

ALLAN FRANKLIN AND COLIN HOWSON*

WHY DO SCIENTISTS PREFER TO VARY THEIR EXPERIMENTS?

I.

IN GENERAL, before any experimental result is regarded as established, the experimental procedure is repeatedly instantiated in a variety of conditions: the instrumentation, apparatus, or analysis procedure is changed or new laboratory personnel are involved. If the variation is sufficiently marked, the experiment ceases to be a mere repetition and becomes a different experiment entirely. It has often been remarked that repeating the same experiment (with the same result) yields diminishing returns of information. This is easily enough explained, for after sufficiently many repetitions, in a variety of circumstances, our confidence that we are genuinely witnessing the phenomenon in question has grown sufficiently great that further tests become otiose.

A rather different question is why, when we have repeatedly performed a test on an hypothesis h and satisfied ourselves that the observed result e is a genuine experimental result,¹ where h together with some background theory T predict e , we should perform, where possible, an entirely different type of test on h , involving say an entirely different range of background theory T' . Suppose we do and obtain a result e' , where $h \wedge T' \vdash e'$. We then regard e' as providing more confirmation of h than would be provided by merely another repetition yielding e . Can we explain this? Yes, and very simply, using a Bayesian analysis of the situation. Let us suppose that we have a conditional probability function relativized to a stock of background information including T and T' . Suppose E and E' are two different experimental procedures, in the

*Department of Physics, University of Colorado, Boulder, CO 80309, U.S.A. and Dept. of Philosophy, London School of Economics, London WC2A 2AE. We are grateful to Dr. Michael Redhead for helpful discussions. Errors and confusions that remain are, of course, our own responsibility. This material is based on work partially supported by the National Science Foundation under Grant No. SES-8204074. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

¹For more general discussions of evidential support for hypotheses and theories see J. Dorling, 'Bayesian Personalism, the Methodology of Scientific Research Programmes, and Duhem's Problem', *Studies in History and Philosophy of Science* 10 (1979), 177; and M. Redhead, 'A Bayesian Reconstruction of the Methodology of Scientific Research Programmes', *Studies in History and Philosophy of Science* 11 (1980), 341; and P. Horwich, *Probability and Evidence*. (Cambridge: Cambridge University Press, 1982); and P. Orbach, 'On the Utility of Repeating the "Same" Experiment', *Australasian Journal of Philosophy* 59 (1981), 151.

sense discussed below, which can test h , each capable (in principle) of being indefinitely instantiated and yielding respectively, $e_1, e_2, \dots, e_n, \dots; e'_1, e'_2, \dots, e'_m, \dots$, where $P(e_i|h) = P(e'_j|h) = 1, i, j, \geq 1$. The fact that E and E' are *different* experimental tests of h would seem to indicate that the results of using E' are less than maximally correlated with those of E and vice versa. We can make this precise by defining E and E' to be distinct experimental procedures iff, for all $m > m_0$, for some m_0 .

$$P(e_{m+1}|e_1 \wedge e_2 \wedge \dots \wedge e_m) > P(e'_{m+1}|e_1 \wedge e_2 \wedge \dots \wedge e_m) \quad (1)$$

and for all $n > n_0$, for some n_0

$$P(e'_{n+1}|e'_1 \wedge e'_2 \wedge \dots \wedge e'_n) > P(e_{n+1}|e'_1 \wedge e'_2 \wedge \dots \wedge e'_n) \quad (2)$$

where P represents the belief structure of the 'ideal' experimenter.

The result we have mentioned, namely that h receives *more* support from a mixture of confirming results relative to E and E' than from E or E' alone is now simple to prove. For

$$\begin{aligned} \frac{P(h|e_1 \wedge \dots \wedge e_k \wedge e'_{k+1})}{P(h|e_1 \wedge \dots \wedge e_k)} &= \frac{P(e_1 \wedge \dots \wedge e_k)}{P(e_1 \wedge \dots \wedge e_k \wedge e'_{k+1})} \\ &= \frac{1}{P(e'_{k+1}|e_1 \wedge \dots \wedge e_k)}. \end{aligned}$$

