0039-3681(94)E0006-T

# How to Avoid the Experimenters' Regress

*Allan Franklin**

## 1. Collins and the Experimenters' Regress

HARRY COLLINS is well known for both his skepticism concerning experimental results and evidence and for what he calls the 'experimenters' regress', the view that a correct outcome is one obtained with a good experimental apparatus, whereas a good experimental apparatus is one that gives the correct outcome. He has expressed this view at length in *Changing Order* (Collins, 1985).

He illustrates these views with his history of the early attempts to detect gravitational radiation, or gravity waves. He argues that the decision between the claimed observation of gravitational waves by Weber and the failure to detect them in six other experiments could not be made on reasonable or rational grounds. This results from the fact that one cannot legitimately regard the subsequent experiments as replications[1] and that one cannot provide independent reasons for belief in either result. He argues that we cannot be sure that we can actually build a gravity wave detector and that we might have been

*Department of Physics, University of Colorado at Boulder, Campus Box 390, Boulder, CO 80309-0390, U.S.A.

*Received 7 February* 1993; *in revised form* 23 *June* 1993.

[1]Collins offers two arguments concerning the difficulty, if not the virtual impossibility of replication. The first is philosophical. What does it mean to replicate an experiment? In what way is the replication similar to the original experiment? A rough and ready answer is that the replication measures the same physical quantity. Whether or not it, in fact, does so can, I believe, be argued for on reasonable grounds, as discussed below.

Collins's second argument is pragmatic. This is the fact that in practice it is often difficult to get an experimental apparatus, even one known to be similar to another, to work properly. Collins illustrates this with his account of Harrison's attempts to construct two versions of a TEA laser (Transverse Excited Atmospheric) (Collins, 1985, pp. 51–78). Despite the fact that Harrison had previous experience with such lasers, and had excellent contacts with experts in the field, he had great difficulty in building the lasers. Hence the difficulty of replication.

Ultimately Harrison found errors in his apparatus and once these were corrected the lasers operated properly. As Collins admits, '. . . in the case of the TEA laser the circle was readily broken. The ability of the laser to vaporize concrete, or whatever, comprised a universally agreed criterion of experimental quality. There was never any doubt that the laser ought to be able to work and never any doubt about when one was working and when it was not (*ibid.* p. 84).'

Although Collins seems to regard Harrison's problems with replication as casting light on the episode of gravity waves, as support for the experimenters' regress, and as casting doubt on experimental evidence in general, it really doesn't work. As Collins admits (see quote in last paragraph), the replication was clearly demonstrable. One may wonder what role Collins thinks this episode plays in his argument.

Pergamon

*Stud. Hist. Phil. Sci.*, Vol. 25, No. 3, pp. 463–491, 1994.

0039-3681/94 $7.00+0.00

fooled into thinking we had the recipe for constructing one, and that 'we will have no idea whether we can do it until we try to see if we obtain the correct outcome. *But what is the correct outcome* (emphasis in original)?'

> What the correct outcome is depends upon whether or not there are gravity waves hitting the Earth in detectable fluxes. To find this out we must build a good gravity wave detector and have a look. But we won't know if we have built a good detector until we have tried it and obtained the correct outcome! But we don't know what the correct outcome is until . . . and so on ad infinitum.
> The existence of this circle, which I call the 'experimenters' regress', comprises the central argument of this book. Experimental work can only be used as a test if some way is found to break into the circle. The experimenters' regress did not make itself apparent in the last chapter because in the case of the TEA-laser the circle was readily broken. The ability of the laser to vaporize concrete, or whatever, comprised a universally agreed criterion of experiment quality. There was never any doubt that the laser ought to be able to work and never any doubt about when one was working and when it was not. Where such a clear criterion is not available, the experimenters' regress can only be avoided by finding some other means of defining the quality of an experiment; a criterion must be found which is independent of the experiment itself (Collins, 1985, p. 84).

More succinctly, 'Proper working of the apparatus, parts of the apparatus and the experimenter are defined by the ability to take part in producing the proper experimental outcome. Other indicators cannot be found (*ibid.* p. 74).'

Collins argues that there are no formal criteria that one can apply to decide whether or not an experimental apparatus is working properly. In particular, Collins argues that calibration of an experimental apparatus cannot provide such a criterion.

> Calibration is the use of a surrogate signal to standardize an instrument. The use of calibration depends on the assumption of near identity of effect between the surrogate signal and the unknown signal that is to be measured (detected) with the instrument. Usually this assumption is too trivial to be noticed. In controversial cases, where calibration is used to determine relative sensitivities of competing instruments, the assumption may be brought into question. Calibration can only be performed provided this assumption is not questioned too deeply (*ibid.* p. 105).

In Collins's view the regress is broken by negotiation within the appropriate scientific community, which does not involve what we might call epistemological criteria, or reasoned judgment. Thus, the regress raises serious questions concerning both experimental evidence and its use in the evaluation of scientific hypotheses and theories. If no way out of the regress can be found then he has a point.

In this paper I will examine Collins's account of the first attempts to detect gravitational radiation. I will then present my own account of the episode, which differs substantially from his, and argue that his account is misleading and provides no grounds for belief in the experimenters' regress. I will show that calibration, although an important component of the decision, was not decisive in this case precisely because the experiments used a new type of apparatus to try to detect a hitherto unobserved phenomenon, and that the case of gravity wave detection is not at all typical of scientific experiments. I will also argue that the regress was broken by reasoned argument.

Before I begin, I would like to address an important methodological difference between Collins's account and my own. Collins bases his account of the episode almost entirely on interviews with some of the scientists involved. They are not named and are identified only by letter. My own account is based on the published literature. A supporter of Collins might argue that the published record gives a sanitized version of the actual history,[2] and that what scientists actually believed is contained in the interviews. I suggest that the interviews do not, in fact, show the scientists' consideration of the issues raised by the discordant results, and that these considerations are contained in the published record. In this particular episode, we have a published discussion among the participants, in which they explicitly addressed the issues as well as each other's arguments. I see no reason to give priority to off-the-cuff comments made to an interviewer, and to reject the accounts that scientists wished to have made as part of the permanent record.[3] There is no reason to assume that because arguments are presented publicly that they are not valid, or that the scientists did not actually believe them. There are, in fact, good reasons to believe that these are the arguments believed by the scientists. After all, a scientist's reputation for good work is based primarily on the published record, and it seems reasonable that they would present their strongest arguments there.[4] In addition, although Collins presents evidence that the various arguments were weighted differently by different scientists, the arguments presented were, in fact, the same as those given in publications. Neither does Collins's account demonstrate that the decision was based on anything other than the combined evidential weight of these arguments. As we shall see, there was considerable interchange between Weber and his critics, and that criticisms were offered by others, answered by Weber, and these answers were themselves

[2]Trevor Pinch recently remarked that an account based only on publications was 'bloodless' (private communication).

[3]Michael Lynch (Lynch, 1991) has, in a somewhat different case, argued that what scientists said when they were recording their data has more importance in evaluating their experimental claims than is their published considerations. This conflates data and experimental results. For a discussion of the general issue see Bogen and Woodward (1988) and for discussion of this specific case see Franklin (1993b).

[4]Someone might object that the scientist is merely putting their best foot forward, and that the public arguments are not those they actually believed. I don't believe this to be the case, and Collins has certainly not presented any evidence to support this view. I have presented evidence that, at least in one case, the arguments offered in private were the same as those offered publicly. In the case of the Fifth Force, a modification of the law of gravity, I have examined the private e-mail correspondence between the proposers of the hypothesis and compared it with the published record. There is no difference in the arguments offered. See Franklin (1993a), pp. 35–48.

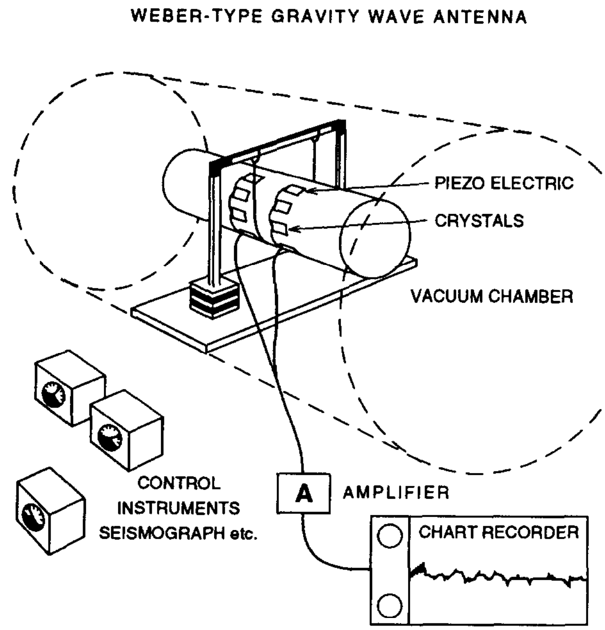


*Fig. 1. A Weber-type gravity wave detector. From Collins (1985), permission requested, but no reply.*

evaluated. The published record indicates that the decision was based on a reasoned evaluation of the evidence.

