$$s(x) = \begin{cases} 1 - a, & \text{for } x < -1.645; \\ 1 - a/2 + ax/3.29, & \text{for } -1.645 < x < 1.645; \\ 1, & \text{for } x > 1.645. \end{cases}$$

This enumeration is not intended to imply that unconscious bias would work by any such explicit or exact means as the example selection functions. Indeed, it seems virtually certain that data selection due to unconscious bias would operate in a messy and inexact fashion driven by a host of psychological considerations. The intent of employing these several families of models, to the contrary, is to provide functional forms that are relatively amenable to calculations, for models that capture reasonable qualitative properties of the unconscious bias process. Taking the last as an example, it seems reasonable to suppose that a biased experimenter would be most eager to retain significant runs, and most willing to discard significantly negative runs, with an intermediate level of preference for intermediate cases. While it would be ridiculous to propose that such an experimenter would unconsciously be calculating the three-part $s(x)$ function given above, this $s(x)$ gives a computable quantification of the specified qualitative features; we therefore reasonably may argue that any psychological process fitting the given qualitative description should produce output statistics similar to those of our exemplar, chosen for its computational convenience.

Figure 1 displays examples of all of these functions. The uppermost plot shows the standard normal distribution in the interval $[-3, 3]$. The next plot is a mean-shifted normal, corresponding to the mean shift model which is our representation for a genuine anomalous effect. The mean of the distribution has been set at 0.3, a size appropriate to the scale of effect in some of the more successful PEAR datasets. The third plot shows the cutoff distribution with the same mean. (All of the selected distributions in Figure 1 have been normalized to have the same area as the shifted normal with which they are compared.) The next three plots show

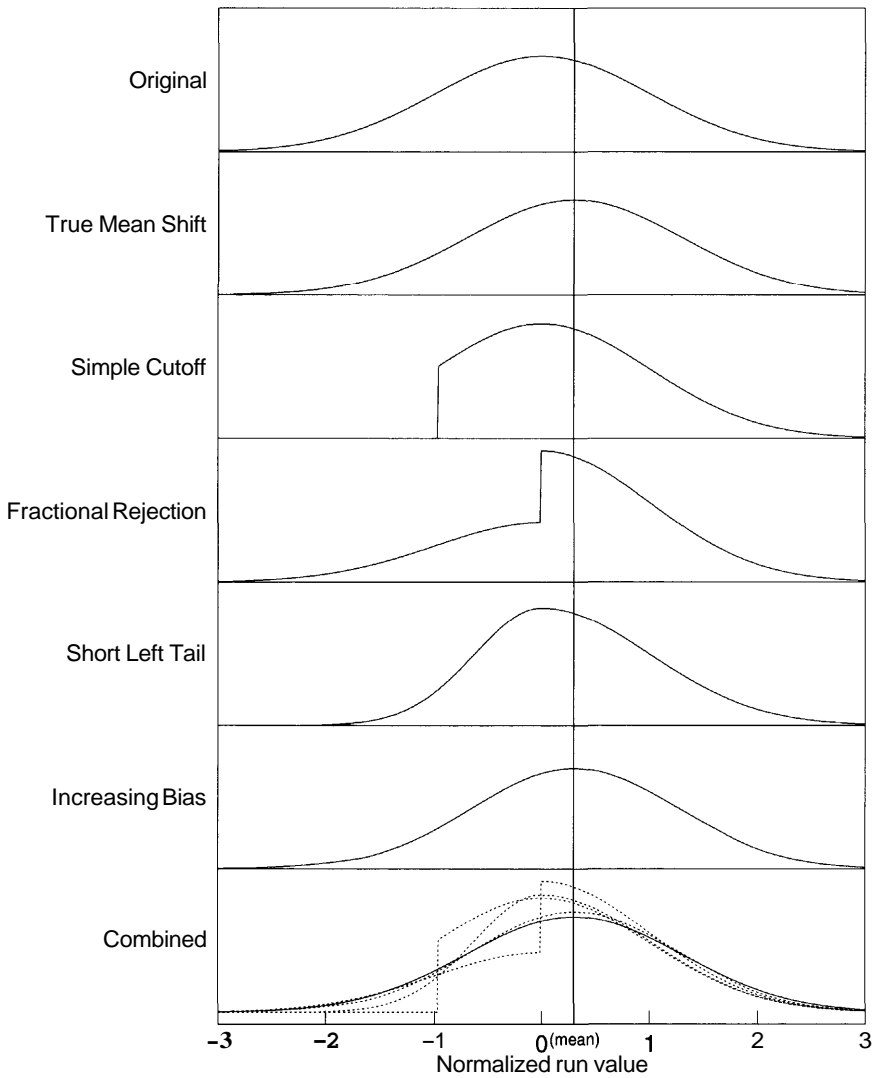


Fig. 1. Selection model distributions.

the fractional rejection model, the short tail model, and the increasing bias model, all for the same mean of 0.3. Finally, the last plot shows all four selection models superimposed as dotted lines on the shifted normal with the same mean.

Comparison Methods and Statistical Power

Most selection processes produce non-normal selected distributions, and even the normality-preserving selection process produces a selected distribu-

TABLE 1
Statistical Power Comparison—Required N

Selection model	χ^2 Goodness-of-fit	σ	Skewness	Kurtosis
Simple cutoff	192	30	44	1.47×10^5
Fractional rejection	1920	638	203	460
Short tail	1040	43	127	7551
Increasing bias	2.28×10^5	638	3510	5870

tion with a smaller variance than its source. Given the illustrations of Figure 1, it might seem natural to test the selection models against the mean shift model by comparing their distribution densities directly, *e.g.* with a χ^2 test for goodness of fit on some appropriate binning scheme. In fact, higher statistical resolution can be achieved by instead computing some of the higher moments of the respective distributions. (See Appendix for a more detailed discussion.) The mean shift model predicts that, regardless of mean shift, the standard deviation of the normalized data should be $a = 1$; the normalized third central moment (skewness) should be $\gamma_3 = 0$; and the normalized fourth central moment (kurtosis) should be $\gamma_4 = 3$. Examining some or all of these parameters will detect the difference between a mean shift distribution and a selected distribution for considerably smaller numbers of data points than a general χ^2 test.

Table 1 shows the statistical power of tests on the higher moments, for the same effect size 0.3 shown in Figure 1, by giving the number of data points required for an $\alpha = 0.05$, $\beta = 0.50$ test of each of the above selection models against the corresponding mean shift model. That is, if the mean shift model for $\mu = 0.3$ is taken as the null hypothesis, and the data actually are being generated by one of the selection models with a chosen to replicate the same value of μ , Table 1 reports the N at which the expected value of the given test statistic is at the $p = 0.05$ significance level ($\alpha = 0.05$).^{*} This corresponds to the point where, if the null hypothesis is false, the test is equally likely to produce results which are significant or nonsignificant under the aforementioned α criterion, leading to $\beta = 0.5$ (probability of erroneously accepting a false null hypothesis). It should be noted that $\beta = 0.5$ is an unsatisfactorily high probability of Type II error; it is used here not prescriptively but for convenience of calculation, since Table 1 is intended solely to demonstrate the relative sensitivity of different tests for detecting the various selection models.