Similarly

$$\frac{P(h|e_1 \wedge \dots \wedge e_k \wedge e_{k+1})}{P(h|e_1 \wedge \dots \wedge e_k)} = \frac{1}{P(e_{k+1}|e_1 \wedge \dots \wedge e_k)}.$$

Hence h receives proportionally *more* support at the $k + 1$ th trial from the experiment E' than from simply repeating E if and only if $P(e'_{k+1}|e_1 \wedge \dots \wedge e_k) < P(e_{k+1}|e_1 \wedge \dots \wedge e_k)$. But this, as we have seen, is precisely the case for all $k > \text{some } k_0$, where E and E' are less than perfectly correlated experimental procedures. And being less than perfectly correlated (in their outcomes) is a good formal explication of what one means by saying that E and E' are different experimental tests of h .

II.

In the previous section we demonstrated the happy conclusion that two 'different' experimental procedures provide more evidence in support of an

hypothesis than two repetitions of the 'same' experiment. In this section we will discuss some of the conditions under which it is reasonable to make such probability assignments, or in other words when we would consider the results of using E and E' less than maximally correlated. The following section will provide evidence, taken from the history of twentieth century physics, to show that such assignments are implicitly, and sometimes explicitly, made by practicing scientists under the conditions we describe.

In this discussion we shall assume that the experiments referred to test the same hypothesis and measure the same quantity. It is, of course, true that the same hypothesis may be tested by measuring two different quantities, as for example the hypothesis 'CP is violated in K_L^0 decays' may be tested by measuring either the rate of $K_L^0 \rightarrow \pi^+ \pi^-$ decays or by measuring the asymmetry in the decays $K_L^0 \rightarrow e^\pm \pi^\mp \nu$ or $K_L^0 \rightarrow \mu^\pm \pi^\mp \nu$. This may be dealt with by further restricting the hypothesis to two separate hypotheses: 'CP is violated in K_L^0 decay into pions' and 'CP is violated in the decay $K_L^0 \rightarrow \mu^\pm \pi^\mp \nu$.' Even without such a restriction these experiments will be 'different' in the sense discussed below. An interesting example of two such 'different' experiments which used the same physical apparatus simultaneously will be discussed in the next section.

Why should two experimental procedures be considered 'different' if the experimental apparatuses are 'different'? That this is not question begging or the beginning of infinite regress will be made clear below. We begin with two idealized and extreme categories of 'different' experimental apparatuses. The first category consists of two apparatuses, A_1 and A_2 , whose proper operation is implied by a single theory, T . These apparatuses may differ in size, material, geometrical arrangement, number of particular pieces of experimental equipment, analysis procedure, etc. The second category consists of two apparatuses, B_1 and B_2 , which depend exclusively on two distinct theories, T and T' , respectively. Such theories may be distinguished by examining the set of statements implied by each of them. An example of this is the bubble chamber, which depends on the theory of bubble formation in superheated liquids and the spark chamber, which depends on the theory of electrical discharges in ionized gases. In practice, two experimental apparatuses may consist of parts which depend on the same theory and parts which depend on different theories. Such hybrid cases may be dealt with by considering the parts separately, using the arguments given below. In the following discussion we shall assume that our conditional probability functions are relativized to a store of background knowledge which includes T and T' , as well as what one might call super theories such as quantum mechanics and special relativity, which are believed to apply to all of nature. The discussion is clearer, however, if one restricts oneself to the specific theories of the apparatus.

Let us first consider a single apparatus A_1 , which depends on a single theory

T and which is operated at two different times, t_1 and t_2 . Then unless some explicit time asymmetry is presupposed, a uniformity of nature assumption is made which implies that $P(e_i|e_j) = 1$, where e_i and e_j are similar results obtained at different times i, j , using A_1 , and thus that $P(e_i \wedge e_j) = P(e_i)$. This assumption is, in fact, made for the vast majority of experiments, for no experiment acquires all of its data instantaneously and there seems to be no useful distinction to be made concerning the time interval between data acquisition runs, or whether or not the apparatus was turned off between these runs. In practice any well-designed experiment will include checks to assure the proper and constant operation of the apparatus. We now consider two apparatuses, A_1 and A_2 , which depend on a single theory T and for which the uniformity of nature assumption is made. We assert here that $P(e|e') < 1$ where e, e' are outcomes of A_1 and A_2 respectively. This is plausible because the successful operation of A_1 will not, in general, guarantee the successful operation of A_2 .² Thus $P(e \wedge e') < P(e)$ or $P(e')$. From our previous discussion it then follows trivially that A_1 and A_2 give rise to different experimental procedures and hence are more informative, if successful, than one, and better than two Heraclitean or identical repetitions of a single experiment. The argument for two apparatuses, dependent respectively on two distinct theories, T and T' , now goes through *a fortiori*.