Let us now consider in detail Collins's discussion of gravity wave detectors.

## 2. Collins's Account of Gravity Wave Detectors

Collins illustrates the experimenters' regress and his skepticism concerning experimental results with the early history of gravity wave detectors.[5] He begins with a discussion of the original, and later to become a standard, apparatus developed by Joseph Weber (Fig. 1). Weber used a massive aluminum alloy bar,[6] or antenna, which was supposed to oscillate when struck by gravitational radiation.[7] The oscillation was to be detected by observing the amplified signal from piezo-electric crystals attached to the antenna. The expected signals were quite small (the gravitational force is quite weak in comparison to electromagnetic force) and the bar had to be insulated from other sources of noise such as electrical, magnetic, thermal, acoustic, and seismic forces. Because the bar was at a temperature different from absolute zero, thermal noise could not be

[5]As discussed earlier, one cannot examine Collins's sources in any detail. Collins uses interviews almost exclusively, and to maintain anonymity he refers to scientists by a letter only. In addition there are no references given to any of the published scientific papers involved, not even to those of Weber.

[6]This device is often referred to as a Weber bar.

[7]Gravitational radiation is produced when a mass is accelerated.

avoided, and to minimize its effect Weber set a threshold for pulse acceptance. Weber claimed to have observed above-threshold pulses, in excess of those expected from thermal noise.[8] In 1969, Weber claimed to have detected approximately seven pulses/day due to gravitational radiation.

The problem was that Weber's reported rate was far greater than that expected from calculations of cosmic events (by a factor of more than 1000), and his early claims were met with skepticism. During the late 1960s and early 1970s, however, Weber introduced several modifications and improvements that increased the credibility of his results. He claimed that above-threshold peaks had been observed simultaneously in two detectors separated by 1000 miles. Such coincidences were extremely unlikely if they were due to random thermal fluctuations. In addition, he reported a 24 hour periodicity in his peaks, the sidereal correlation, that indicated a single source for the radiation, perhaps near the center of our galaxy. These results increased the plausibility of his claims sufficiently so that by 1972 three other experimental groups had not only built detectors, but had also reported results. None was in agreement with Weber.

At this point Collins invokes the experimenters' regress cited earlier. He argues that if the regress is a real problem in science then scientists should disagree about what constitutes a good detector, and that this is what his fieldwork shows. He presents several excerpts from interviews with scientists working in the field that show differing opinions on the quality of detectors.[9] There were also different reasons offered for scientists' belief in Weber's claims. These included the coincidences between two separated detectors, the fact that the coincidence disappeared when one detector signal was delayed relative to the other, and Weber's use of the computer for analysis.[10] Not everyone agreed. Collins argues that these differing opinions demonstrate the lack of any consensus over formal criteria for the validity of gravitational wave detectors. According to Collins, the decision as to what counts as a competently performed experiment is coextensive with the debate about what the paper outcome of the experiment is.

Collins notes that after 1972 Weber's claims were less and less favored. During 1973 three different experimental groups reported negative results and subsequently these groups, as well as three others, reported further negative results. No corroboration of Weber's results was reported during this period. Although in 1972 approximately a dozen groups were involved in experiments

[8]Given any such threshold there is a finite probability that a noise pulse will be larger than that threshold. The point is to show that there are pulses in excess of those expected statistically.

[9]This might also be expected when a new detector is first proposed and there has been little experience in its use. Although one may think about sources of background in advance, it is the actual experience with the apparatus that often tells scientists which of them are present and important.

[10]Weber originally analysed the data using his own observation of the output tapes.

aimed at checking Weber's findings, by 1975 no one, except Weber himself, was still working on that particular problem. Weber's results were regarded as incorrect. There were, however, at least six groups working on experiments of much greater sensitivity, designed to detect the theoretically predicted flux of gravitational radiation.

The reasons offered by different scientists for their rejection of Weber's claims were varied, and not all of the scientists engaged in the pursuit agreed about their importance. During the period 1972–1975 it was discovered that Weber had made several serious errors in his analysis. His computer program for analysing the data contained an error and his statistical analysis of residual peaks and background was questioned and thought to be inadequate. Weber also claimed to find coincidences between his detector and another distant detector when, in fact, the tapes used to provide the coincidences were actually recorded more than 4 hours apart. Weber had found a positive result where even he would not expect one. Others cited the failure of Weber's signal to noise ratio to improve, despite his 'improvements' to his apparatus. In addition, the sidereal correlation disappeared.

Perhaps most important was the uniformly negative results obtained by six other groups. Collins points out that only one of these experimental arrangements was not criticized by other groups, and that all of these experiments were regarded as inadequate by Weber.

> Under these circumstances it is not obvious how the credibility of the high flux case fell so low. In fact, it was not the single uncriticized experiment that was decisive; . . . Obviously the sheer weight of negative opinion was a factor, but given the tractability, as it were, of all the negative evidence, it did not *have* [emphasis in original] to add up so decisively. There was a way of assembling the evidence, noting the flaws in each grain, such that outright rejection of the high flux claim was not the necessary inference (*op. cit.* p. 91).

If Collins is correct in arguing that the negative evidence provided by the replications of Weber's experiment, the application of what we might call epistemological criteria, combined with Weber's acknowledged errors is insufficient to explain the rejection of Weber's results then he must provide another explanation. Collins offers instead the impact of the negative evidence provided by scientist Q.[11] Collins argues that it was not so much the power of Q's experimental result, but rather the forceful and persuasive presentation of that result and his careful analysis of thermal noise in an antenna that turned the tide. Q was also quite aggressive in pointing out Weber's mistakes. After Q's second negative result, no further positive results were reported.[12]

[11]Any reader of the literature will easily identify Q as Richard Garwin.

[12]Collins does not imply that there was anything wrong with the behavior of Q and his group. 'There is no reason to believe that they had anything but the best motives for these actions but they pursued their aim in an unusually vigorous manner (*ibid.*, p. 95).'

Actually, no positive results, other than Weber's, were reported before Q's publication. In fact, I have found no reports of positive results with a Weber bar detector by anyone other than Weber and his collaborators. Collins regards Q's work as the explanation of how the experimenters' regress was solved in this case. 'The growing weight of negative reports, all of which were indecisive in themselves, were crystallized, as it were, by Q. Henceforward, only experiments yielding negative results were included in the envelope of serious contributions to the debate (*ibid.* p. 95)'.

Collins concludes, 'Thus, Q acted as though he did not think that the simple presentation of results with only a low key comment would be sufficient to destroy the credibility of Weber's results. In other words, he acted as one might expect a scientist to act who realized that evidence and arguments alone are insufficient to settle unambiguously the existential status of a phenomenon (*ibid.* p. 95)'.

Scientists did offer other explanations of the discordant results of Weber and his critics. These included possible differences in the detectors; i.e. piezo-electric crystals or other strain detectors, the antenna material, and the electronics; different statistical analysis of the data, the pulse length of the radiation, and calibrations of the apparatus. These last three figure prominently in the subsequent history. Finally there was the invocation of a new, 'fifth force', the possibility that the gravity wave findings were the result of mistakes, deliberate lies, or self-deception, and the explanation by psychic forces. Collins notes that by 1975 all of these alternative explanations, except for the accepted view that Weber had made an error, had disappeared from the scientific discussions. 'This is exactly the sort of change we would expect to take place as the field reached consensus (*ibid.* p. 99)'. Collins suggests that this was not a necessary conclusion, and that scientists might reasonably investigate these more radical possibilities.

Finally, Collins deals with the attempt to break the experimenters' regress by the use of experimental calibration. (See the earlier discussion of calibration). Experimenters calibrated their gravity wave detectors by injecting a pulse of known electrical energy at one end of their antenna and measuring the output of their detector. This served to demonstrate that the apparatus could detect energy pulses and also provided a measure of the sensitivity of the apparatus. One might, however, object that the electrostatic pulses were not an exact analog of gravity waves. Another experimenter did use a different method of calibration. He used a local, rotating laboratory mass to more closely mimic gravity waves.[13]

[13]A local oscillating mass is also not an exact analog. Although it produces tidal gravitational forces in the antenna, it does not produce gravity waves. Only a distant source could do that. Such a mass would, however, have a gravitational coupling to the antenna, rather than an electromechanical one.

According to Collins, Weber was initially reluctant to calibrate his own antenna electrostatically, but did eventually do so. His observations included, however, a quite different method of analysing the output pulses. He used a non-linear, energy algorithm, whereas his critics used a linear, amplitude algorithm. (For a discussion of this difference see Appendix 1). The critics argued that one could show quite rigorously, and mathematically, that the linear algorithm was superior in detecting pulses. The issues of the calibration of the apparatus and the method of analysis used were inextricably tied together. When the calibration was done on Weber's apparatus, it was found that the linear algorithm was 20 times better at detecting the calibration signal than was Weber's non-linear algorithm. For the critics, this established the superiority of their detectors. Weber did not agree. He argued that the analysis and calibration applied only to short pulses, those expected theoretically and used in the calibration, while the signal he was detecting had a length and shape that made his method superior.

