SLAC-PUB-12051

Blind Analysis in Nuclear and Particle Physics

Joshua R. Klein
Department of Physics
University of Texas

Aaron Roodman
Stanford Linear Accelerator Center
Stanford University

Abstract

Over the past decade or so, blind analysis has become a widely used tool in nuclear and particle physics measurements. A blind analysis avoids the possibility of experimenters biasing their result toward their own preconceptions by preventing them from knowing the answer until the analysis is complete. There is at least circumstantial evidence that such a bias has affected past measurements, and as experiments have become costlier and more difficult and hence harder to reproduce, the possibility of bias has become a more important issue than in the past. We describe here the motivations for performing a blind analysis, and give several modern examples of successful blind analysis strategies.

Submitted to Annual Reviews of Nuclear and Particle Science

Work supported in part by Department of Energy contract DE-AC02-76SF00515

1 Introduction

Hans von Osten, who lived at the beginning of the twentieth century, could do math. Given a pair of single digit numbers written on a blackboard, Hans could add them together correctly nearly all of the time. What was remarkable about this ability was the fact that Hans—often called 'Clever Hans'—was a horse. He would demonstrate his skill by pawing the ground with his hoof until he had reached the sum of the two numbers, while those who had presented the problem looked on. Critics of Hans's ability tried to determine whether his trainer was providing him with signals, but could find none. Eventually, they asked the trainer to leave the room, but still Hans managed to add the numbers correctly more often than not. The mystery of Hans's ability was only solved in 1907 when the psychologist Oskar Pfungst proposed that a trial be done in which no one in the room with Hans knew both of the numbers presented. With all the observers 'blind' to the answer, Hans was unable to produce a correct result. The conclusion was that Hans was indeed clever: he had been using subtle non-verbal cues from those in the room—cues his observers were not even aware they were providing—to decide when to stop pawing the ground.

The 'Clever Hans Effect'[1] has left its impression on modern science, particularly on medicine. Most large-scale clinical trials of new drugs not only require that the patients be unaware of whether they are receiving a placebo or not, but that those who administer the drug also be kept blind as to which patients are in the control sample (the trials are thus 'doubly blind')[2].

By contrast, throughout most of their history nuclear and particle physics have run 'open' experiments, in which an estimate of the final answer is known well before the analysis is complete. Adjustments to cuts, measurements of backgrounds and acceptances, and evaluations of systematic uncertainties are routinely made with full knowledge of the current value of the intended measurement and the effects any changes have on the measurement. Such an approach makes a great deal of sense in a physics experiment, as it allows us to bring to bear some of our most commonly used techniques: we check that the answer makes sense; we give particular scrutiny to results which contradict established models, previous measurements, or conventional wisdom; we are conservative in our estimates of systematic uncertainties so as not to mislead others about the significance of our result.

In his 1932 paper presenting the results of his measurement of the electron charge to mass ratio e/m [2], Frank Dunnington concludes with a warning born of his own experience applying these common sense principles:

[1]The effect of experimenter expectations on the behavior of subjects is more often referred to as the 'Hawthorne Effect', named after a factory at which worker productivity was being studied. The productivity was a strong function of what the experimenters indicated they thought would be important (for example, light levels)

[2]Nevertheless, in some extreme cases, the results of a clinical trial have been used by the experimenters as an unacknowledged criterion for their publication: no one wants to publish the fact that their new drug does not work. In such cases, the trial may be repeated until a desired result is obtained, thus leading to a bias in the published data. The problem has so concerned the medical field that the International Committee of Medical Journal Editors [1] now formally requires all clinical trials to first be registered in a public trials registry, *before* the trial begins. With such registration, the investigators essentially commit to publishing the results of their experiments, regardless of whether the outcome is favorable to them or their sponsors.

> It is also desirable to emphasize the importance of the human equation in accurate measurements such as these. It is easier than is generally realized to unconsciously work toward a certain value. One cannot, of course, alter or change natural phenomena...but one can, for instance, seek for those corrections and refinements which shift the results in the desired direction. Every effort has been made to avoid such tendencies in the present work.

At least part of Dunnington's effort 'to avoid such tendencies' was to keep himself from actually knowing the results of his measurement until he was finished with it. To do this, he kept the value of the angle between the electron source and his detector hidden from himself, by asking his machinist to build something close—but not exactly at—the 340° that was needed [3]. Without exact knowledge of this angle, he could not make the final calculations which would change his data into a measurement of e/m.