Our discussion suggests that only identical, or Heraclitean, repetitions of a single experiment may be considered the same. Experiments involving different theories are usually thought of as giving rise to stronger inequalities, $P(e|e') \ll 1$ than those involving the same theory. We have not attempted here to give any numerical estimates of these probabilities or to deal with the question of how much support is given to an hypothesis by a given experimental result. This is a complex problem involving not only the precision of the result but also how much confidence one has that the apparatus is actually measuring the quantity involved, *cf.* the Galilean telescope and Aristotelian natural philosophers.³

In addition, the degree of 'difference' between two experimental results may very well depend on the theoretical context at a given time. To illustrate this let us consider an ahistorical and idealized account of experimental tests of the Newtonian addition of velocities law prior to 1905, when classical mechanics was the only theory, and after 1905, when Einstein's special theory of relativity emerged as a serious rival. Prior to 1905 there would have been no reason to suspect that testing this law at speeds close to c , the speed of light, would have been very 'different' than tests at low speeds, because classical mechanics made no such distinction. After 1905, when special relativity gave a velocity addition

²A case in point was the construction of the Dubna 10 GeV synchrotron which was a scaled up version of the 3 GeV Brookhaven Cosmotron. The Dubna machine never worked as well as it was supposed to because certain difficulties which were negligible at 3 GeV became important at 10 GeV.

³For discussion of some of these questions see Redhead, Horwich and Dorling, Note 1, and C. Glymour, *Theory and Evidence* (Princeton: Princeton University Press, 1980).

law, which differed significantly from the Newtonian one at speeds close to c , we would think that experiments at high and low speeds would be quite 'different'. In fact, it seems quite plausible, that had such a high velocity test given results in agreement with classical mechanics, then classical mechanics would have been more strongly confirmed after 1905 than before.

Our model can accommodate this quite easily. Let e_1, \dots, e_n be experimental results at low speeds in agreement with classical mechanics, and let e'_{n+1} be a high speed result in agreement with classical mechanics. It seems clear that prior to 1905 $P(e'_{n+1}|e_1 \wedge \dots \wedge e_n) \approx P(e_{n+1}|e_1 \wedge \dots \wedge e_n)$, and thus such a high velocity test provides little additional confirmation, as compared to a repetition at low speeds.

Let us now consider the situation after 1905. The background knowledge (relative to which our probabilities are defined) has now loosened up a bit, because special relativity has now emerged as a serious rival to classical mechanics. Let CM stand for classical mechanics and SR for special relativity. Further define $P^*(\phi) = P(\phi|e_1 \wedge \dots \wedge e_n)$ and let $P^*(\text{CM}) = p$ and $P^*(\text{SR}) = q$, where p, q are both considerably greater than 0.

$$P^*(e'_{n+1}) = P^*(e'_{n+1}|\text{CM})P^*(\text{CM}) + P^*(e'_{n+1}|\text{SR})P^*(\text{SR}) + \varepsilon_1$$

where ε_1 is quite small. (This appears to be justified historically.) But $P^*(e'_{n+1}|\text{SR}) = 0$ because special relativity predicts a very different result at speeds close to c than does classical mechanics.

Thus $P^*(e'_{n+1}) \approx P^*(\text{CM}) = p$, since $P^*(e'_{n+1}|\text{CM}) = 1$ or *i.e.* $P(e'_{n+1}|e_1 \wedge \dots \wedge e_n) \approx p$.