From Table 1 we see several features important to the identification of the most sensitive test:

^{*} When the test statistic is a moment parameter, its expected value is found by applying Equations 1–5, given in the next section. The expected value of the χ^2 is found by direct computation of the difference between each of the selected distributions and a normal distribution with the same mean and unit variance.

- At least one of the three central-moment tests always outperforms the distribution-based χ^2 by a large margin, as would be expected from the analysis in the Appendix.
- In three of the four cases, the most sensitive moment test is on the standard deviation σ ; in the remaining case it is on skewness. The kurtosis test is never the most sensitive.
- The most sensitive moment test, for each selection model, requires appreciably less data than the second-best test; the difference is sometimes considerable.

A feature which does not appear in Table 1 is the fact that the identity of the most sensitive moment parameter depends on the scale of the effect. For example, in Table 1 the cutoff model can most readily be detected by its change in the standard deviation; however, for a mean shift of 0.03, an order of magnitude smaller than that used in Table 1 (and more typical of general PEAR databases), the most sensitive test for the cutoff model becomes the skewness, γ_3 . In light of this variability, the best procedure for testing whether data are consistent with a particular selection model would seem to be to compute the statistical power of each test for the given effect size and use the most sensitive.

Selection Model Moment Parameters

Both the statistical power calculations discussed in the previous sections and actual comparisons with empirical data require that we calculate the higher moments of a selected distribution from its mean shift. All of the selection rules allow the mean and higher moments to be calculated from a given a ; although the mean shift in terms of a is seldom given as an invertible function, procedures such as a numerical binary search readily can be used to find the a that corresponds to a desired mean shift, and the higher moments then can be calculated directly. Figure 2 illustrates the results of such a process for the cutoff distribution, with the cutoff parameter, standard deviation, skewness, and kurtosis presented as functions of the mean shift.

The functional forms of the mean and higher moments of each selection model, as functions of the free parameter a , can be calculated by straightforward integrations. For example, for a given cutoff parameter a the cutoff distribution has the following distribution moments and filedrawer parameters:

$$\begin{aligned}
 \text{Mean: } m &= f(a)/(1 - F(a)) = f(a)/F(-a) \\
 \text{Variance: } \sigma^2 &= 1 + ma - m^2 \\
 \text{Skewness: } \gamma_3 &= \frac{2m^3 - 3am^2 + (a^2 - 1)m}{\sigma^3} \\
 \text{Kurtosis: } \gamma_4 &= \frac{3 + m(a^3 + 3a) - m^2(4a^2 + 2) + 6m^3a - 3m^4}{\sigma^4} \\
 \text{Filedrawer: } S &= 1 - F(a); \quad Q = \frac{F(a)}{1 - F(a)}.
 \end{aligned} \tag{1}$$

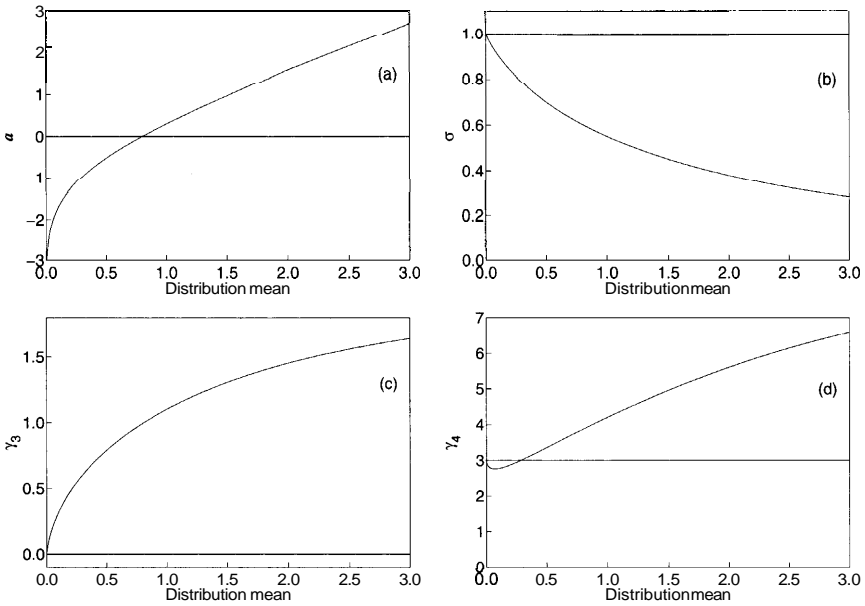


Fig. 2. Evolution of cutoff distribution with mean shift (a) cutoff parameter, (b) standard deviation, (c) skewness, (d) kurtosis.

It can be seen that, even for this relatively simple function, the functional form of the higher central moments becomes somewhat involved. For the remaining distributions it will be somewhat more straightforward to report their moments $\langle x^2 \rangle, \langle x^3 \rangle, \langle x^4 \rangle$ rather than the statistical parameters, which are determined by the central moments. That is, for any distribution whatsoever, the variance, skewness, and kurtosis are given by:

$$\begin{aligned}
 \sigma^2 &= \langle (x - \langle x \rangle)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2 \\
 \gamma_3 &= \frac{\langle (x - \langle x \rangle)^3 \rangle}{\sigma^3} = \frac{1}{\sigma^3} [\langle x^3 \rangle - 3\langle x \rangle \langle x^2 \rangle + 2\langle x \rangle^3] \\
 \gamma_4 &= \frac{\langle (x - \langle x \rangle)^4 \rangle}{\sigma^4} = \frac{1}{\sigma^4} [\langle x^4 \rangle - 4\langle x \rangle \langle x^3 \rangle + 6\langle x \rangle^2 \langle x^2 \rangle - 3\langle x \rangle^4].
 \end{aligned}
 \tag{2}$$

The standard normal distribution has $\langle x \rangle = 0, \langle x^2 \rangle = 1, \langle x^3 \rangle = 0,$ and $\langle x^4 \rangle = 3,$ hence its expected values of mean, standard deviation, skewness, and kurtosis are 0, 1, 0, and 3, respectively.