Is there any evidence over the history of nuclear and particle physics that experimentalists, in Dunnington's words, 'unconsciously work toward a certain value'? It would be almost impossible to say definitively, since we of course do not know what any particular experimentalist was thinking (unconsciously or otherwise) during the process of a measurement. We might expect that if such a bias exists, then over time new measurements would tend to agree better with prior measurements than with their modern and much more precise value.

Perhaps the classic example of a measurement suspected of experimenter's bias is the speed of light. Measurements of the speed of light span an entire century, with improvements of five orders of magnitude in the experimental uncertainties. A summary of measurements[4], using a variety of techniques, made prior to 1960 is shown in Fig 1, along with the much more accurate value from later results using a methane absorption line frequency[5]. One striking feature is the 17 km/sec shift between the series of experiments from 1930-1940 and later determinations. A fascinating post-mortem on the systematic uncertainties in these experiments[6], noting the different techniques used in the four "low" results, speculates about one of the sources of bias:

> the investigator searches for the source or sources of such errors, and continues to search until he gets a result close to the accepted value. *Then he stops!*

We have done a brief investigation of this issue in more modern particle physics experiments, using a selected (and hence biased!) set of historical measurements typically compiled by the Particle Data Group (PDG). The PDG's own history plots, which they have published from time to time since 1975 as part of their Reviews of Particle Physics, depict only what the PDG itself published in each year. Their numbers are typically averages over several measurements, and so the published values over time are by construction correlated with one another. They therefore do not by themselves necessarily indicate that there is any bias in the data. We have looked instead at the individual measurements, compiled by the PDG or by other reviews, and compared them to the published values which existed at the time of the measurement.

Figure 2 shows the four measurements we have examined: the neutron lifetime, the $K^0_{\rm S}$ lifetime, the mass of the Λ, and the value of the ratio $g_{\rm A}/g_{\rm V}$ determined from neutron β

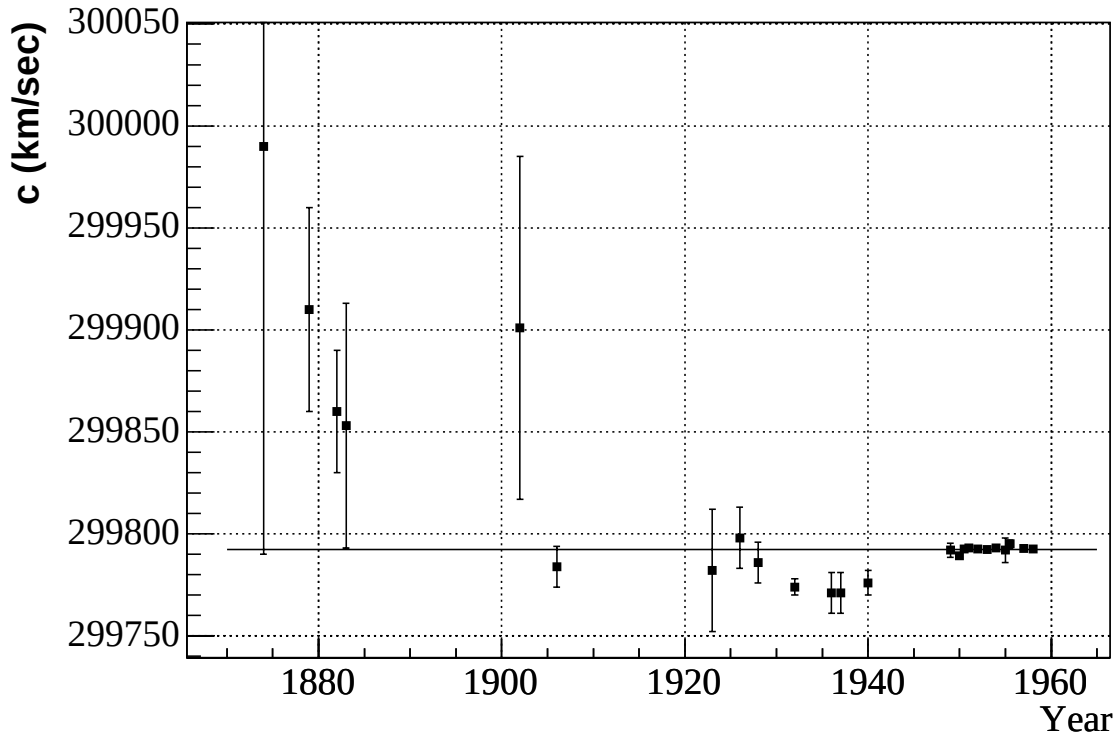

Figure 1: Summary of speed of light measurements. The line indicates the ultimate experimental value. Among other interesting features, the series of four measurements from 1930-1940 displays a $17 km/sec$ systematic shift from the true value.