Collins regards Weber's agreement to the calibration procedure as a mistake. He had, by agreeing to it, also accepted two assumptions. The first was that gravitational radiation interacted with the antenna in the same way as electrostatic forces. Second, he accepted that the localized insertion of an energy pulse at the end of the antenna had a similar effect to that of a gravity wave that interacted with the entire antenna from a great distance.

Collins concludes,

> The anomalous outcome of Weber's experiments could have led toward a variety of heterodox interpretations with widespread consequences for physics. They could have led to a schism in the scientific community or even a discontinuity in the progress of science. Making Weber calibrate his apparatus with the electrostatic pulses was one way in which his critics ensured that gravitational radiation remained a force that could be understood within the ambit of physics as we know it. They ensured physics' continuity—the maintenance of links between past and future. Calibration is not simply a technical procedure for closing debate by providing an external criterion of competence. In so far as it does work this way, it does so by controlling interpretive freedom. It is the control on interpretation which breaks the circle of the experimenters' regess, not the 'test of a test' itself (*ibid.* pp. 105–106).

Collins states that the purpose of his argument is to demonstrate that science is uncertain. He concludes, however, 'For all its fallibility, science is the best institution for generating knowledge about the natural world that we have (*ibid.* p. 165)'.

## 3. Discussion

Although I agree with Collins concerning the fallibility of science and on its status as 'the best institution for generating knowledge about the natural world we have', I believe there are serious problems with his argument. These are particularly important because the argument, despite Collins's disclaimer, really seems to cast doubt on experimental evidence and on its use in science, and therefore on the status of science as knowledge.

Collins's argument can be briefly summarized as follows. There are no other rigorous independent criteria for either a valid result or for a good experimental apparatus, independent of the outcome of the experiment. This leads to the experimenters' regress in which a good detector can only be defined by its obtaining the correct outcome, whereas a correct outcome is one obtained using a good detector. This is illustrated by the discussion of gravity wave detectors. In practice the regress is broken by negotiation within the scientific community, but the decision is not based on anything that one might call epistemological criteria. This casts doubt not only on the certainty of experimental evidence, but on its very validity. Thus, experimental evidence cannot provide grounds for scientific knowledge.

### *3.1. Gravity Wave Detection*[14]

Collins might correctly argue that the case of gravity wave detectors is a special case, one in which a new type of apparatus was being used to try to detect a hitherto unobserved quantity. I agree.[15] I do not, however, agree that one could not present arguments concerning the validity of the results, or that one could not evaluate the relative merits of two results, independent of the outcome of the two experiments. The regress can be broken by reasonable argument. I will also demonstrate that the published record gives the details of that reasoned argument. Collins's view that there were no formal criteria, applied to deciding between Weber and his critics, may be correct. But, the fact that the procedure was not rule-governed, or algorithmic, does not imply that the decision was unreasonable. (See discussion in Galison (1987), pp. 276–277).

Let us now examine the early history of attempts to observe gravity waves. As we shall see, it was not a question of what constituted a good gravity wave detector, but rather a question of whether or not the detector was operating properly and whether or not the data were being analysed correctly. There is a distinction between data and results, or phenomena, as Bogen and Woodward (1988) have pointed out. All of the experiments did, in fact, use variants of the

[14]I will rely, primarily, on a panel discussion on gravitational waves that took place at the Seventh International Conference on General Relativity and Gravitation (GR7), Tel-Aviv University, 23–28 June, 1974. The panel included Weber and three of his critics, Tyson, Kafka, and Drever, and included not only papers presented by the four scientists, but also discussion, criticism, and questions. It includes almost all of the important and relevant arguments concerning the discordant results. The proceedings were published as Shaviv and Rosen (1975). Unless otherwise indicated all quotations in this section are from Shaviv and Rosen (1975). I shall give the author and the page numbers in the text.

[15]One might then wonder why he uses such an atypical example as his illustration of the experimenters' regress.

Weber antenna, and, with the exception of Weber, similar analysis procedures. The discordant results reported by Weber and his critics are not unusual occurrences in the history of physics, particularly at the beginning of an experimental investigation of a phenomenon.[16]

There was a clear claim by Weber that gravity waves had been observed. There were several other results of experiments to detect such waves that were negative. In addition, there were admitted errors made by Weber and serious questions raised concerning Weber's analysis and calibration procedures. To be fair, not everyone working in the field, particularly Weber, agreed about the importance of these problems. Collins expressed some surprise that the credibility of Weber's results fell so low. '. . . given the tractability, as it were, of all the negative evidence, it did not have to add up so decisively (Collins *op. cit.*, p. 91)'. I am not surprised. I believe that Collins has seriously overstated the tractability of the negative results and understated the weight of the evidence against Weber's results. The fact that Weber's critics might have disagreed about the force of particular arguments does not mean that they did not agree that Weber was wrong. To decide the question we must look at the history of the episode as given in published papers, conference proceedings, and public letters. I believe that the picture these give is one of overwhelming evidence against Weber's result, and that the decision, although not rule governed, was reasonable, and based on epistemological criteria.

I begin with the issue of calibration and Weber's analysis procedure. The question of determining whether or not there is a signal in a gravitational wave detector, or whether or not two such detectors have fired simultaneously is not easy to answer. There are several problems. One is that there are energy fluctuations in the bar due to thermal, acoustic, electrical, magnetic, and seismic noise, etc. When a gravity wave strikes the antenna its energy is added to the existing energy. This may change either the amplitude or the phase, or both, of the signal emerging from the bar. It is not just a simple case of observing a larger signal from the antenna after a gravitational wave strikes it. This difficulty informs the discussion of which was the best analysis procedure to use.

The non-linear, or energy, algorithm preferred by Weber was sensitive only to changes in the amplitude of the signal. The linear algorithm, preferred by everyone else, was sensitive to changes in both the amplitude and the phase of the signal. (See discussion in Appendix 1). Weber preferred the non-linear procedure because it resulted in proliferation, several pulses exceeding threshold for each input pulse to his detector. 'We believe that this kind of cascading may result in observation of a larger number of two-detector coincidences for algorithm (6) [non-linear] than for (7) [linear], at certain

[16]For a discussion of other similar episodes, that of experiments on atomic parity violation and on the Fifth Force in gravity see Franklin (1990, 1993a, and 1993b).

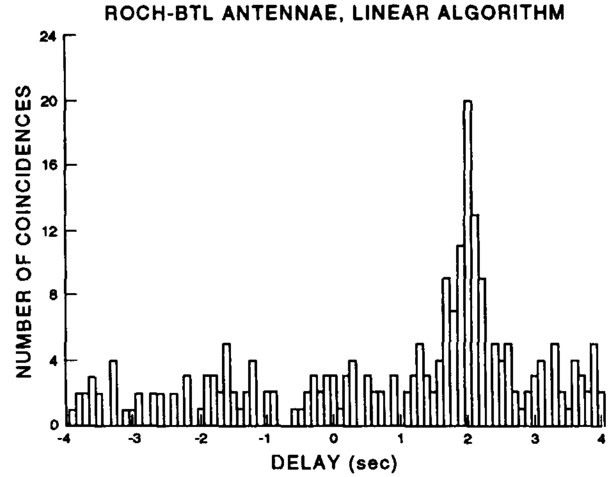


*Fig. 2. A plot showing the calibration pulses for the Rochester-Bell Laboratory collaboration. The peak due to the calibration pulses is clearly seen. From Shaviv and Rosen (1975), permission requested, but no reply.*

energies (Weber, p. 246)'.[17] Weber admitted, however, that the linear algorithm, preferred by his critics, was more efficient at detecting calibration pulses. He stated, 'It is found for pulses which increase the energy of the normal mode from zero to $kT$ that algorithm (7) [linear] gives a larger amount of response pulses exceeding thresholds, than algorithm (6) [non-linear]. Perhaps this is the reason that algorithm (7) is preferred by a number of groups (Weber, p. 247)'. (I note here that Weber's earlier statement indicated that more than one pulse was detected for a single input pulse using the non-linear algorithm. His second statement refers to the efficiency of detecting individual calibration pulses. The language is somewhat confusing). Similar results on the superiority of the linear algorithm for detecting calibration pulses were reported by both Kafka (pp. 258–259) and Tyson (pp. 281–282). Tyson's results for calibration pulse detection are shown for the linear algorithm in Fig. 2, and for the non-linear algorithm in Fig. 3. There is a clear peak for the linear algorithm, whereas no such peak is apparent for the non-linear procedure. (The calibration pulses were inserted periodically during data taking runs. The peak was displaced by 2 seconds by the insertion of a time delay, so that the calibration pulses would not mask any possible real signal, which was expected at zero time delay).

Nevertheless, Weber preferred the non-linear algorithm. His reason for this was that this procedure gave a more significant signal than did the linear one. This is illustrated in Fig. 4, in which the data analysed with the non-linear algorithm are presented in (a) and for the linear procedure in (b). 'Clearly these

[17]One might worry that this cascading effect would give rise to spurious coincidences.