With Equations 2 in place, we may describe the parameters of the other distributions somewhat more concisely. For the fractional rejection distribution, described by the rejection rate $a,$ the first four moments and filedrawer quotient are given by:

$$\begin{aligned}
 \langle x \rangle &= \frac{af(0)}{1-a/2} & \langle x^2 \rangle &= 1 \\
 \langle x^3 \rangle &= \frac{2af(0)}{1-a/2} & \langle x^4 \rangle &= 3 \\
 Q &= \frac{a}{2-a}.
 \end{aligned} \tag{3}$$

For the short-left-tail distribution, described in terms of the contraction factor a , the corresponding values are

$$\begin{aligned}
 \langle x \rangle &= 2f(0)(1-a) & \langle x^2 \rangle &= 1-a+a^2 \\
 \langle x^3 \rangle &= 4f(0)(1-a+a^2-a^3) & \langle x^4 \rangle &= 3(1-a+a^2-a^3+a^4) \\
 Q &= \frac{1-a}{1+a}.
 \end{aligned} \tag{4}$$

And finally, for the increasing bias distribution, if for conciseness we define $b = 1.645$ for the transition points of $s(x)$, the moments and filedrawer quotient are

$$\begin{aligned}
 \langle x \rangle &= \frac{a}{\gamma} (2F(b) - 1) & \langle x^2 \rangle &= 1 \\
 \langle x^3 \rangle &= \frac{a}{(2-a)b} (6F(b) - 2bf(b) - 3) & \langle x^4 \rangle &= 3 \\
 Q &= \frac{a}{2-a}.
 \end{aligned} \tag{5}$$

Relation to General Results

Figure 3 applies Equations 1-5 to illustrate the ease of detection, and the filedrawer parameter required, for the four models as the mean shift μ is increased from 0 to 0.5. The "distortion index" plotted in the top figure is the rate of growth with N (the number of observed data points) of a Z -score describing the departure of the most sensitive parameter from its theoretical value. (When N is large enough for the usual normal approximations to the variation in a , γ_3 , and γ_4 to be useful, then the distortion index, multiplied by \sqrt{N} , gives the expected value of $|Z|$ for the most sensitive statistic.) The bottom graph simply plots the filedrawer quotient for that model at that mean shift.

It is conspicuous from the figure that the models, despite having grossly different distributions and being best detected by different distribution parameters, obey a generalized form of the results rigorously derived for normality-preserving selection functions. As μ increases, both the distortion and the filedrawer of any particular model increase monotonically. At every μ , the order of models ranked by increasing distortion is exactly the reverse of their order ranked by increasing filedrawer; changing to a model with less statistical distortion always increases the filedrawer quotient.

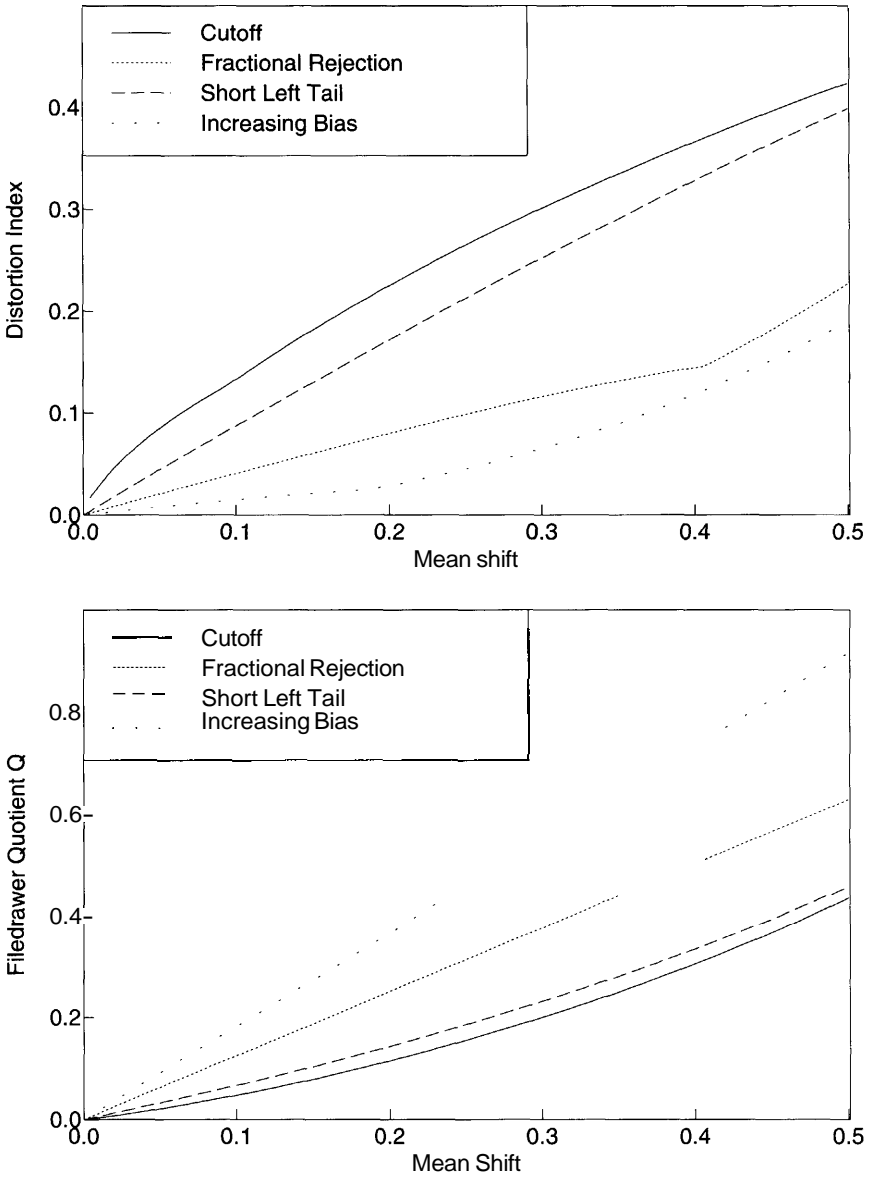


Fig. 3. Distortion and filedrawer for four selection models.

REG Data

The PEAR REG experiment involved the collection of data from a device with no known non-anomalous channels for the operator's influence. While full

details of the experimental protocols and controls are available elsewhere (Jahn *et al.*, 1987; Dunne & Jahn, 1995), a brief summary may be in order.

- **Redundant Recording.** The raw data were printed on a continuous paper tape, and concurrently entered into a computer file. Summary data were also recorded by the operator in a logbook. In any case where a discrepancy appeared among the three records, the paper tape record was given precedence.
- **Advance Designation of Intention.** The operator was required to declare an intention before data were generated. Data collection could not be initiated until the intention was entered and logged in the experimental computer.
- **Continuity of Record.** The paper tape record was required to be continuous, without gaps or breaks, as a safeguard against precisely the sort of selection discussed in this analysis. As an aside, it should be noted that this requirement of physical integrity of the paper tape and the primacy of the tape record over the other redundant records in case of disagreement also provided strong safeguards against the alteration of extant data or the introduction of spurious data. Such interventions would be considerably more difficult than the already challenging task of making data disappear from the records.

Data were collected in runs of 50, 100, or 1000 trials, where one trial is the sum of successes in 200 binary $p = 0.5$ events. If spurious data selection were attempted, the finest possible scale of intervention would have been the run level; picking and choosing what data to retain at the level of individual trials would require a massive invasion of all the data recording systems, both hardcopy and electronic, and involve far more labor than fabrication of the entire database from whole cloth.

The three different run lengths mentioned above must be treated separately in order to discriminate properly between selection and mean shift models. The reason is that a mean shift model, in the absence of additional qualifying hypotheses, predicts an effect that is constant at the trial level, and therefore predicts that the average Z-scores of runs depend on how many trials comprise them. Therefore, a mixture of Z-scores for runs of different lengths would be an intrinsically heterogeneous database under one of the hypotheses being compared, rendering all statistical comparisons suspect.