decay [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 7, 8]. The measurements are shown as open dots, with the error bars depicting the uncertainties published along with the measurement. The published average which existed at the time each measurement was made (that is, not including the measurement itself) is represented by the dashed curves, where the spacing between the curves represents the published 1σ uncertainty on those averages. The horizontal dotted lines show the 2004 PDG averages [40].

Although the effect is not striking, the measurements do tend to cluster nearer the prior published averages than the 'final' value. Grouping all the data points together, the χ^2 for the hypothesis that the measurements are normally distributed around the prior averages is 131.2 for 83 degrees of freedom, while the χ^2 for the hypothesis that they are normally distributed about the final average is 249.7 for ~ 82 degrees of freedom.

Even if Figure 2 showed a strong correlation between the measurements and previously published averages, that would not necessarily mean that the experimental results were biased by experimenters' concerns about contradicting conventional wisdom. Many other explanations could account for the behavior—common techniques that were used and were later found to have systematic effects which had been neglected, for example, or physical corrections which were unknown for many years and which would have shifted all the values nearer the modern averages. Other examples often cited as possible evidence of bias in measurements are the unusually low χ^2 values found in global fits to data sets, such as the the average B meson lifetime $\chi^2 = 4.5$ for 13 degrees of freedom [3], or global fits to solar neutrino data which have for example $\chi^2 = 70.2$ for 81 degrees of freedom [41].

Although we cannot say conclusively whether bias has influenced measurements in nuclear and particle physics, the way to avoid even the possibility is to follow Dunnington's and Pfungst's examples and perform measurements while staying blind to the value of our answer. Blind analysis in nuclear and particle physics experiments has its modern origins around 1990 with experiment E791 at Brookhaven National Laboratory, a search for the rare decay $K_L \to \mu e$, although the idea had been discussed at least 10 years earlier [42]. As a rare process experiment, E791 had good motivation to use a blind analysis: a potential discovery could easily be missed if they allowed flexibility in their final cuts to remove events 'suspiciously' close to the edge of the signal box. We discuss this kind of 'hidden signal box' technique in more detail in Section 3.1.

The number of experiments which analyze their data blindly has grown steadily since E791's example, and the approaches to how to successfully do a blind analysis are as varied as the experiments themselves. In the next sections we describe the fundamental philosophy behind blind analysis, and detail examples of experiments which have published blind results. We have restricted our discussion only to techniques aimed at avoiding the kind of unintentional bias which concerned both Clever Hans's critics and Dunnington. We do not consider here intentional bias or the bias resulting from systematic effects in instrumentation or technique, none of which can be removed through blind analysis techniques.