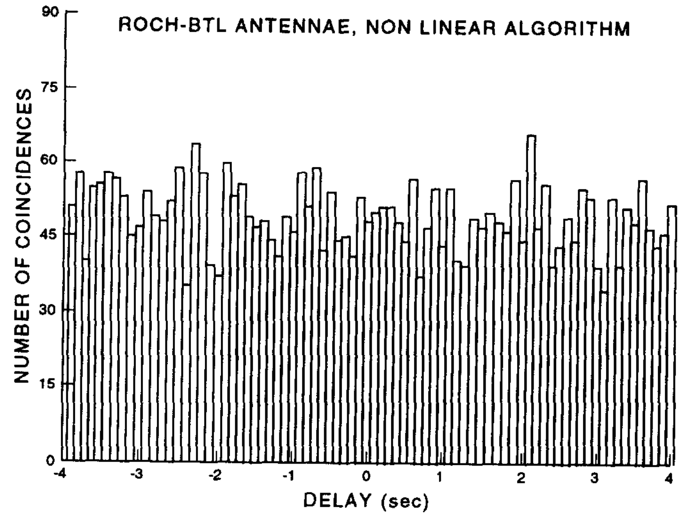


*Fig. 3. A time-delay plot for the Rochester–Bell Laboratory collaboration, using the non-linear algorithm. No sign of any zero-delay peak is seen. From Shaviv and Rosen (1975), permission requested, but no reply.*

results are inconsistent with the generally accepted idea that $\dot{x}^2+\dot{y}^2$ [the linear algorithm] should be the better algorithm (Weber, pp. 251–252)'. Weber was, in fact, using the positive result to decide which was the better analysis procedure. If anyone was 'regressing', it was Weber.

Weber's failure to calibrate his apparatus was criticized by others. 'Finally, Weber has not published any results in calibrating his system by the impulsive introduction of known amounts of mechanical energy into the bar, followed by the observation of the results either on the single detectors or in coincidence (Levine and Garwin, 1973, p. 177)'.

His critics did, however, analyse their own data using both algorithms. If it was the case that, unlike the calibration pulses where the linear algorithm was superior, using the linear algorithm either masked or failed to detect a real signal, then using the non-linear algorithm on their data should produce a clear signal. None appeared. Typical results are shown in Figs 3 and 5. Figure 3, which is Tyson's data analysed with the non-linear algorithm, not only shows no calibration peak, but it does not show a signal peak at zero time delay. It is quite similar to the data analysed with the linear algorithm shown in Fig. 5. (I note that for this data run no calibration pulses were inserted).[18] Kafka also reported the same result, no difference in signal between the linear and the non-linear analysis.

[18]Collins does not discuss the fact that Weber's critics exchanged both data and analysis programs, and that they analysed their own data with Weber's preferred non-linear analysis algorithm and failed to find a signal. This fact, as documented in the published record, would seem to argue for the use of epistemological criteria in the evaluation of the discordant experimental results.

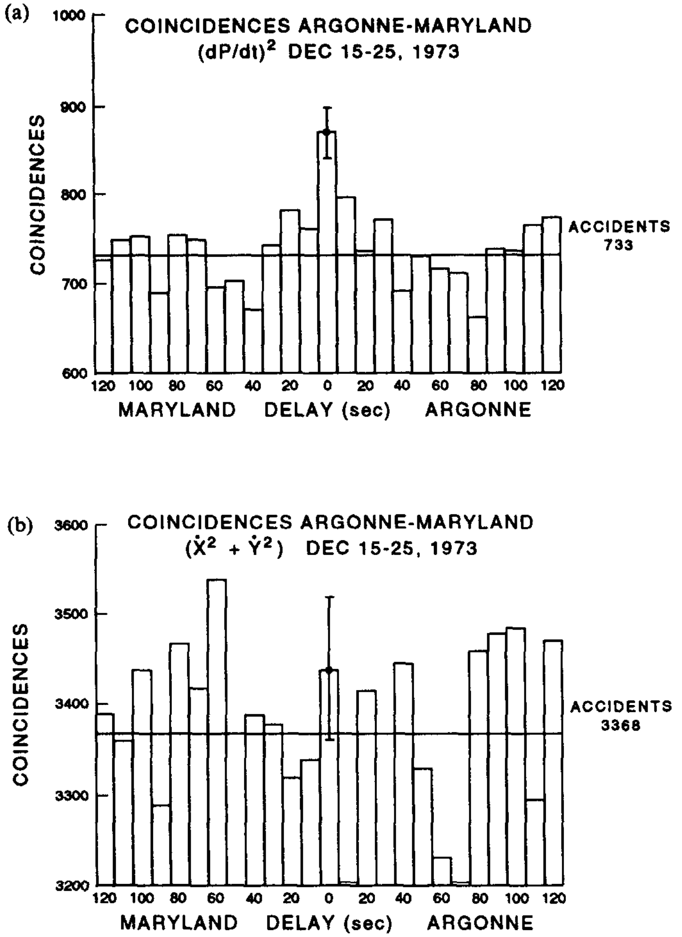


*Fig. 4. Weber's time-delay data for the Maryland–Argonne collaboration for the period 15–25 December, 1973. The top graph uses the non-linear algorithm, whereas the bottom uses the linear algorithm. The zero-delay peak is seen only with the non-linear algorithm. From Shaviv and Rosen (1975), permission requested, but no reply.*

Weber had an answer. He suggested that although the linear algorithm was better for detecting calibration pulses, which were short, the real signal of gravitational waves was a longer pulse than most investigators thought. He argued that the non-linear algorithm that he used was better at detecting these longer pulses. The critics did think that gravitational radiation would be produced in short bursts. For example, Douglass and others (1975) remarked that, 'the raw data are filtered in a manner optimum for short pulses (p. 480)'. 'The filter was chosen and optimized on the basis of optimal filter theory and the assumption that bursts of gravitational radiation would be much shorter in

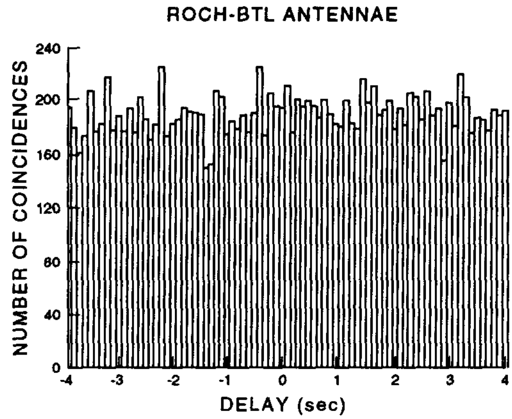


*Fig. 5. A time-delay plot for the Rochester–Bell Laboratory collaboration, using the linear algorithm. No sign of a zero-delay peak is seen. From Shaviv and Rosen (1975), permission requested, but no reply.*

duration then the 0.1 sec response time of the detector electronics (*ibid.* pp. 480–481).'

Still, if the signal was longer, one would have expected it to show up when the critics's data was processed with the non-linear algorithm. It didn't. (See Fig. 3.) Tyson remarked,

> I would merely like to comment that all the experiments of the Weber type, where you have an integrated calorimeter which asks the question: 'Did the energy increase or decrease in the last tenth of a second?'—all those experiments, of which my own, Weber's, and Kafka's are an example—would respond in a similar manner to a given pulse shape in the metric given the same algorithm. I think it must be something which only your (Weber) detector is sensitive to and not ours (Tyson, p. 288).

Drever also reported that he had looked at the sensitivity of his apparatus with arbitrary waveforms and pulse lengths. Although he found a reduced sensitivity for longer pulses, he did analyse his data to explicitly look for such pulses. He found no effect (Fig. 6). He also found no evidence for gravity waves using the short pulse (linear) analysis (Fig. 7).

Drever summarized the situation in June 1974 as follows.

> Perhaps I might just express a personal opinion on the situation because you have heard about Joseph Weber's experiments getting positive results, you have heard about three other experiments getting negative results and there are others too getting negative results, and what does this all mean? Now, at its face value there is obviously a strong discrepancy but I think it is worth trying hard to see if there is any way to fit all of these apparently discordant results together. I have thought about this very hard, and my conclusion is that in any one of these experiments relating to Joe's one, there is always a loophole. It is a different loophole from one experiment to the next. In the case of our own experiments, for example, they are not very sensitive for long pulses. In the case of the experiments described by Peter Kafka and Tony Tyson, they

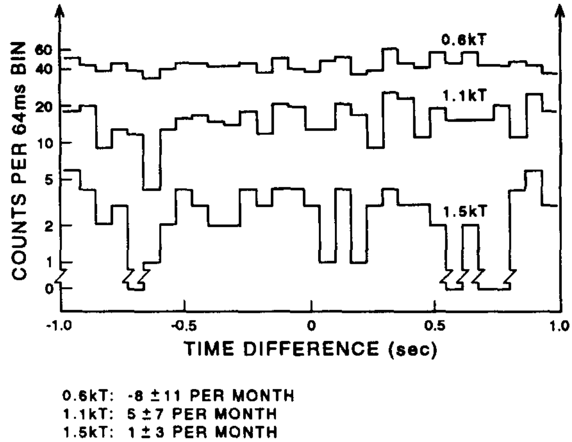


*Fig. 6. Drever's data looking for long pulses. No zero-delay peak is seen. From Shaviv and Rosen (1975), permission requested, but no reply.*