Table 2 displays the statistics for the three run lengths present in the primary REG database. The two active intentions have been combined, with the sign of deviations in the low intention reversed to produce a uniform measure of deviation in the direction of intention. The statistical uncertainty (1σ) for each parameter also is given, along with the Z-score for its deviation from the expected value. Note that all of the Z-scores for higher moments are nonsignificant, indicating that the data are at least consistent with a mean shift model.

In accordance with the discussions in "Comparison Methods and Statistical Power", we may proceed now to calculate the statistical power of testing the various moments of the selection models on each of these databases. Table 1 presented a minimum N for having a probability $\beta = 0.5$ of failing to distinguish

TABLE 2
REG Run Distributions

Run length	N_{runs}	Mean	SD	Skewness	Kurtosis
50	12,849	0.0277 ± 0.0088 $Z = 3.1369$	1.0077 ± 0.0062 $Z = 1.2316$	0.0085 ± 0.0216 $Z = 0.3939$	2.9553 ± 0.0432 $Z = -1.0344$
100	3340	0.0402 ± 0.0173 $Z = 2.3252$	0.9932 ± 0.0122 $Z = -0.5593$	-0.0371 ± 0.0424 $Z = -0.8763$	2.9514 ± 0.0848 $Z = -0.5732$
1000	700	0.0485 ± 0.0378 $Z = 1.2828$	0.9881 ± 0.0267 $Z = -0.4449$	-0.0052 ± 0.0926 $Z = -0.0561$	3.2692 ± 0.1852 $Z = 1.4536$

a selection model from the undisturbed mean shift hypothesis. Since the amount of formal data in hand is fixed and cannot be modified, Table 3 instead presents β for each parameter test, on each database, where a standard $\alpha = 0.05$ criterion for rejecting the mean shift hypothesis is assumed.

The listing of β in Table 3 values allows easy identification of the most powerful test for detecting the presence of each selection model in a given dataset: simply choose the parameter with the smallest β value for that model in that dataset. Unfortunately it also is clear that, for these effect sizes and database sizes, the statistical power is too low to distinguish some of the models from the mean shift model. In particular, the lowest β value that appears for the increasing bias model is 0.923; that is, even with the most sensitive test on the best database for the purpose, there is a 92.3% likelihood that a database actually produced by the increasing bias process would fail to produce a test statistic significantly different from the expected value for a mean shift model.

Table 4 lists the predictions for the various models on these datasets. In each case the single free parameter of the model is used to fit the observed mean; the standard deviation, skewness, and kurtosis then follow from the functional form of the model. In addition to the model predictions for these parameters, Table 4 gives the Z -scores for the empirical value of the given parameter (as reported in

TABLE 3
Statistical Power β

Dataset	Parameter	Cutoff	Fractional rejection	Short tail	Increasing bias
50-trial runs	SD	4.5×10^{-4}	0.950	0.209	0.950
	Skewness	3.7×10^{-6}	0.751	0.744	0.923
	Kurtosis	3.2×10^{-3}	0.949	0.950	0.950
100-trial runs	SD	0.042	0.950	0.708	0.950
	Skewness	0.012	0.900	0.898	0.943
	Kurtosis	0.241	0.950	0.950	0.950
1000-trial runs	SD	0.497	0.950	0.796	0.950
	Skewness	0.408	0.918	0.916	0.946
	Kurtosis	0.755	0.950	0.950	0.950

TABLE 4
Model Predictions

Model	σ	$Z(\sigma)$	Skewness (γ_3)	$Z(\gamma_3)$	Kurtosis (γ_4)	$Z(\gamma_4)$
50-trial runs						
Mean shift	1.0000	1.2316	0.0000	0.3939	3.0000	-1.0344
Cutoff	0.9670	6.5141	0.1393	-6.0508	2.7973	3.6551
Fractional rejection	0.9996	1.2930	-0.0277	1.6740	3.0031	-1.1053
Short tail	0.9827	4.0005	0.0282	-0.9091	3.0006	-1.0474
Increasing bias	0.9996	1.2930	-0.0104	0.8752	3.0012	-1.0611
100-trial runs						
Mean shift	1.0000	-0.5593	0.0000	-0.8763	3.0000	-0.5732
Cutoff	0.9549	3.1240	0.1794	-5.1088	2.7743	2.0892
Fractional rejection	0.9992	-0.4931	-0.0402	0.0722	3.0065	-0.6496
Short tail	0.9749	1.4893	0.0413	-1.8497	3.0012	-0.5875
Increasing bias	0.9992	-0.4931	-0.0151	-0.5206	3.0024	-0.6019
1000-trial runs						
Mean shift	1.0000	-0.4449	0.0000	-0.0561	3.0000	1.4536
Cutoff	0.9474	1.5217	0.2029	-2.2479	2.7655	2.7200
Fractional rejection	0.9988	-0.4009	-0.0484	0.4669	3.0094	1.4028
Short tail	0.9698	0.6838	0.0500	-0.5959	3.0018	1.4440
Increasing bias	0.9988	-0.4009	-0.0181	0.1395	3.0035	1.4346

Table 1) relative to the model prediction. It is divided into three sections, for the three different datasets.

It is clear that the cutoff model is rejected for all three datasets. The short tail model is also strongly rejected for the largest dataset, that of 50-trial runs. Both the fractional rejection model and the increasing bias model are consistent with the statistics of the actual data, just as the mean shift model is. This ambiguous result is only to be expected, given the β values listed in Table 3, where the most sensitive tests have $\beta = 0.751$ for the fractional rejection model and $\beta = 0.923$ for the increasing bias model. We simply do not have enough data to distinguish these models reliably from the mean shift model on the strength of any statistical parameter of the distribution. Therefore, the resolution of the possibility of data selection must turn, not on the statistical parameters of the formal data distributions, but on the filedrawer quotients describing the amount of absent data.

Missing Data: Void Runs

Aside from the published data reported in Table 2, some data have of course been discarded. The formal protocol, although it has changed over time, always has mandated the invalidity of data collected under certain protocol-violating conditions. To the greatest extent possible, these void criteria have been designed with the intent of eliminating the human decision factor, and therefore the possibility of biased preferences.

The standard protocol requires that there be some record of the existence of every occasion on which formal data were generated, or even when an attempt was made to generate formal data. Violation of this protocol condition would require a deliberate attempt to deceive; the consequences of such efforts will be discussed further below, but for the moment we are concerned only with the possible impact of bias on the decision to reject data.

In the majority of cases, the void run generated data which were recorded, and statistical summaries were computed. Generally these are cases where some protocol violation mandated that the data be considered invalid; *e.g.*, during a period when the formal protocol required a minimum of 5 runs per session, some sessions of fewer than 5 runs were generated due to operator misunderstandings of the protocol. These performances were declared void and excluded from the formal database. During the same period, operators were permitted to complete a series over the course of multiple laboratory visits (since a series might include up to 300 runs, requiring over 5 hours of the operator's time). Some operators never returned to complete a series, and in these cases also the experimenters were obliged to mark the data as void.