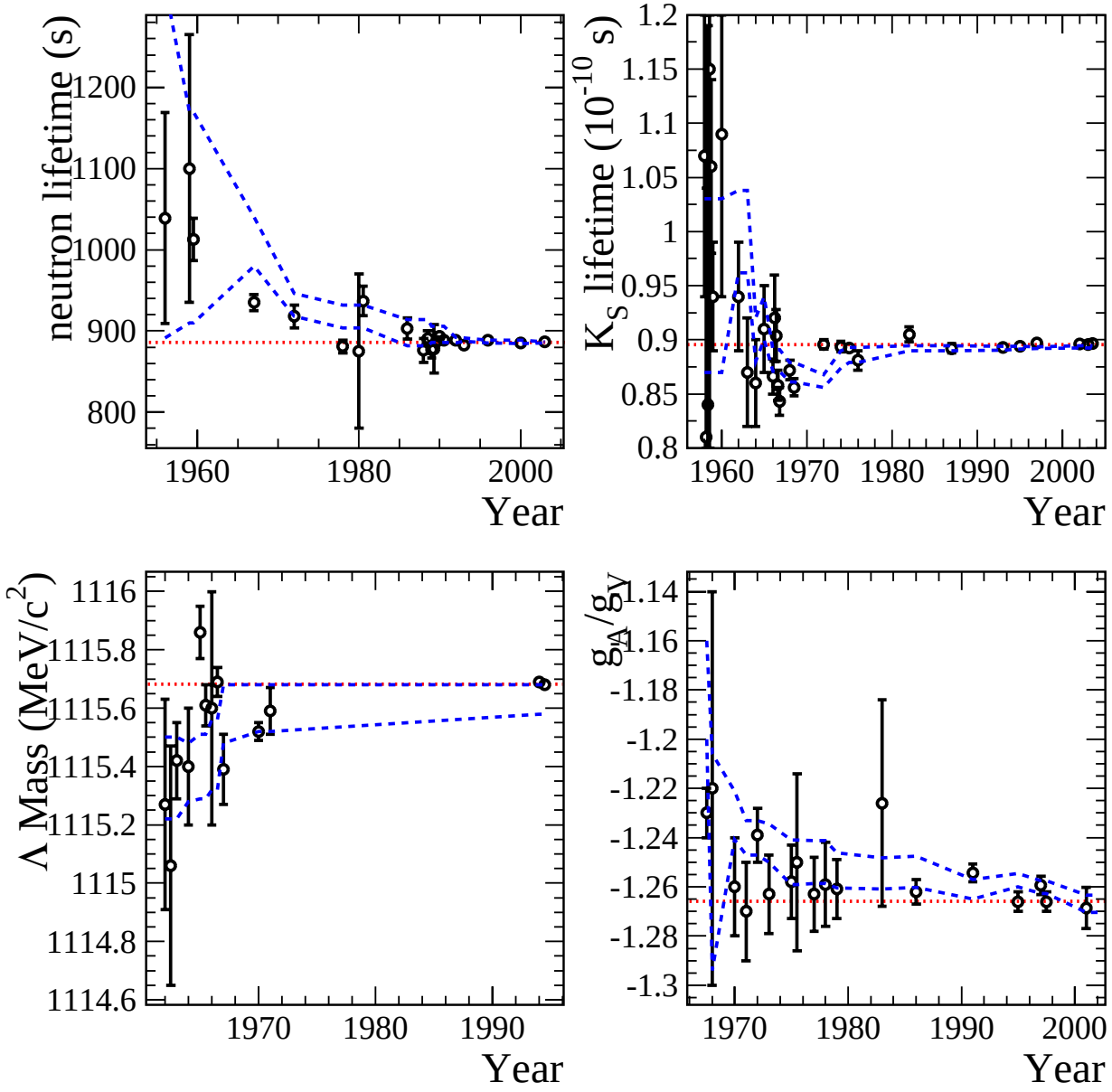

Figure 2: The history of four measurements compared to published averages before each measurement was made (dashed lines) and the currently accepted value (dotted lines).

2 Experimenter's Bias and the Motivations for Blind Analyses

The Oxford English Dictionary defines bias as *A systematic distortion of an expected statistical result due to a factor not allowed for in its derivation.* In a nuclear or particle physics measurement there are many potential sources of bias: the trigger for the experiment, the algorithms used to reconstruct events or extract signals, or the particular instrumentation used. The accepted procedure, when such biases cannot be directly measured or eliminated, is to estimate their size and include the estimates as systematic uncertainties on the measurement. *Experimenter's bias* differs from these biases because its source is the human being making the measurement, who may (in Dunnington's words) "unconsciously work toward a certain value". The bias may be in the direction of previous measurements, prior theoretical expectations, or some other preconception. The crucial difference between experimenter's bias and any other bias in a measurement is that the size of an experimenter's bias cannot be estimated. Thus the only available approach is to use a methodology that prevents or suppresses it.

Experimenter's bias can creep into a measurement in several ways. The first scenario is subtle, in that the biases it may produce are of the order of the measurement's statistical uncertainty. In general, the data used for a measurement is isolated with a series of selection requirements, or cuts. While the value of these cuts on a particular quantity may be chosen to maximize the sensitivity of the result, very often there is a wide plateau in the value of the cut, over which the quality of the result varies little. The cartoon shown in Fig. 3 illustrates this point. In the case shown, the value of the cut may be chosen arbitrarily within the sensitivity plateau. The numerical value of the result, however, may vary as a function of the cut value, especially for cuts that significantly alter the signal efficiency or background contamination. Such variations will be statistical, with a magnitude on the order of the statistical uncertainty of the measurement. If the value of the cut is chosen with the knowledge of how that value affects the final answer, then the measurement can be biased toward the expected (or desired) result.