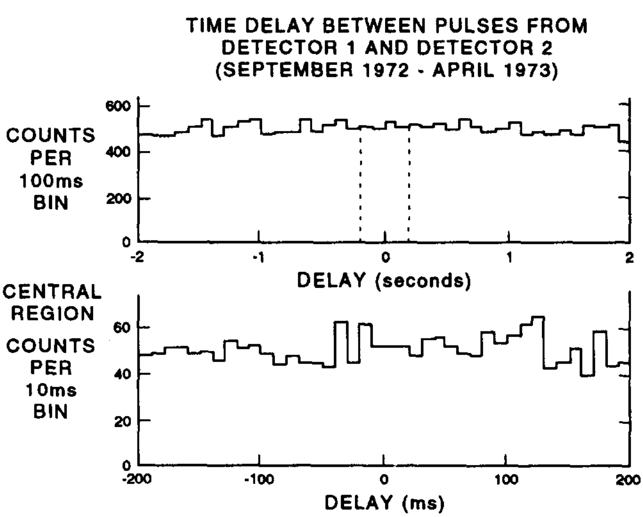


*Fig. 7. Drever's time delay plot. No sign of a peak at zero-delay is seen. From Shaviv and Rosen (1975), permission requested, but no reply.*

used a slightly different algorithm which you would expect to be the most sensitive, but it is only the most sensitive for a certain kind of waveform. In fact, the most probable waveforms. But you can, if you try very hard, invent artificial waveforms for which this algorithm is not quite so sensitive.[19] So it is not beyond the bounds of possibility that the gravitational waves have that particular kind of waveform. However, our own experiment would detect that type of waveform; in fact, as efficiently as it would the more usually expected ones, so I think we close that

[19]Weber did, in fact, report such a waveform (Weber, 1975).

loophole. I think that when you put all these different experiments together, because they are different, most loopholes are closed. It becomes rather difficult now, I think, to try and find a consistent answer. But still not impossible, in my opinion. One cannot reach a really definite conclusion, but it is rather difficult, I think to understand how all the experimental data can fit together (Drever, pp. 287–288).

There was considerable cooperation among the various groups. They exchanged both data tapes and analysis programs. 'There has been a great deal of intercommunication here. Much of the data has been analysed by other people. Several of us have analysed each other's data using either our own algorithm or each other's algorithms (Tyson, p. 293)'. This led to the first of several questions about possible serious errors in Weber's analysis of his data. Douglass first pointed out that there was an error in one of Weber's computer programs.

The nature of the error was such that any above-threshold event in antenna A that occurred in the last or the first 0.1 sec time bin of a 1000 bin record is erroneously taken by the computer program as in coincidence with the next above-threshold event in channel B, and is ascribed to the time of the later event. Douglass showed that in a four-day tape available to him and included in the data of [Weber, 1973], nearly all of the so-called 'real' coincidences of 1–5 June (within the 22 April to 5 June 1973 data) were created individually by this simple programming error. Thus not only some phenomenon besides gravity waves *could*, but in fact *did* cause the zero-delay excess coincidence rate (Garwin, 1974, p. 9, emphasis in original).

Weber admitted the error, but did not agree with the conclusion.

This histogram is for the very controversial tape 217. A copy of this tape was sent to Professor David Douglass at the University of Rochester. Douglass discovered a program error and incorrect values in the unpublished list of coincidences. Without further processing of the tape, he (Douglass) reached the incorrect conclusion that the zero delay excess was one per day. This incorrect information was widely disseminated by him and Dr R. L. Garwin of the IBM Thomas J. Watson Research Laboratory. After all corrections are applied, the zero delay excess is 8 per day. Subsequently, Douglass reported a zero delay excess of 6 per day for that tape (Weber, p. 247).

Although Weber reported that his corrected result had been confirmed by scientists at other laboratories and that copies of the documents had been sent to editors and workers in the field I can find no corroboration of any of Weber's claims in the published literature. At the very least, this error raised doubts about the correctness of Weber's results (shown in Fig. 8).

Another serious question was raised concerning Weber's analysis of his data. This was the question of selectivity and possible bias. Tyson characterized the difference between Weber's methods and those of his critics.

I should point out that there is a very important difference in essence in the way in which many of us approach this subject and the way Weber approaches it. We have taken the attitude that, since these are integrating calorimeter type experiments which are not too sensitive to the nature of pulses put in, we simply maximize the sensitivity

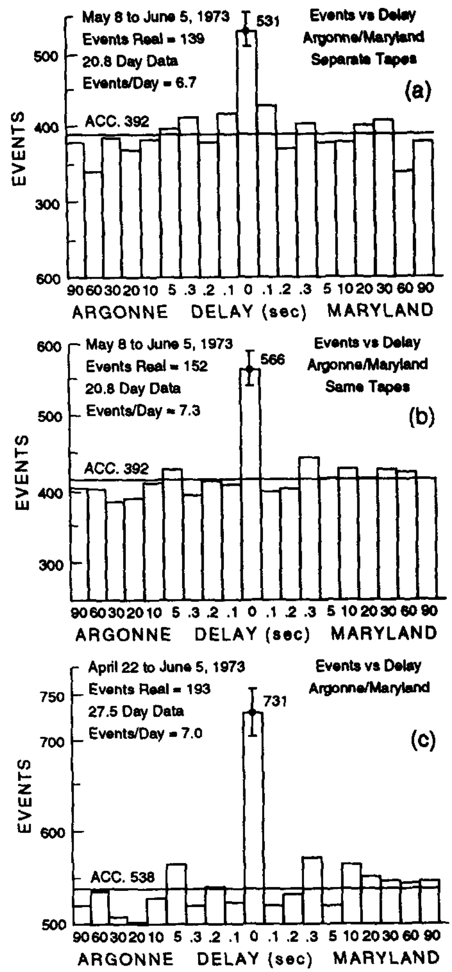


*Fig. 8. Weber's results. The peak at zero time delay is clearly seen. From Weber and others (1973), permission requested, but no reply.*

and use the algorithms which we found maximized the signal to noise ratio, as I showed you. Whereas Weber's approach is, he says, as follows. He really does not know what is happening, and therefore he or his programmer is twisting all the adjustments in the experiment more or less continuously, at every instant in time locally maximizing the excess at zero time delay. I want to point out that there is a potentially serious possibility for error in this approach. No longer can you just speak about Poisson statistics. You are biasing yourself to zero time delay, by continuously modifying the experiment on as short a time scale as possible (about four days), to maximize the number of events detected at zero time delay. We are taking the opposite approach, which is to calibrate the antennas with all possible known sources of excitation, see what the result is, and maximize our probability of detection. Then we go through all of the data with that one algorithm and integrate all of them. Weber

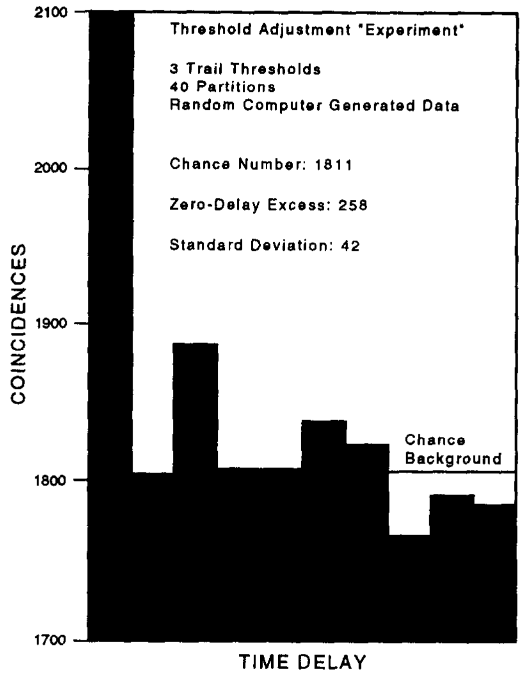


*Fig. 9. The result of selecting thresholds that maximized the zero-delay signal, for Levine's computer simulation. From Garwin (1974), permission requested, but no reply.*

made the following comment before and I quote out of context: 'Results pile up'. I agree with Joe (Weber). But I think you have to analyse all of the data with one well-understood algorithm (Tyson, p. 293).

A similar criticism was offered by Garwin, who also presented evidence from a computer simulation to demonstrate that a selection procedure such as Weber's could indeed produce his positive result.

Second, in view of the fact that Weber at CCR-5 [a conference on General Relativity held in Cambridge][20] explained that when the Maryland group failed to find a positive coincidence excess 'we try harder', and since in any case there has clearly been selection by the Maryland group (with the publication of data showing positive coincidence excesses but with no publication of data that does not show such excesses),[21] James L. Levine has considered an extreme example of such selections. In Figure [9] is shown the combined histogram of 'coincidences' between two independent streams of random computer-generated data. This 'delay histogram' was obtained by partitioning the data into 40 segments. For each segment, 'single events' were defined in each 'channel' by assuming one of three thresholds a, b, or c. That combination of thresholds was chosen for each segment which gave the maximum

[20] I have been unable to find a published proceedings of this conference.