Now let e_{n+1} be a result in agreement with the Newtonian law at low speeds. At such speeds both classical mechanics and special relativity predict the same results, within observational error. Thus we have, where ε_2 is also small in comparison with p and q ,

$$\begin{aligned} P^*(e_{n+1}) &= P^*(e_{n+1}|\text{CM})P^*(\text{CM}) + P^*(e_{n+1}|\text{SR})P^*(\text{SR}) + \varepsilon_2 \\ &\approx 1 \times P^*(\text{CM}) + 1 \times P^*(\text{SR}) \\ &\approx p + q \end{aligned}$$

i.e. $P(e_{n+1}|e_1 \wedge \dots \wedge e_n) \approx p + q$.

Hence $P(e'_{n+1}|e_1 \wedge \dots \wedge e_n) < P(e_{n+1}|e_1 \wedge \dots \wedge e_n)$. Thus we see that our model indicates that the emergence of rivals to an hypothesis h may very well determine what is or is not merely a repeated test of h .⁴ We note also that these experiments will differ in respect of experimental apparatus as discussed above.

⁴P. Horwich, Note 1, p. 118.

III.

In this section we present evidence, taken from the history of twentieth century physics, to support the view that practicing physicists do, in fact, use the procedures for classifying experiments as the 'same' or 'different', outlined in the previous sections.

Perhaps the clearest examples of almost Heraclitean repetitions (we neglect here different data runs with a single apparatus) are the multi-group bubble chamber collaborations common in the 1960's.⁵ In these experiments a set of bubble chamber photographs were taken under presumably identical conditions. In order to analyze the data more quickly, portions of these sets were then given to several groups of physicists at different locations. The results were then combined in a single report. In these cases track measurements were made on similar measuring machines and the track reconstruction and event fitting would be done with standard computer programs (e.g. TVGP, Three View Geometrical Program and SQUAW, an event fitting program). Although some subjective judgements were made, for example track identification by ionization estimates, these were regarded as insignificant differences. The data were combined at one location and the paper written. Copies of the paper were circulated to members of the groups for comment and a final version agreed upon. This led to the somewhat odd result that papers were published whose co-authors may not have ever met. The existence of such combined reports illustrates not only the assumption of the uniformity of nature in time, [e.g. $P(e_i \wedge e_j) = P(e_i) = P(e_j)$], but also perhaps the assumption of the uniformity of physicists.

A second rather interesting example of experiments which differ only slightly is provided by the Lunar Laser Ranging experiments.⁶ In this case identical sets of data were analyzed independently by two groups. The reasons for such analysis are stated quite clearly by the Committee on Gravitational Physics:

[This] leads to computer programs with upwards of 100,000 Fortran statements. Verification of the reliability of such enormous programs is practically impossible. It would appear that only two sets of software, each developed completely independently, can serve as an adequate check on reliability.⁷

Thus independent analyses (or 'different' experiments) are regarded as providing more evidential weight than a single analysis. It is interesting to note, however,

⁵For an example of one such collaboration see J. Lynch *et al.*, 'Peculiarities Observed in the Reaction $\pi p \rightarrow K n$ ', *Phys. Lett.* **35B** (1971), 457.

⁶J. G. Williams *et al.*, 'New Test of the Equivalence Principle from Lunar Laser Ranging', *Phys. Rev. Lett.* **36** (1976), 551; I. I. Shapiro, C. C. Counselman, and R. W. King, 'Verification of the Principle of Equivalence from Massive Bodies', *Phys. Rev. Lett.* **36** (1976), 555.

⁷Committee on Gravitational Physics, Space Science Board, *Strategy for Space Research in Gravitational Physics in the 1980's* (Washington, D.C.: National Academy Press, 1981), p. 44.

that prior to publication of the results, but before all checks had been made, there was a discrepancy between the two results:

A valuable cross check for model or software errors has been provided by an MIT – AFCRL analysis of the lunar-ranging data with separately developed software. Both analyses presently give null results. However, up until the time when the MIT – AFCRL reported their null result to us, our solutions erroneously indicated a 1m amplitude for the Nordtvedt term. The error was traced to our truncation of some apparently small relativistic terms in the equations of motion. We planned to add all remaining relativistic terms of order $1/c^2$ before publication of our results, but we had not expected that these additions would affect the Nordtvedt term significantly.⁸

Even if the planned check had not been made the discrepancy between the two results would no doubt have uncovered the error.