In some cases data were declared void due to an equipment malfunction of such nature as not to preclude data generation or recording. For example, there were several occasions on which runs were generated with internal (inter-trial) standard deviations of 14–18 rather than the theoretical $\sqrt{50} = 7.071$; this grossly aberrant output was taken as sufficient demonstration that the noise source had suffered a breakdown and was no longer emitting properly conditioned random values. (Indeed, in these cases physical intervention was required to restore proper operation of the device.) In other cases, individual runs were declared void due to protocol violations such as the unexpected and disruptive arrival of visitors.

In some cases, no data values were recorded for the void runs. Much of the time this was due to equipment failures that made it impossible to record data, as for example during a period when the automated data collection was handled not by a local computer but by a remote connection to a departmental server, which was prone to unpredictable downtime and sometimes failed during an experiment. Sometimes, however, these episodes were due to operator errors in the conduct of the experiment, or were caused by a problem such as a disruptive visit actually prevented the recording of data rather than merely interrupting the operator.

The formalism discussed above addresses the statistical features left behind in a population of observed and recorded data, as a consequence of the construction of that population, by discarding and concealing a selected component of the total source distribution. It is clearly fatuous to apply this technique to the population of void runs with recorded values: the impact that their removal has had on the data can be calculated directly, simply by restoring them to the experimental population. On the other hand, since at least some of the void data with known values are products of a random source known to be malfunctioning

at the time, including them as part of the data under analysis violates one of the assumptions of the formalism, namely that the output of the experimental apparatus follows a standard normal distribution.

The best resolution of the situation with the two classes of void runs would seem to be as follows. The distribution of those voids *with* values can be computed; it can be compared, both with the null hypothesis of zero effect and with the observed effect size in the formal data, for any evidence of bias in its removal from the database, and it can be recombined with the formal data to establish its impact, if any, on the scale and significance of the anomalous effect. The voids *without* values, in contrast, comprise a population of missing data which properly should be compared with the filedrawer quotients predicted for the various selection models. These predictions should, however, be based on the mean shift and population size of the formal data alone, not the formal data recombined with voids of known value, since these latter are not in all cases drawn from the same distribution.

Table 5 summarizes the void data present in each of the three databases. It gives the numbers of each type of void run; for the voids with values, it additionally gives the number of voids with results in and contrary to the direction of intention and the Z-score of this count imbalance. (These totals do not add to the total count of voids with values in the 50-trial runs, because six of these runs had means of exactly 100.00 which is neither in nor contrary to the direction of intention.) Also, the mean and standard deviation of the population of voids with values, the Z-score of this population against the null hypothesis, and a two-population T-score for the difference between the voids and the formal data are given. Finally, the overall Z-score for the anomalous mean shift is recomputed with the void data added to the formal data.

The uniformly negative means of the void populations suggest that, despite all efforts, some degree of bias was present in the rejection of these data. None of the three void populations differs significantly from a null hypothesis, however.

TABLE 5
Void Runs, by Database

	50-trial runs	100-trial runs	1000-trial runs
N without values	184	1	3
N with values	590	80	24
... Matching intention	277	44	9
... Against intention	307	36	15
Z of count imbalance	-1.2414	0.8944	-1.2247
Mean Z of voids	-0.0422	-0.0080	-0.2752
σ	1.0504	1.0604	0.7319
Z vs. null	-1.0247	-0.0712	-1.3483
T vs. formal	-1.5824	-0.4023	-2.1020
Z, formal only	3.1369	2.3252	1.2828
Z, formal + void	2.8526	2.2869	1.0158

The composite Z for a difference from the null, across all three sets, is -1.2197 , entirely consistent with chance variation.

The population counts of void runs in and contrary to the direction of intention provide a secondary check of the existence of simple forms of bias. It may be noted that in the 100-trial run length the number of void runs in the direction of intention actually exceeds the number of void runs contrary to intention. Of the total population of void runs across all three categories, there are 358 contrary to intention, 330 in the direction of intention, and 6 null, producing a net $Z = -1.0675$ against a hypothesis that a void run is equally likely to be in, or contrary to, the direction of intention.

The void 1000-trial runs do differ significantly ($p = 0.036$, two-tailed) from the formal 1000-trial runs, and the meta-analytic combination of these three T -scores produces a marginally significant composite $Z = -1.9867$ ($p = 0.047$, two-tailed). While this significant difference could be taken as evidence for real bias in the void selection process, it also should be noted that the combination of no significant deviation from the null with a significant deviation from the formal data is consistent with the existence of a genuine effect in the formal data, provided that the circumstances which led to the rejection of a run as void also were such as to impede any anomalous effect. Since the most important criteria for voiding involved equipment breakdowns of various sorts, and unavoidable external interruption or distraction of the operator, this last would seem a reasonable expectation.

The recalculated Z -scores that include the voids along with the formal data are, for obvious reasons, slightly decreased. As a result, the composite Z -score representing the overall evidence for an anomaly, which is 3.8087 for the formal data, is reduced to 3.4439 when the voids are included, a reduction of approximately 9.6%. We may conclude from this that there is marginal evidence for a bias in the selection of void runs with values, but that it does not substantially impact the evidence for the existence of an effect. While it might produce a small distortion in the observed effect size, this distortion is smaller than the statistical uncertainty in that observation.

Having resolved the interpretation of those void runs which have recorded values, the next stage is to consider the voids with no values by applying the selection formalism. As discussed above, these voids without values are the appropriate population of missing data to be compared with the filedrawer quotients computed for each model. Table 6 compares the actual filedrawer quotients Q , as computed from the formal data populations in Table 2 and the void populations in Table 5, with the theoretical values required by the fractional rejection, short tail, and increasing bias models. (The cutoff model has been dropped from Table 6 since its statistical predictions have already been shown to be completely incompatible with all datasets.)

The Z -scores presented in Table 6 are computed from the mean and standard deviation of the binomial distribution for the theoretical rejection rate mandated by the model. That is, the predicted rejection rate for a given model, in

TABLE 6
 Filedrawer Populations

Item	50-trial runs	100-trial runs	1000-trial runs
Number of formal runs	12,849	3340	700
Void runs without values	184	1	3
Real filedrawer Q	0.0143	0.0003	0.0043
Fractional rejection model			
Predicted Q	0.0347	0.0504	0.0608
Predicted void population	446	168	43
Z, real vs. prediction	12.3182	12.8824	6.0869
Short tail model			
Predicted Q	0.0176	0.0259	0.0313
Predicted void population	226	87	22
Z, real vs. prediction	2.7740	9.2188	4.0422
Increasing bias model			
Predicted Q	0.0506	0.0735	0.0886
Predicted void population	650	245	62
Z, real vs. prediction	18.1485	15.5863	7.4770

conjunction with the known population of retained data, gives us both an expected population of voids, and a standard deviation for that expected population. The observed population of void runs can then be compared with that theoretical mean and standard deviation to obtain a Z-score, which is the source of the Z values listed in Table 6. The sign of the difference was ignored in computing these Z-scores, so if p-values are calculated for them, the two-tailed form must be used.