[21] As Weber answered, the Maryland group had presented data showing no positive coincidence excess at GR7. Garwin was not, however, at that meeting, and the proceedings were not published until after Garwin's 1974 letter appeared.

'zero delay coincidence' rate for that segment. The result was 40 segments selected from one of nine 'experiments'. The 40 segments are summarized in Figure [9], which shows a 'six-standard-deviation' zero-delay excess (Garwin *op. cit.*, pp. 9–10).

Weber denied both charges.

It is not true that we turn our knobs continuously. I have been full time at the University of California at Irvine for the last six months, and have not been turning the knobs by remote control from California [Weber's group and one of his antennas was located at the University of Maryland]. In fact, the parameters have not been changed for almost a year. What we do is write the two algorithms on a tape continuously. The computer varies the thresholds to get a computer printout which is for 31 different thresholds. The data shown are not the results of looking over a lot of possibilities and selecting the most attractive ones. We obtain a result that is more than three standard deviations for an extended period for a wide range of thresholds. I think it is very important to take the point of view that the histogram itself is the final judge of what the sensitivity is (Weber, pp. 293–294).

Weber did not, however, specify his method of data selection for his histogram. In particular, he did not state that all of the results presented in a particular histogram had the same threshold.

Interestingly, Weber cited evidence provided by Kafka as supporting a positive gravity wave result. Kafka did not agree. This was because the evidence resulted from performing an analysis using different data segments and different thresholds. Only one showed a positive result, indicating that such selectivity could produce a positive result. Kafka's results are shown in Fig. 10. Note that the positive effect is seen in only the bottom graph. 'The very last picture (Fig. 10) is the one in which Joe Weber thinks we have discovered something, too. This is for 16 days out of 150. There is a 3.6 $\sigma$ [standard deviation] peak at zero time delay, but you must not be too impressed by that. It is one out of 13 pieces for which the evaluation was done, and I looked at least at 7 pairs of thresholds. Taking into account selection we can estimate the probability to find such a peak accidentally to be of the order of 1% (Kafka, p. 265)'.

There was also a rather odd result reported by Weber.

First, Weber has revealed at international meetings (Warsaw, 1973, etc.) that he had detected a 2.6 standard deviation excess in coincidence rate between a Maryland antenna [Weber's apparatus] and the antenna of David Douglass at the University of Rochester. Coincidence excess was located not at zero time delay but at '1.2 seconds', corresponding to a 1-sec intentional offset in the Rochester clock and a 150-millisecond clock error. At CCR-5, Douglass revealed, and Weber agreed, that the Maryland Group had mistakenly assumed that the two antennas used the same time reference, whereas one was on Eastern Daylight Time and the other on Greenwich Mean Time. Therefore, the 'significant' 2.6 standard deviation excess referred to gravity waves that took four hours, zero minutes and 1.2 seconds to travel between Maryland and Rochester (*op. cit.*, p. 9).

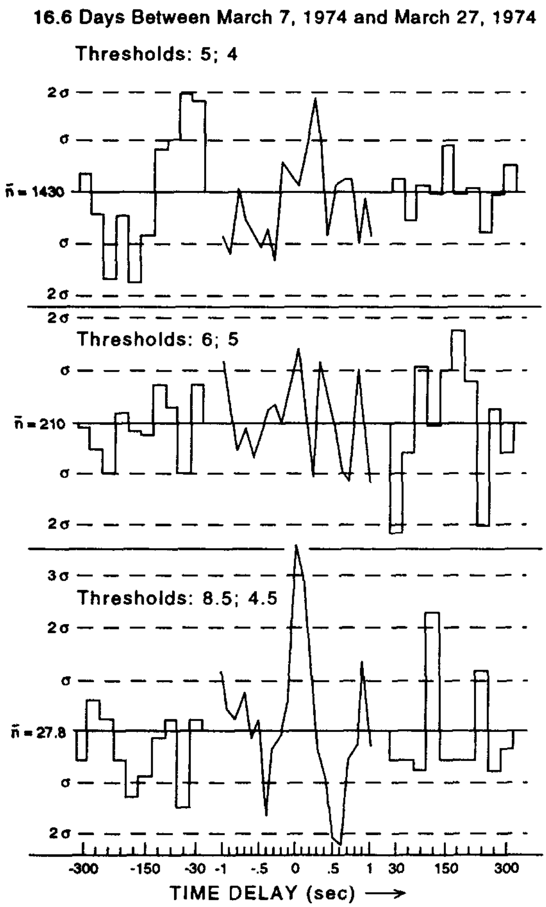


*Fig. 10. Kafka's results using varying thresholds. A clear peak is seen at zero-delay. From Shaviv and Rosen (1975), permission requested, but no reply.*

Weber answered that he had never claimed that the 2.6 standard deviation effect he had reported was a positive result. By producing a positive result where none was expected, Weber had, however, certainly cast doubt on his analysis procedures.

Levine and Garwin (1974) and Garwin *op. cit.* (1974) raised yet another doubt about Weber's results. This was the question of whether or not Weber's apparatus could have produced his claimed positive results. Here again, the evidence came from a computer simulation.

Figure [11(b)] shows the 'real coincidences' confined to a single 0.1 sec bin in the time delay histogram. James L. Levine and I observed that the Maryland Group used a 1.6 Hz bandwidth 'two-stage Butterworth filter'. We suspected that mechanical excitations of the antenna (whether caused by gravity waves or not) as a consequence of the 1.6 Hz bandwidth would not produce coincident events limited to a single

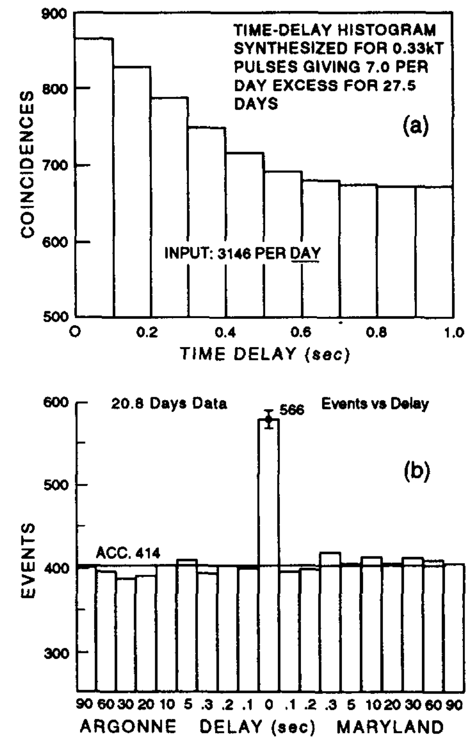


*Fig. 11. (a) Computer simulation result obtained by Levine for signals passing through Weber's electronics. (b) Weber's reported result. The difference is clear. From Levine and Garwin (1974), permission requested, but no reply.*

> 0.1 sec time bin. Levine has simulated the Maryland apparatus and computer algorithms to the best of the information available in Weber and others (1973) and has shown that the time-delay histogram for coincident pulses giving each antenna 0.3 $kT$ is by no means confined to a single bin, but has the shape shown in Fig. [11(a)] (Garwin *op. cit.*, p. 9).

Let us summarize the evidential situation concerning gravity waves at the beginning of 1975. There were discordant results. Weber had reported positive results on gravitational radiation, whereas six other groups had reported no evidence for such radiation. The critic's results were not only more numerous, but had also been carefully cross-checked. The groups had exchanged both data and analysis programs and confirmed their results. The critics had also investigated whether or not their analysis procedure, the use of a linear algorithm, could account for their failure to observe Weber's reported results. They had used Weber's preferred procedure, a non-linear algorithm, to analyse their data, and still found no sign of an effect. They had also calibrated their experimental apparata by inserting electrostatic pulses of known energy and found that they could detect a signal. Weber, on the other hand, as well as his critics using his analysis procedure, could not detect such calibration pulses.

There were, in addition, several other serious questions raised about Weber's analysis procedures. These included an admitted programming error that generated spurious coincidences between Weber's two detectors, possible selection bias by Weber, Weber's report of coincidences between two detectors when the data had been taken 4 hours apart, and whether or not Weber's experimental apparatus could produce the narrow coincidences claimed.

It seems clear that the critic's results were far more credible than Weber's. They had checked their results by independent confirmation, which included the sharing of data and analysis programs. They had also eliminated a plausible source of error, that of the pulses being longer than expected, by analysing their results using the non-linear algorithm and by looking for such long pulses. They had also calibrated their apparata by injecting known pulses of energy and observing the output.

In addition, Weber's reported result failed several tests. Weber had not eliminated the plausible error of a mistake in his computer program. It was, in fact, shown that this error could account for his result. It was also argued that Weber's analysis procedure, which varied the threshold accepted, could also have produced his result. Having increased the credibility of his result when he showed that it disappeared when the signal from one of the two detectors was delayed, he then undermined his result by obtaining a positive result when he thought two detectors were simultaneous, when, in fact, one of them had been delayed by 4 hours. As Garwin also argued, Weber's result itself argued against its credibility. The coincidence in the time delay graph was too narrow to have been produced by Weber's apparatus. Weber's analysis procedure also failed to detect calibration pulses.