A similar case has been presented by Aronson *et al.*⁹ in discussing the possible energy dependence of η_{+-} , the CP violating parameter in $K_L^0 \rightarrow \pi^+ \pi^-$ decay. Here the same data were analyzed in two different ways by the same group of physicists. That these two independent analyses (different experiments in our sense) were regarded as providing more evidence in favor of the hypothesis of energy dependence than merely rechecking one analysis, is shown in their statement:

The results obtained by Methods A and B are in good qualitative agreement, the principal difference being that Method B gives slope parameters which are somewhat smaller in magnitude and also in statistical significance. *Notwithstanding the agreement between the results of these two different methods* (emphasis added), the possibility remains that the apparent energy dependence of the $K^0 - \bar{K}^0$ parameters is due to some unknown systematic effect in the data.¹⁰

Two ‘different’ experiments may even be performed using the same apparatus at the same time. A case in point is the experiments testing CP violation in K_L^0 decays by either observing $K_L^0 \rightarrow \pi^+ \pi^-$ ¹¹ or by observing an asymmetry in $K_L^0 \rightarrow u^\pm \pi^\mp \nu$ decay.¹² Here the apparatus guaranteed that two charged particles from K_L^0 decay traversed a magnetic spectrometer. A muon filter, consisting of one meter of lead and two banks of scintillation counters was located behind the spectrometer. Only muons could penetrate the lead and count in both counter

⁸J. G. Williams *et al.*, Note 6, p. 553.

⁹S. H. Aronson *et al.*, ‘Determination of the Fundamental Parameters of the $K^0 - \bar{K}^0$ System in the Energy Range 30 – 110 GeV’, *Phys. Rev. Lett.*, **48** (1982), 1306.

¹⁰Note 9, p. 1308.

¹¹R. Messner *et al.*, ‘New Measurement of the $K_L^0 \rightarrow \pi^+ \pi^-$ Branching Ratio’, *Phys. Rev. Lett.*, **30** (1973), 876.

¹²R. Piccioni *et al.*, ‘Measurement of the Charge Asymmetry in the Decay $K_L^0 \rightarrow \pi^\pm u^\mp \nu$ ’, *Phys. Rev. Lett.*, **29** (1972), 1412.

banks. The information as to whether or not these banks had fired was recorded for each event. In looking for $K_L^0 \rightarrow \pi^+ \pi^-$, only those events in which the counter banks had not fired, those without a muon, were used, although further analysis was required to identify $K_L^0 \rightarrow \pi^+ \pi^-$. To look for an asymmetry in $K_L^0 \rightarrow \mu^\pm \pi^\mp \nu$ decay, only events which contained a muon, *i.e.* the counter banks had fired, were used. These experiments could, of course, have been performed separately but that would have wasted valuable beam time.

The examples above refer to 'different' experiments in the sense that different analysis procedures or information were used. We now turn to cases which illustrate our remarks concerning the theoretical basis of the experimental apparatus. This discussion will use evidence from the *Review of Particle Properties*,¹³ which is generally recognized as the standard source of information on particle properties for the physics community. We see clearly the importance of type of apparatus used in a measurement by the fact that it is given for each measurement of a quantity. Table 1, (Ref. 13, p. S61), gives a partial list of measurement techniques used (specific devices such as the DESY Pluto detector are omitted). These may depend on different theories of operation, *i.e.* counters, spark chambers, emulsions and bubble chambers. Within these groups we see differences in recording techniques, *i.e.* optical spark chambers where the data is recorded on film as compared with automatic or wire spark chambers where the information is recorded electronically, or in material used, *i.e.* deuterium, freon, hydrogen, helium, heavy liquid, propane, or xenon bubble chambers.