It does not take much of a statistical bias to produce a surprisingly large signal. A typical analysis sensitive to statistical bias might be the search for a peak in an invariant mass distribution. If the cuts are chosen in a biased way, then the actual fraction of events accepted in the region of a potential signal may be artificially high. Consider m sequential cuts with true acceptance A_i and a biased fraction of events in the signal region A_i'. The significance, defined as $S = N_{\rm signal}/\sqrt{N_{\rm background}}$, is given by

$$S = (\prod_{i=1}^{m} \frac{A_i'}{A_i} - 1) \times \sqrt{N \prod_{i=1}^{m} A_i}. \qquad (1)$$

where N is the number of total number of events in the signal region. As a numerical example, if an analysis begins with $N = 2500$, and uses a set of 10 cuts each with acceptance $A_i = 0.9$, a geometric average bias of just 1% ($(\prod_{i=1}^{m} A_i'/A_i)^{1/m} = 1.01$) will lead to an apparent signal above background of roughly 3 σ.

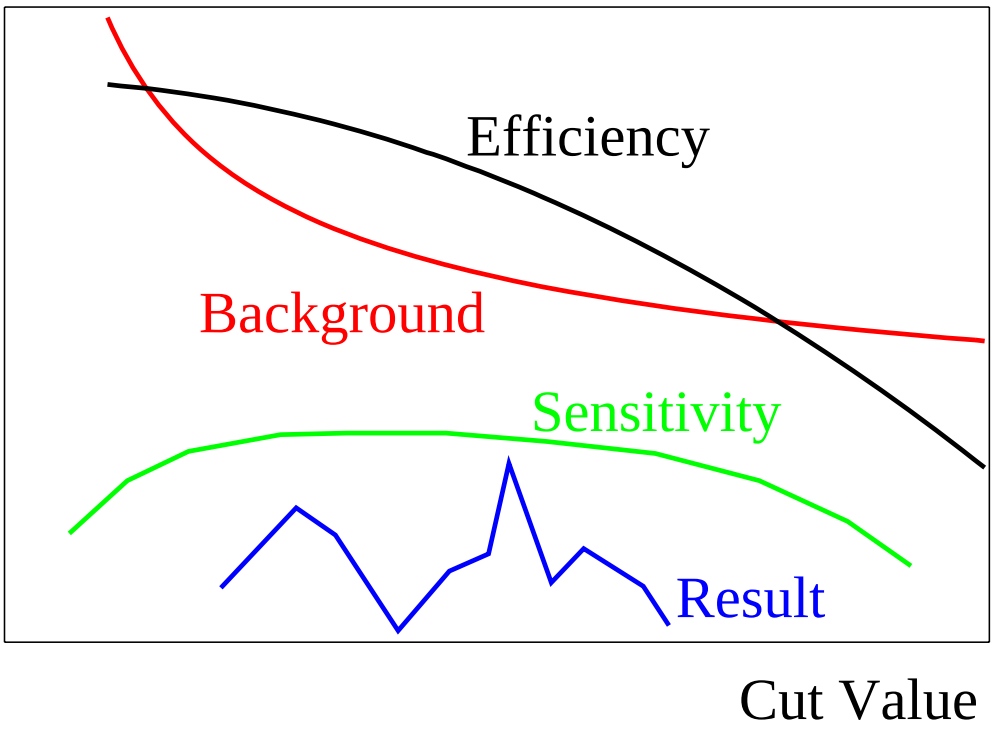

Figure 3: A cartoon demonstrating the variation in the central value of a measurement due to fluctuations even when the cut acceptance for signal is reasonably flat.

Even though the cuts may be asymptotically unbiased—applied to an infinite data set, $A_i' = A_i$—by having been chosen based in part on how many more apparent signal events they accept in the signal region in *this data set*, they can still be biased. Put another way, if an ensemble of experiments with the same number of events is analyzed using an identical set of cut values (tuned perhaps, on just one of the experimental data sets) then the mean number of observed signal events will tend toward the true value (which may be zero). But if the ensemble of experiments is analyzed and in each experiment a new value of the cuts can chosen based on the observation of events in the signal bin, the mean number of observed events in the signal bin can always be larger than the true number of signal events.