It is evident that all of the models predict rejection rates grossly in excess of the actual number of void runs without recorded values. In many cases the mismatch is so extreme that p-values cannot be calculated readily by conventional techniques. The least significant result is the value of 2.774 for the short tail model on the 50-trial runs, with a two-tailed p -value of 0.006. This particular model, however, can already be rejected on the basis of its distribution statistics as discussed above. No model examined here can plausibly accommodate both the observed distribution statistics and the known rate of run rejection.

Additional General Arguments: Models with More Parameters

Except for the cutoff model, which demonstrably has the smallest possible filedrawer for a given mean shift, the selection models examined above are not obviously optimal or extremal in their properties. While they show the same relationship between distortion and filedrawer expected from the analysis of normality-preserving selection, it seems worthwhile to sample the space of possible $s(x)$ in somewhat more detail to test the generality of this property.

An immediate generalization can be obtained by noting that the cutoff model and the fractional rejection model are two extreme members of a single, two-parameter family of selection processes. This family of "fractional cutoff"

models rejects a fraction r of all data falling below a minimal cutoff boundary b . The simple cutoff model is then the fractional cutoff model with $b = a$ and $r = 1$; the fractional rejection model is the fractional cutoff model with $b = 0$ and $r = a$. A choice of a specific mean shift does not, of course, identify a specific member of this two-parameter family, but rather determines a one-parameter subset of it.

For more general investigations, one may begin to approximate the freedom of the full, arbitrary $s(x)$ by adopting a sectional selection model. For some n , choose $n - 1$ boundaries b_1, \dots, b_{n-1} in the real line, along with the notional "boundaries" $b_0 = -\infty$ and $b_n = +\infty$. For convenience, and without appreciable loss of generality (since n is not fixed), these boundaries can be placed at uniform quantiles of the inverse normal distribution, so that $\int_{b_{k-1}}^{b_k} f(x)dx = 1/n$ for each $k \in 1, \dots, n$. The selection function $s(x)$ is then taken as the piecewise continuous function defined by $s(x) = s_k, b_{k-1} < x < b_k$, for each $k \in 1, \dots, n$.

It is obvious that for this n -section selection function $S = (1/n)\sum_{k=1}^n s_k$. The higher moments are given by $\langle x^m \rangle = (1/S)\sum_{k=1}^n s_k I(b_{k-1}, b_k, m)$, where the integral function I is defined as $I(a, b, m) = \int_a^b x^m f(x)dx$; note that the I terms depend only on the set of b s and are the same for any selection function using the same set of sections.

Figure 4 shows numerous selection models plotted against their distortion index D and filedrawer quotient Q . All models in this figure are constructed to have $\mu = 0.0277$, identical to the 50-trial runs subset. The shaded area in the lower left corner identifies the limits of Q and D imposed by a $Z < 2$ criterion of consistency with the observed characteristics of the 50-trial runs; any model outside that area will be rejected at $p < 0.05$ (two-tailed) or better for its distribution shape, its filedrawer prediction, or both.

The four filled markers show the four one-parameter models, as labeled on the figure. The smooth solid line shows the behavior of the normality-preserving selection model, where the distortion index is purely driven by the change in variance. The dotted line shows the evolution of the two-parameter fractional cutoff model as it is smoothly changed from the cutoff model to the fractional rejection model while maintaining constant mean shift. It is noteworthy that this curve passes very close to the point characterizing the short left tail model, despite their very different functional forms. It is also intriguing that as the cutoff boundary b migrates toward zero, the index of distortion passes through a minimum and then begins to increase again, while the filedrawer quotient increases monotonically throughout.

The remaining elements of this figure are based on sectional selection models with 100 sections. The cross icon is the sectional model that most nearly matches the optimal (maximal S) normality-preserving selection; as one might expect it lies on the curve for such models, at its minimum in Q (which is its maximum in S given $Q = (1 - S)/S$). The open circle is the 100-section sectional model defined by $s_k = \alpha_0 + \alpha_1 k$, where the linear parameters α_0 and α_1 are determined by the joint constraints $s_{100} = 1$ and $\mu = 0.0277$.

The dots are the result of an iterative optimization procedure applied to a 100-

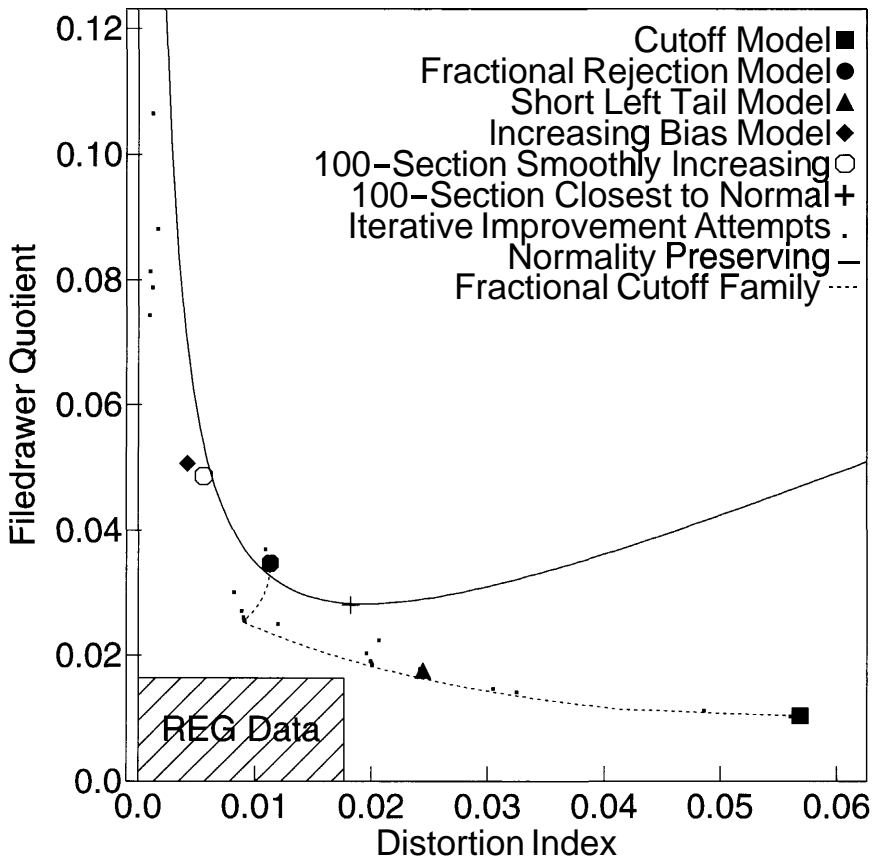


Fig. 4. Distortion vs. filedrawer tradeoff.

section model. At several starting points, including the smoothly increasing model, the normal model, and three models close to the different members of the fractional cutoff family, a gradient-following optimization algorithm was used to try to reduce D , reduce Q , or both. For each starting point one algorithm attempted to reduce D without regard to effects on Q , another minimized Q without regard to D , and others attempted to minimize linear combinations of Q and D with varying weights. These optimization algorithms halt when their attempt to follow a downward gradient produces an increase rather than a decrease in the target parameter, indicating that a local minimum lies within the algorithm's finest resolution for adjustments to the individual s_k .