Table 1. Measurement techniques

ASPK	Automatic spark chambers
CC	Cloud chamber
CNTR	Counters
DBC	Deuterium bubble chamber
ELEC	Electronic combination
EMUL	Emulsions
FBC	Freon bubble chamber
HBC	Hydrogen bubble chamber
HEBC	Helium bubble chamber
HLBC	Heavy liquid bubble chamber
HYBR	Hybrid: Bubble chamber and electronics
MMS	Missing mass spectrometer
OSPK	Optical spark chamber
PBC	Propane bubble chamber
PLAS	Plastic detector
SPEC	Spectrometer
SPRK	Spark chamber
STRC	Streamer chamber
WIRE	Wire chamber
XEBC	Xenon bubble chamber

¹³Particle Data Group, 'Review of Particle Properties', *Rev. Mod. Phys.* 52 (1980), S1.

Table 2. Mean life difference $\pi^+ - \pi^-$ /average ($\times 10^{-1}$)

Value	Error	Measurement technique
0.23	0.40	CNTR
0.4	0.7	CNTR
-0.14	0.29	CNTR
0.055	0.071	CNTR
Average 0.053	0.068	

Table 3. Charged pion mean life (units 10^{-9} s)

Value	Error	Measurement technique	Charge
25.6	0.8	CNTR	
25.46	0.32	CNTR	+
26.02	0.04	CNTR	+
25.6	0.3	CNTR	
25.9	0.3	CNTR	
26.67	0.24	CNTR	
26.04	0.05	CNTR	+
26.02	0.02	CNTR	+ -
26.09	0.08	CNTR	+
Average 26.030	0.023		

Tables 2 – 4 provide examples of measurements where the apparatus used, in this case scintillation counters, depends on a single theory, but in which increasing specificity and differentiation are noted. Table 2 (Ref. 13, p. S68) gives the measurements of the percentage difference between the mean lives of the π^+ and π^- mesons divided by their average. Although scintillation counters were used for all the measurements the geometry, beam conditions, *etc.* were different. In Table 3 (Ref. 13, p. S69), measurements of the mean life of the charged π mesons are given. Counters were again used, but now, in the light of possible, although not observed, differences in lifetime between the two charges, the charge of the meson beam used is now shown. A further difference is shown in Table 4 (Ref. 13, p. S65) which lists measurements of the ratio of muon to proton magnetic moments. These are all counter experiments but now not only is the charge of the muon given but the specific phenomenon used in the measurement is also shown *i.e.* spin resonance, phase of precession, hyperfine splitting.

Measurements using apparatuses which depend on different theories are illustrated in Table 5 (Ref. 13, p. S71), giving the values of the left – right asymmetry in $\eta \rightarrow \pi^+ \pi^- \pi^0$. The apparatus used were bubble chambers, both hydrogen and deuterium and spark chambers, optical and automatic.

It is, however, true that experimental apparatuses thought to be the same, or at least very similar, because a single theory was thought to govern their

Table 4. Muon to proton magnetic moment ratio

Value	Error	Measurement technique	Charge	Comments
3.1865	.0022	CNTR	+	Spin resonance
3.1830	.0011	CNTR	+	Precession strobe
3.176	.013	CNTR	-	Precession strobe
3.1834	.0002	CNTR	+	Precession phase
3.18336	.00007	CNTR	+	Precession strobe
3.1808	.0004	CNTR	-	Precession strobe
3.18338	.00004	CNTR	+	Precession phase
3.183351	.000016	CNTR		Hyperfine splitting
3.183314	.000034	CNTR		Hyperfine splitting
3.183330	.000044	CNTR	+	Precession phase
3.183347	.000009	CNTR	+	Precession phase
3.183336	.000013	CNTR		Hyperfine splitting
3.183349	.000015	CNTR		Hyperfine splitting
3.183326	.000013	CNTR		Hyperfine splitting
3.1833467	.0000082	CNTR	+	Precession phase
3.1833299	.0000025	CNTR		
3.1833403	.0000044	CNTR	+	Hyperfine splitting
3.1833448	.0000029	CNTR	+	Precession strobe
Average				
3.1833371	.0000039			

Table 5. Left-right asymmetry parameter for $\eta \rightarrow \pi^+ \pi^- \pi^0$ ($\times 10^{-2}$)