Of course in practice *finding* a set of cuts whose geometric average bias is as large as 1% is not necessarily easy. For example, for the first of the $A_i = 0.9$ cuts, the variance on the number of accepted events is less than 1%, and so either one needs to work hard to tune the cuts, or else another factor is at work, such as an initial statistical fluctuation upward, or one is operating in a particularly sensitive region of the analysis where small variations in the cut position for background events lead to much larger variations in the number of accepted events in the signal region.

The next bias scenario, typically involving the search for rare processes or decays, is much less subtle. Experiments searching for small signals, at the edge of detectability in statistics

or above backgrounds, are especially dependent on the exact values of the selection cuts. If the values of the cuts are chosen with the knowledge of which events are included or excluded, the results may be biased toward either observation or elimination of a signal. Such choices can be easy to make if each event is examined individually. Nearly every observed event can be found to have something unusual about it, and so can be included in a signal sample (sometimes a sample of one event) or excluded as an unexpected background. In one extreme, cuts chosen to remove individual events will yield a better upper limit than is deserved. In the other extreme, cuts chosen to retain individual events may produce a signal where none is warranted. This *selection* bias is perhaps the most dangerous. produced by

The last way in which experimenter's bias may affect a measurement is in the decision that a measurement has been completed, as the warning from the speed of light study indicates. Galison[43] notes in his historical study *How Experiments End*: "... there is no strictly logical termination point inherent in the experimental process," instead "the decision to end an investigation draws on the full set of skills, strategies, beliefs, and instruments at the experimentalist's disposal." If the decision to stop analyzing and publish relies on the value of the result—in particular how close it adheres to the experimentalist's preconceptions—the result may be biased toward the preconceived value. The danger of continuing the data analysis, finding mistakes or improving the analysis, until the result agrees with expectations, is well known. This *stopping* bias may affect any kind of measurement and may be a small effect or a large one. It is also probably the most common kind of bias found in nuclear or particle physics.

A *blind analysis* is a method that hides some aspect of the data or result to prevent experimenter's bias. There is no single blind analysis technique, nor is each technique appropriate for all measurements. Instead the blind analysis method must carefully match the experiment, both to prevent experimenter's bias and to allow the measurement to be made unimpeded by the method. There are several blind analysis methodologies that will be described in this review, each appropriate for a certain kind of measurement. These methods can be grouped according to exactly what is kept hidden in the measurement:

1. the signal events, when the signal occurs in a well defined region of the experiment's phase space
2. the result, when the numerical answer can be separated from all other aspects of the analysis
3. the number of events in the data set, when the answer relies directly upon their count
4. a fraction of the entire data set

While blind analysis techniques may not be feasible or necessary in all measurements, as a general rule the possibility of experimenter's bias should be considered in all experiments. Typical objections to the use of blind analysis techniques are that it slows the pace of data analysis, that certain aspects of the analysis become difficult, and that unexpected phenomena can only be found by full exploration of the data. The latter issue is a serious one. Consider the following anecdote from a well-known physicist[44]:

> While looking for the decay $\pi^+ \to e^+\nu_e$, we focused all our attention on reducing backgrounds, since a prior experiment had set a limit at the level of 10^{-6} on the branching ratio. When we heard that an experiment at CERN had seen a signal around 10^{-4} I switched from delayed to prompt. The signal was right there, and could have been seen on the first day.

While some blind techniques are susceptible to this pitfall, not all are. In such cases, a method allowing a full exploration of a data subsample would have not missed such a large signal.

In general, it is crucial that the blind analysis technique be designed as simply and narrowly as possible. A good method, appropriately used, minimizes delays or difficulties in the data analysis. In some cases, a blind analysis may delay certain aspects of an analysis until the blind procedure has been removed, or *unblinded*. For example, certain cross-checks may only be possible after the blind procedure has been removed. The trade-offs involved must be considered according to the individual merits of each case. However, by blinding only a very narrow aspect of the analysis, the methods described in this review minimize the scope of the data analysis needed after unblinding. In general, for the examples described in this review, the pace of data analysis has only been slowed to the extent that individual data analysts have worked to check their measurement more carefully before unblinding their result. We note that the none of the blind techniques we describe here—and perhaps no blind technique—can be applied to an analysis in which backgrounds are cut or signals identified by event-by-event human inspection.