Contrary to Collins, I believe that the scientific community made a reasoned judgment and rejected Weber's results and accepted those of his critics. Although no formal rules were applied, i.e. if you make four errors, rather than three, your results lack credibility; or if there are five, but not six, conflicting results, your work is still credible; the procedure was reasonable.

I also question Collins's account of Garwin's role (Scientist Q). Although Garwin did present strong and forceful arguments against Weber's result, the same arguments were being made at the time by other scientists, albeit in a somewhat less aggressive manner. Collins's point that Garwin behaved as if he thought that a reasoned argument would not be sufficient to destroy the credibility of Weber's result also seems questionable. Garwin's behavior could also be that of a scientist who believed that Weber's results were wrong, and that valuable time and resources were being devoted to the investigation of an incorrect result, and who thought that Weber's adherence to his incorrect result was casting doubt on all of the good work being done in the field.[22] It might also just be the case that Garwin is a forceful and powerful polemicist.

I also question the role of Garwin as the crystallizer of the opposition to Weber. As we have seen, other scientists were presenting similar arguments against Weber. At GR7, Garwin's experiment was mentioned only briefly, and although the arguments about Weber's errors and analysis were made, they were not attributed to the absent Garwin.[23]

For those who prefer a theory-first view of science, I note that although disagreement with theoretical predictions may have played a role in the skepticism about Weber's initial results, it played no major role in the later dispute. Once Weber had established the credibility of his results by varying the time delay and seeing the effect disappear and by observing the sidereal correlation the argument became almost solely experimental. Was Weber really observing gravitational radiation?

### *3.2. Calibration*

A point that should be emphasized is that although calibration, and its success or failure, played a significant role in the dispute, it was not decisive, as Collins correctly points out. Other arguments were needed. In most cases, failure to detect a calibration signal would be a decisive reason for rejecting an experimental result. In this case it was not. The reason for this was precisely because the scientists involved seriously considered the question of whether or not an injected electrostatic energy pulse was an adequate surrogate for a gravity wave. It was doubts as to its adequacy that led to the variation in analysis procedures and to the search for long pulses.

The detection of gravitational radiation is not a typical physics experiment. Although experiments may find new phenomena it is not usual to have an experiment in which a new type of apparatus is used to search for a hitherto unobserved phenomenon. In a typical physics experiment there is usually little question as to whether or not the calibration signal is an adequate surrogate for the signal one wishes to detect. It is usually the case that calibration of the apparatus is independent of the phenomenon one wants to observe. A few illustrative cases will help.

Consider the problem I faced as an undergraduate assistant in a research laboratory. I was asked to determine the chemical composition of the gas in a discharge tube and I was given an optical spectroscope. The procedure followed was to use the spectroscope to measure the known spectral lines from various sources such as hydrogen, sodium, and mercury. The fact that I could measure

[22]Several scientists working on gravitational radiation mentioned that they thought Weber had, at least to some extent, discredited work in the field (private communication).

[23]The panel discussion on gravitational waves covers 56 pages (243–298) in Shaviv and Rosen (1975). Tyson's discussion of Garwin's experiment occupies one short paragraph (approximately one quarter of a page) on p. 290.

these known lines accurately showed that the apparatus was working properly.[24] In addition to providing a check on whether or not I could measure spectral lines of optical wavelengths, this procedure also provided a calibration of my apparatus. I could determine small corrections to my results as a function of wavelength. I then proceeded to determine the composition of the gas in the discharge tube by measuring the spectral lines emitted and comparing them with known spectra. There was no doubt that the calibration procedure was adequate. The calibration lines measured spanned the same wavelength region as the ones I used to determine the composition.

Let us consider a more complex experiment, that of the Princeton group (Christenson and others, 1964) that observed the decay $K^{o}_{L} \rightarrow 2\pi$ and established the violation of CP symmetry (combined particle–antiparticle and space–reflection symmetry). The decay was detected by measuring the momenta of the two charged decay particles and reconstructing their invariant mass, assuming the decay particles were pions, and reconstructing the direction of the decaying particle relative to the beam. If it was a $K^{o}_{L}$ decay into two pions the mass should be the mass of the $K^{o}_{L}$ and the angle should be zero. An excess of events was indeed found at the $K^{o}_{L}$ mass and at zero angle to the beam. In order to demonstrate that the apparatus was functioning properly and that it could detect such decays, it was checked by looking at the known phenomenon of the regeneration of $K^{o}_{S}$ mesons, followed by their decay into two pions. If it was operating properly the distributions in mass and angle in the case of both the $K^{o}_{S}$ decays and the proposed $K^{o}_{L}$ decays should have been identical. They were. [For details of this experiment see Franklin (1986, Chapters 3, 6 and 7).] Here too there was no doubt that the surrogate and the phenomenon were sufficiently similar. Both detected two particle decays of particles which had the $K^{o}$ mass, and which were travelling parallel to the beam.

A somewhat different example is provided by an experiment to measure the $K^{+}_{e2}$ branching ratio (Brown and others, 1967). [For further discussion of this experiment see Franklin (1990, Chapter 6).] In this case the decay positron resulting from the decay was to be identified by its momentum, its range in matter, and by its counting in a Cerenkov counter set to detect positrons. The proper operation of the apparatus was shown, in part, by the results themselves. Because the $K^{+}_{e2}$ decay was very rare (approximately $10^{-5}$) compared to other known $K^{+}$ decay modes such decays in coincidence with noise in the Cerenkov counter would be detected. In particular, the muon from $K^{+}_{\mu 2}$ decay, which had a known momentum of 236 MeV/$c$, was detected. A peak was observed at the predicted momentum, establishing that the apparatus could measure momentum accurately. In addition, the width of the peak determined the experimental momentum resolution, a quantity needed for the analysis of the

[24]It also showed the experimenters that I was working properly.

experiment. The Cerenkov counter was checked, and its efficiency for positrons measured, by comparing it to a known positron detector in an independent experiment. The apparatus was also sensitive to $K^+_{e3}$ decay. This decay produced high energy positrons with a maximum momentum of 227 MeV/*c*, which was quite close to the 246 MeV/*c* momentum expected for $K^+_{e2}$ decay. High energy $K^+_{e3}$ positrons were used to determine the range in matter expected for the $K^+_{e2}$ positrons, and to demonstrate that the apparatus could indeed measure the range of positrons in that energy region. The approximately 10% difference in momentum was considered small enough, given the known behavior of positrons in this energy region. In this case, too, there was no doubt as to the adequacy of the calibration.

In all three cases the calibration of the apparatus did not depend on the outcome of the experiment in question. In these cases proper operation of the experimental apparatus was demonstrated independently of the composition of the gas discharge, whether or not the $K^\circ_L$ actually decays into two pions, or what the $K^+_{e2}$ branching ratio was. Clearly, three examples do not demonstrate that calibration always works, but they are, I believe, far more typical of the calibration procedures used in physics than is gravity wave detection. I also believe that in cases such as these they are legitimately more decisive. Had any of these calibration procedures failed, then the results of the experiments would have been rejected. In the case of gravity waves, as we have seen, calibration, while important, was not decisive. Scientists are quite good at the pragmatic epistemology of experiment.

Collins also claims that calibration is not a 'test of a test', but rather breaks the circle of the experimenters' regress by its control of the interpretation of experimental results. He offers Weber's failed calibration as an explanation of why alternative explanations of the discordant results of Weber and his critics were not offered after 1975.

There is a simpler explanation for the lack of alternatives. Weber's result was reasonably regarded as wrong. There is no need to explain an incorrect result.

## 4. Conclusion

I have argued that Collins's argument for the experimenters' regress is wrong. He conflates the difficulty of getting an experiment to work with the problem of demonstrating that it is working properly. This leads him, particularly in the case of the TEA laser, to argue against the possibility of the replication of an experiment. (See discussion in note 1.) The impossibility of replication, combined with what he claims is the lack of formal criteria for the proper operation of an experimental apparatus leads to the experimenters' regress. Gravity wave detection is then used to illustrate the regress.

I believe that I have shown that his account of gravity waves is incorrect. Epistemological criteria were reasonably applied to decide between Weber's result and those of his critics. I have also argued that although calibration was not decisive in the case of gravity wave detectors, nor should it have been, it is often a legitimate and important factor, and may even be decisive, in determining the validity of an experimental result.

Both the argument about the impossibility of replication and the lack of criteria in deciding the validity of experimental results fail. The history of gravity wave detectors does not establish what Collins claims it does. There are no grounds for belief in the experimenters' regress.