Number of events	Value	Error	Measurement technique
1351	7.2	2.8	DBC
1300	5.8	3.4	HBC
10,665	0.3	1.0	OSPK
705	-6.1	4.0	HBC
36,800	1.5	0.5	ASPK
10,709	.3	1.1	OSPK
1138	-1.4	3.0	HBC
349	3.2	5.4	DBC
220,000	-0.05	0.22	ASPK
165,000	0.28	0.26	OSPK
Average	0.12	0.17	

operation, have turned out to differ in significant ways. A case in point is the experiments on the double scattering of electrons in the 1920's and thereafter.¹⁴ The electrons used in these experiments came from both β -decay and thermionic sources and physicists of the day believed that the type of source used made no essential difference.¹⁵ Later work, however, showed that β -decay electrons

¹⁴For a detailed history of these experiments see A. Franklin, 'The Discovery and Nondiscovery of Parity Nonconservation', *Stud. Hist. Phil. Sci.* 10 (1979), 201.

¹⁵R. Cox, private communication. Professor Cox performed experiments during this period using both types of source.

Table 6. η_{+-} , CP Violating parameter
($\times 10^{-3}$)

Value	Error	Measurement technique
1.95	0.20	OSPK
1.99	0.16	OSPK
1.92	0.13	OSPK
1.95	0.04	OSPK
2.00	0.09	OSPK
1.94	0.08	OSPK
Average		
1.95	0.03	
2.23	0.05	ASPK
2.30	0.035	ASPK
2.25	0.05	SPEC
2.27	0.12	ASPK
Average		
2.273	0.022	

were longitudinally polarized while thermionic electrons were unpolarized, which resulted in significantly different experimental results. In this same episode, another difference, thought to be insignificant at the time, was also shown later to be important. This involved the question of whether the electron beam scattered from the front surface of the foil targets (a reflection experiment) or passed through the foils (a transmission experiment).¹⁶ At the time almost all experiments were of a reflection type which avoided multiple scattering problems. Later work, in which only the reflection and transmission character of the experiment was changed, showed a clear difference in the results. In fact, the reflection type of apparatus masked an important experimental effect, expected on the basis of a theoretical calculation, which was observed only in transmission experiments after the significance of the difference was realized. The fact that scientists, being fallible, are occasionally wrong in their judgements about significant differences between experiments does not argue against our model presented above. It shows only that differences may be hard to find.

An interesting illustration of 'different' experiments is shown in the current anomaly concerning the value of η_{+-} , the CP violating parameter in $K_L^0 \rightarrow \pi^+ \pi^-$ decay. Prior to 1973, η_{+-} had been measured six times. The results were in good statistical agreement and had a mean of $(1.95 \pm 0.03) \times 10^{-3}$ (See Table 6). Since 1973 there have been four additional measurements, which have a mean of $(2.27 \pm 0.022) \times 10^{-3}$. As shown in Table 6, the first set of measurements all used optical spark chambers, while the second set each used automatic spark chambers (the spectrometer listed included such chambers). These sets are

¹⁶Ref. 14, p. 246.

'different' in the sense discussed earlier, a point noted in the 1976 *Review of Particle Properties*:

The newer experiments are *in principle superior* [emphasis added, also read different] (higher statistics, better acceptance, easier trigger conditions).¹⁷ The large discrepancy between the two sets of measurements is still unresolved and unexplained.

The examples given above indicate that the physics community seems to follow procedures in agreement with our categorization of 'same' and 'different' experiments. While we do not wish to argue that all institutionalized behavior of scientists has methodological justification, we do believe that this evidence does reinforce our view. The examples have been taken from the *Review of Particle Properties*, the standard reference for physicists, as well as from leading research journals. If the accepted 'good' practice of science does not provide clues as to what 'good' science should be, then what will? It seems doubtful that a priori reasoning will suffice. After all, it is generally agreed that science progresses. It seems unlikely that a random walk of good and bad methodological decisions would lead to a science which progresses, at least most of the time. We have also presented reasonable, independent justification for our model. In addition, our view leads quite simply to the conclusion, already intuitively appealing, that two 'different' experiments provide more support for an hypothesis, than two repetitions of the 'same' experiment.

¹⁷Particle Data Group, 'Review of Particle Properties', *Rev. Mod. Phys.*, **48** (1976), S1.