It is notable that the scatter of dots at the upper left represent one model—the one minimizing D without regard to Q —from each of the five starting points. Similarly, several of the larger dots appearing along the "fractional cutoff family" dotted curve result from efforts starting at different points to minimize Q without regard to D . One, obscured by the large block showing the simple

cutoff model, essentially rediscovered that model to within the resolution of the 100-section parameterization. Other dots represent other local minima for the different linear combinations $\alpha D + \beta Q$ being minimized.

The endpoints of these optimization attempts, along with the parametric curve of the fractional cutoff model family, strongly indicate the existence of some limiting curve below which the distortion index and the filedrawer parameter cannot simultaneously be reduced. Moreover, these facts suggest that the fractional cutoff family curve is either on or very near that limit at least up to its inflection point, and that the limiting curve can be expected to have some smooth continuation into the region populated by the D-minimizing endpoints at the upper left. Finally, it is clear that the four one-parameter models examined in detail are all quite near the joint lower limit of distortion and filedrawer, and span much of the possible range of distortion values.

Given the failure of any selection model to occupy the shaded region statistically consistent with the observed REG data, even if the selection model is allowed to optimize 99 free parameters in an effort to achieve such consistency, it seems reasonable to conclude that the REG observations are inconsistent with any form of selection hypothesis whatever. A graph similar to Figure 4 could be drawn for the 100-trial runs; it would show even more dramatic inconsistencies due to the larger effect size and smaller filedrawer. The same cannot be said for the 1000-trial runs due to the poor statistical resolution resulting from their relatively small population. However, even if this subset is regarded as suspect due to the possibility of selection, the 50-trial and 100-trial runs between them produce an aggregate composite $Z = 3.9044$ (using the standard per-trial weighting), a result actually slightly stronger than that of the REG database as a whole.

Conclusions from Model Comparisons

On the basis of the statistical parameters of the data distribution, we are able to reject the data-selection models producing the strongest distortion of the source statistics as an explanation for the anomalous mean shift in the REG data (Table 5). When we examine models with lower distortion indices, we find that the distribution statistics become indistinguishable from a mean shift hypothesis, but the required rejection rate for these models is so large as to be completely incompatible with experimental records (Table 6).

It has been proven above that for selection models preserving normality of the output statistics, there is a tradeoff between the variance (the only free parameter once the mean is fixed) and the filedrawer; the more closely the selection process tries to preserve the original variance while imposing a nonzero mean shift, the more data it must discard. Figures 3 and 4 illustrate that the tradeoff between increasing filedrawer vs. increasing distortion appears to be a general feature of selection models, even when the number of adjustable model parameters is increased and when optimizing searches are made to try to

improve their performance in these features. If this generalization, supported by all currently available evidence, is valid, then the fact that every selection model considered is either strongly rejected by its distribution statistics, strongly rejected by its predicted filedrawer population, or both, means that no form of selection model can account for the data.

As noted above, the experimental protocol involved logging and recording all rejected or missing data along with the reasons for their rejection. Any experimenter being misled by bias into making an invalid decision to discard a dataset thus would leave a record of this act, even if the data values themselves went unrecorded. The only possibility for selection rates large enough to induce the effects therefore requires deliberate deception on the part of the experimenters, rather than simple bias. Experimenter fraud of this sort is frequently invoked as a last-resort accusation for explaining away anomalous results. A drawback of this "explanation" is that it is innately unfalsifiable: once it has been decided that a given experimenter is fraudulent, there is no reason to believe anything that experimenter says, nor any argument that the experimenter can make to refute the accusation. Perhaps more to the point, experimenters who set out to conduct a fraudulent experiment have far less labor-intensive ways to do so than carefully hiding a selected subset of the experimental data after they were generated.

A Subset of Special Interest

Aside from the general question of data selection in the REG experiment, there is a specific subset where the issue is of extra interest. As has been noted in the past (Dunne & Jahn, 1995; Dobyms & Nelson, 1998), the operator assigned to ID code 010 produced impressively large effect sizes in the early period of the experiment. This early period is distinguished from this operator's later data, not only by a temporal hiatus of over a year in which no data were generated, but also by a change in device (a portion of the hiatus, for this operator and all others, was caused by the delay of qualifying and calibrating the replacement REG machine), and by a change in protocol (it was decided during this period that all secondary parameters, such as volitional vs. instructed assignment of trial intentions, must be held constant throughout a series, rather than being variable on a session-by-session basis as had been the case previously). The early data for Operator 010 have an effect size more than an order of magnitude larger than any other database; they are statistically distinguishable not only from the vast bulk of other operator performances, but from the contemporaneous early data of other operators, and from the later performance of the same operator as well. The reasons for the larger effect are not clearly understood.

The distinctive character of this early 010 dataset, and its lack of explanation, mandate that all reasonable hypotheses for its outcome must be carefully scrutinized. Therefore, it seems appropriate to apply the formalism developed herein to the possibility that this extraordinary database was produced by deliberate deception on the part of the operator.

The formal data in this set comprise 503 runs, with a mean Z-score of 0.2556 (composite $Z=5.732$). All runs are in the 50-trial length. There are 38 void runs with values in this database; these voids have a mean value of -0.2037 (composite $Z=-1.256$, nonsignificant). Combining the voids with known values with the formal data would lower the effect size (mean Z-score) to 0.2233, with an associated composite Z-score of 5.194. As in the general database examination, the voids with known values have an apparent negative bias that is nevertheless well within the range of plausible chance variation; the change between effect size with and without the void data is within the statistical uncertainty of the measured effect size in the formal data, and the Z-score of the recombined data remains highly significant. We may conclude that, as in the general analysis, the voids with values do not appreciably impact the experimental conclusions, and they need not be considered further.

The database also contains 10 voids with no recorded values. Unfortunately, since we are here considering the possibility of deliberate concealment of data in addition to experimental bias, this does not give us a value for the actual filedrawer. Indeed, the actual filedrawer in a case of deliberate data selection is both unknown and unknowable. Our interpretation of theoretical filedrawer predictions of model fits must, rather, be based on the credibility of the operator's having managed to discard the requisite amount of data without detection by the experimenters—assuming a selection process that is consistent with the statistical parameters of the data.

Table 7 presents the standard deviation, skewness, and kurtosis of the actual database, as contrasted with the selection models developed previously. The filedrawer quotient Q required for the selection model to fit the observed mean shift is also reported.