## 5. Epilogue

At the present time gravity waves have not been detected by either the use of Weber bar antennas or by the newer technique of using an interferometer, in which the gravitational radiation will have a differential effect on the two arms of the interferometer and thus change the observed interference pattern. The radiation has not been detected even though current detectors are several orders of magnitude more sensitive than those in use in 1975.[25]

Gravity waves have, however, been observed. They have been detected by measuring the change in orbital period of a binary pulsar. Such a binary system should emit gravitational radiation, thereby losing energy and decreasing the orbital period. This effect was measured using the two results of Hulse and Taylor (1975), which provided the initial measurement of the period, and of Weisberg and Taylor (1984), which measured the period at a later time. The measured change in the period was $(-2.40 \pm 0.09) \times 10^{-12}$ s s$^{-1}$, in excellent agreement with the theoretical prediction of $(-2.403 \pm 0.002) \times 10^{-12}$ s s$^{-1}$. 'As we have pointed out before most relativistic theories of gravity other than general relativity conflict strongly with our data, and would appear to be in serious trouble in this regard. It now seems inescapable that gravitational radiation exists as predicted by the general relativistic quadrupole formula (Weisberg and Taylor, 1984, p. 1350)'.[26] If General Relativity is correct, Weber should not have observed a positive result.

*Acknowledgements* — Part of this work was supported by a Faculty Fellowship and grant-in-aid from the Council on Research and Creative Work, Graduate School, University of Colorado. I thank the Council for its support. This material is based on work partially supported by the National Science Foundation under Grant No. DIR-9024819. Any opinions, findings, and conclusions or recommendations are those of the author and do not necessarily reflect the views of the National Science Foundation. I am also grateful to an anonymous referee for helpful comments and suggestions.

[25]An account of an experiment using such a detector appears in Astone and others (1993). Using a very sensitive cryogenic antenna they set a limit of no more than 0.5 events/day, in contrast to Weber's claim of approximately seven events/day.

[26]More recent measurements and theoretical calculations give $(2.427 \pm 0.026) \times 10^{-12}$ s s$^{-1}$ (Measured) (Taylor and Weisberg, 1989) and $(2.402576 \pm 0.000069) \times 10^{-12}$ s s$^{-1}$ (Theory) (Damour and Taylor, 1991).

**Appendix 1**

'Let the output voltage of the gravitational radiation antenna amplifier be given by

$$A = F(t) \sin(\omega_o t + \varphi) \tag{1}$$

where $\omega_o$ is the normal mode angular frequency. The amplitude $F(t)$ and the phase $\varphi$ have values characteristic of signals and noise. It is now common practice to obtain from (1) the amplitude and phase by combining (1) with local reference oscillator voltages $\sin \omega_o t$ and $\cos \omega_o$ to obtain:

$$A \cos \omega_o t = \tfrac{1}{2}F(t)\,[\sin(2\omega_o t + \varphi) + \sin \varphi] \tag{2}$$

$$A \sin \omega_o t = \tfrac{1}{2}F(t)\,[\cos \varphi - \cos(2\omega_o t + \varphi)]. \tag{3}$$

After filtering with a time constant short compared with the antenna relaxation time, (2) and (3) become the averages

$$x = \langle F(t) \cos \varphi/2 \rangle \tag{4}$$

$$y = \langle F(t) \sin \varphi/2 \rangle. \tag{5}$$

An incoming signal may change phase and amplitude of the detector voltage, depending on the initial noise-induced phase relations. The detector output voltage includes narrow band noise of the normal mode of the antenna $V_{ANT}$ and relatively wide band noise $V_N$ from transducers and electronics. To search for sudden changes in amplitude we may observe a function of the derivative of the power $P$ which for convenience is taken as the (positive) quantity:

$$(dP/dt)^2 = [\Delta(x^2 + y^2)/\tau]^2 = [\Delta(V_{ANT} + V_N)^2/\tau]^2 \rightarrow [2\Delta(V_{ANT}V_N)/\tau]^2. \tag{6}$$

(6) is independent of the phase. Incoming signals which change only the phase would therefore be missed and to include such cases we may search for sudden changes in the quantity

$$(dx/dt)^2 + (dy/dt)^2 = [(\Delta(V_{ANT} + \Delta V_N)^2{}_x + (\Delta V_{ANT} + V_N)^2{}_y]/\tau^2. \tag{7}$$

Suppose we insert a sequence of calibration test pulses with the short duration $\Delta t$ at times $t_1, t_2, t_3 \ldots t_n$ and search for the single pulse detector response only at times $t_1 + \Delta t$, $t_2 + \Delta t$, $t_3 + \Delta t$, $\ldots$ $t_n + \Delta t$. It is found for pulses which would increase the energy of the normal mode from zero to $kT$ that algorithm (7) gives a larger amount of response pulses exceeding thresholds, than algorithm (6). Perhaps this is the reason that algorithm (7) is preferred by a number of groups.

However, a study of chart records shows that algorithm (7) produces single response pulses for each test pulse while algorithm (6) may produce a sequence with more than 20 pulses following insertion of a single test pulse, many of them large enough to cross thresholds. This is a consequence of occurrence of the term $\Delta(V_{ANT}V_N)$ in (6). The single pulse excites the antenna and $V_{ANT}$ remains large for the antenna relaxation time. The rapidly varying wide band noise $V_N$ then produces the sequence of large pulses. This does not occur in (7) because $\Delta V_{ANT}$ instead of $V_{ANT}$ is combined with $\Delta V_N$. For very weak signals the term $2V_{ANT}\Delta V_{ANT}$ may be important for (6).

In one series of observations 50 single $kT$ pulses were introduced at 2-minute intervals. One hundred and ninety-two response pulses exceeding threshold set at 5 per minute, were emitted by the receiver for algorithm (6) in consequence of the proliferation process. (Weber in Shaviv and Rosen (1975), pp. 245–246)'.

## References

Astone, P., Bassan, M. and Bonifazi, P. *et al.* (1993), 'Long-term Operation of the Rome "Explorer" Cryogenic Gravity Wave Detector', *Physical Review D* **47**, 362–375.

Bogen, J. and Woodward, J. (1988), 'Saving the Phenomena', *The Philosophical Review* **97**, 303–352.

Bowen, D. R., Mann, A. K. and McFarlane, W. K., *et al.* (1967), 'Measurement of the $K^{+}_{e2}$ Branching Ratio', *Physical Review* **154**, 1314–1322.

Christenson, J. H., Cronin, J. W. and Fitch, V. L., *et al.* (1964), 'Evidence for the $2\pi$ Decay of the $K^{\circ}_{2}$ Meson', *Physical Review Letters* **13**, 138–140.

Collins, H. (1985), *Changing Order* (London: Sage Publications).

Damour, T. and Taylor, J. H. (1991), 'On the Orbital Period Change of the Binary Pulsar PSR 1913+16', *The Astrophysical Journal* **366**, 501–511.

Douglass, D. H., Gram, R. Q. and Tyson, J. A., *et al.* (1975), 'Two-Detector-Coincidence Search for Bursts of Gravitational Radiation', *Physical Review Letters* **35**, 480–483.

Franklin, A. (1986), *The Neglect of Experiment* (Cambridge: Cambridge University Press).

Franklin A. (1990), *Experiment, Right or Wrong* (Cambridge: Cambridge University Press).

Franklin, A. (1993a), *The Rise and Fall of the Fifth Force: Discovery, Pursuit, and Justification in Modern Physics*. New York: American Institute of Physics.

Franklin, A. (1993b). 'Discovery, Pursuit, and Justification', *Perspectives on Science* **1**, 252–284.

Galison, P. (1987), *How Experiments End*. Chicago: University of Chicago Press.

Garwin, R. L. (1974), 'Detection of Gravity Waves Challenged', *Physics Today* **27**, 9–11.

Hulse, R. A. and Taylor, J. H. (1975). 'A Deep Sample of New Pulsars and Their Spatial Extent in the Galaxy', *The Astrophysical Journal* **201**, L55–L59.

Levine, J. L. and Garwin, R. L. (1973), 'Single Gravity-Wave Detector Results Contrasted with Previous Coincidence Detections', *Physical Review Letters* **31**, 176–180.

Levine, J. L. and Garwin, R. L. (1974). 'New Negative Result for Gravitational Wave Detection and Comparison with Reported Detection', *Physical Review Letters* **33**, 794–797.

Lynch, M. (1991), 'Allan Franklin's Transcendental Physics', in A. Fine, M. Forbes and L. Wessels (eds) *PSA 1990*, Vol. 2 (East Lansing, MI: Philosophy of Science Association), 471–485.

Shaviv, G. and Rosen, J. (eds) (1975), *General Relativity and Gravitation: Proceedings of the Seventh International Conference (GR7), Tel-Aviv University, 23–28 June, 1974* (New York: John Wiley).

Taylor, J. H. and Weisberg, J. M. (1989), 'Further Experimental Tests of Relativistic Gravity Using the Binary Pulsar PSR 1913+16', *The Astrophysical Journal* **345,** 434–450.

Weber, J. (1975), 'Weber Responds', *Physics Today* **28,** 13.

Weber, J., Lee, M. and Gretz, D. J., *et al.* (1973), 'New Gravitational Radiation Experiments', *Physical Review Letters* **31,** 779–783.

Weisberg, J. M. and Taylor, J. L. (1984), 'Observations of Post-Newtonian Timing Effects in the Binary Pulsary PSR 1913+16', *Physical Review Letters* **52,** 1348–1350.