An analysis of statistical power indicates that the most sensitive test parameter is the skewness, in the case of the fractional rejection model, and the standard deviation for all other models. In contrast to the examination of the general

TABLE 7
Model Comparisons for Special Subset

	σ	γ_3	γ_4	Q
Actual data	1.0402	0.1037	0.1538	NA
Cutoff	0.8100	0.5547	-0.0482	0.160
Z-score	-7.300	4.129	0.925	
Fractional rejection	0.9668	-0.2459	0.2698	0.320
Z-score	-2.329	-3.201	0.531	
Short tail	0.8467	0.2980	0.0639	0.191
Z-score	-6.136	1.779	-0.412	
Increasing bias	0.9668	-0.0697	0.0834	0.467
Z-score	-2.329	-1.587	-0.322	

database, we generate a $p < 0.05$ rejection of the selection model in every case. Moreover, those models for which the rejection is weakest ($p = 0.0013$ for fractional rejection, $p = 0.020$ for increasing bias, both two-tailed) predict large filedrawer quotients; generating these data by the increasing bias selection model would require discarding very nearly one run for every two which were recorded. While experimenter vigilance might be less than perfect, the redundant measures deployed to prevent operators from concealing the fact that they have generated data (the continuous hardcopy of the data is perhaps the most relevant to the current instance) make it difficult to credit that an operator could succeed in concealing one experimental run out of every three.

Final Summary

The space of possible selection processes is, essentially, the space of all functions of x bounded by $[0, 1]$. A set of four relatively simple one-parameter families of selection processes nevertheless allows some conclusions to be drawn. The cutoff model provably demonstrates the existence of a minimum level of discarded data for any target level of mean shift. The performance of numerous sectional selection models allowing multi-parameter optimization indicates a lower limit to the filedrawer for any given level of statistical distortion, and indicates further that the minimal filedrawer increases as the distortion decreases, and that the four one-parameter models are close to if not actually occupying this minimal limit.

For the effects in the database as a whole, the effect size is small enough that some selection models can produce statistics indistinguishable from those observed in the data. The possibility that the apparent effect was constructed by biased selection can be refuted, however, by comparing the actual population of discarded runs with the population required by those selection models that produce adequate fits to the data statistics. Surveying the space of optimized multi-parameter selection models confirms that the conclusions drawn from the single-parameter models can be generalized to all selection models with high confidence. Since failure to record the existence of a discarded run would require the deliberate circumvention of protocol rather than mere biases of judgment, the thesis that the anomalous effect as a whole could be due to unconscious selection of favorable data can be rejected.

For a database that has been regarded as suspect due to its origin in the exceptional performance of a single operator, all models examined predict statistics significantly different from those actually present. Even the least ill-fitting model requires a "filedrawer" of discarded data so large that its successful concealment from the experimenters by a malevolent operator becomes incredible. Since improving the statistical fit would require an even larger filedrawer quotient, the hypothesis that Operator 010 produced a spurious effect by concealing negative runs also can be regarded as refuted.

Appendix: Relative Statistical Power of Tests

It is common practice to compare distributions by using distribution tests such as χ^2 goodness-of-fit tests or Kolmogorov-Smirnov tests, to the extent that the use of a moment-based test may strike some readers as archaic. However, moment tests offer better statistical power than such distribution tests when a specific hypothesis regarding a moment value is available.

Since in the worst case of a selection effect we may be confronted with a normal distribution that merely has a smaller standard deviation than expected, the detection of reduced variance is used as a test case. Table A.1 presents, for a range of N , the probability of Type II error for an optimally sensitive χ^2 test, given a change in σ such that a simple variance test has $\alpha = \beta = 0.05$. In other words, each line of the table was computed by calculating, for that N , the value of σ that would produce a 5% chance of failing to be rejected by a $p < 0.05$ criterion on a variance test. The χ^2 tests were based on uniform-population binning; the number of bins was chosen by finding the bin number which minimized the Type II error probability β . The β values given in the table assume $\alpha = 0.05$, that is, that the χ^2 test will reject the null hypothesis on a $p < 0.05$ criterion.

Recalling that $\alpha = \beta = 0.05$ for the direct moment test on the variance, it is obvious that the χ^2 test has a much higher chance to overlook the same effect on the same data. It is notable that as N increases, the loss of performance in the χ^2 test becomes worse.

Similar considerations apply to the Kolmogorov-Smirnov test for distribution differences. For a simple demonstration, a Monte Carlo test was run by generating 10,000 sample distributions, each comprising 100 normal deviates with $\alpha = 0.767$. As noted above a simple variance test will reject such samples at the $p < 0.05$ level 95% of the time. The K-S test, in contrast, produced $p < 0.05$ rejection on only 2164 of the samples, indicating a Type II error probability of approximately 78%. It was considered redundant to extend this investigation to larger sample sizes.

TABLE A.1
Sensitivity of Optimal χ^2

N	σ	Bins	β
100	0.767	5	0.363
200	0.836	6	0.449
500	0.896	7	0.511
1000	0.926	7	0.535
5000	0.967	8	0.565
10,000	0.977	8	0.569

Acknowledgments

This work was supported by donations from the Fetzer Institute, the Institut für Grenzgebiete der Psychologie und Psychohygiene, and numerous private donors including Laurance Rockefeller, George Ohrstrom, and Donald Webster.

References

- Dobyns, Y. H., & Nelson, R. D. (1998). Empirical evidence against decision augmentation theory. *Journal of Scientific Exploration*, 12, 231-257.
- Dunne, B. J., & Jahn, R. G. (1995). *Consciousness and Anomalous Physical Phenomena*. Technical note PEAR 95004, May 1995.
- Gould, S. J. (1996). *The Mismeasure of Man*. Norton.
- Jahn, R. G., Dunne, B. J., & Nelson, R. D. (1987). Engineering anomalies research. *Journal of Scientific Exploration*, 1, 21-50.
- Jahn, R. G., Dunne, B. J., Nelson, R. D., Dobyns, Y. H., & Bradish, G. J. (1997). Correlations of random binary sequences with pre-stated operator intention: A review of a 12-year program. *Journal of Scientific Exploration*, 11, 345-367.

Comments from Mikel Aickin

The Dobyns article gives the impression that all reasonable subconscious ("unconscious" in his terms) distortions of the PEAR REG data through data selection have been ruled out. I believe this implication is untrue. I programmed the simulation of a very simple automated strategy, which is oriented toward moving the mean of the data, while arranging thin gs so that a statistical test of Normality would be passed. I used a better Normality test than Dobyns did, so that my simulation provides stronger evidence than his. In my simulations, I counted a success when I could produce the results cited by Dobyns for the data that he retained, without non-Normality being detected. I then recorded the percent of data that had to be deleted, among the successful cases. Here are the results for the data presented by Dobyns.

Type of Experiment	Reported % Data Deleted	Simulated chance of successfully distorting the data	Simulated % of data deleted among successful distortions
50-run	5.7%	1.00	1.0%
100-run	2.3%	0.09	0.5%
1000-run	3.7%	0.33	1.0%

It seems clear from this that it is possible to produce the PEAR REG results through data selection. This says almost nothing, of course, about whether any such selective distortions occurred. I would rely on the professional reputation and integrity of the PEAR investigators, which I regard as beyond question, for the validity of the data. Further, since Dobyns reports that, whether one includes or excludes the "void" data, the overall conclusions about the experiments are the same; it is not clear to me why any of this is of any importance.