It is also important to realize that blind analyses solve one and only one problem, the influence of experimenter's bias on the measurement. Other biases in the measurement caused by the general approach or the instrumentation are not avoided by any of the techniques we describe here. For such biases, either a correction based on a measurement or the inclusion of the bias as a systematic uncertainty still needs to be made, whether the analysis has been done blindly or not.

All sizeable collaborations have internal data analysis and publication review processes. A blind analysis, and the associated division of the data analysis into a *blind* and an *unblind* phase, gives the collaboration at large an opportunity to review the work before the transition of looking at the answer. Today large collaborations are grappling with the issue of vetting many publications for both quality and correctness. Collaborations can require that the decision to unblind a result be made as a part of the internal review process, with the consultation or approval of a wider sub-group, and not by the data analysts alone.[45] This procedure improves the effectiveness of the internal review.

The last issue to confront in using blind analyses as a technique is what do to if the analysis strategy breaks-down. For example, what should an experiment do if, after all selections cuts have been set, the events in the nominal signal region are clearly background, and additional selections to remove such background were simply omitted? It is not necessary in the blind analysis approach to insist that, because an analysis was done blindly, no additional selections may be applied. Ideally, an experiment should consider such situations in advance, to prepare for such cases. One useful principle that may be adopted is that the publication

simply describe the full analysis procedure, in this case to be explicit about which selections were applied after the unblinding. The blind analysis method does not require that data analysis stop after unblinding, nor does it ensure that the results of the analysis are correct. There is no reason to publish a result known to be wrong, just because the analysis was done blindly.

Multiple, independent analyses are occasionally suggested as a way to prevent experimenter's bias. While independent analyses can be a powerful tool for preventing errors in a measurement, in our opinion blind analyses prevent experimenter's bias much more directly than redundant analyses. The two methods can, however, easily be used together.

As experiments have become larger, lengthier, more expensive, and therefore harder to reproduce, the issue of whether to do a blind analysis has perhaps become more important than it was a few decades ago. In some cases, a biased result may stand for many years, possibly causing the expense of a new experiment to be built, or leading theorists and experimentalists down paths which are not productive. Without the luxury of having new and important results verified quickly, the assurance that experimentalist's bias does not contribute to the many other possibilities for error is generally worth the additional time and effort.

3 Blind Analysis Methods

3.1 Hidden Signal Box

Perhaps the most straightforward blind analysis method is the *hidden signal box*. In this technique, a subset of the data, containing the potential signal, is kept hidden until all aspects of the analysis are complete. Often the signal region is defined in terms of two experimental parameters, chosen to separate the signal from backgrounds, and this two-dimensional signal region comprises a *signal box*. Only after the data selection requirements, the signal efficiency, and the estimated background are determined is the hidden signal box opened.

This method is very well suited to measurements searching for rare signals, as long as two criteria are met. First, the signal characteristics and location must be known. In rare decay searches, such as $K_L \to \mu^\pm e^\mp$ or $B^0 \to \mu^+\mu^-$, the signal may be simulated, the efficiency determined, and an appropriate hidden box defined using the invariant mass and whatever other relevant kinematic variable. Second, the experiment must be able to independently estimate the size of the background expected in the signal box. Ideally, this may be accomplished by understanding the source of background events near the signal box, and extrapolating from this sideband region into the hidden signal box. In particular, the background estimate cannot depend on the characteristics of any events that may be inside the signal box. With these conditions, the dependence of the signal to background ratio on the selection requirements is known, and the cuts may be optimized as desired, again without reference to the events in the hidden box.

The hidden box method was first used in a search for the rare decay $K_L \to \mu^\pm e^\mp$ by the E791 experiment at BNL.[46]. They formed a hidden signal box in a region of the invariant mass, $M_{\mu e}$ and the momentum transverse to the kaon beam direction, P_T^2, as shown in Fig. 4. Also visible are a population of background events at lower $M_{\mu e}$, primarily from $K_L \to \pi e \nu_e$ decays where the pion decays in flight. The many cuts applied to remove backgrounds were optimized using the sideband region, $P_T^2 > 144\,\mathrm{MeV}/c^2$, and the signal box was not opened until the cuts were determined. No events were observed, so an upper limit was set on this lepton number violating process.

The hidden signal box method is now a standard technique for searches of rare decays from known particles, and has been used by many particle physics experiments (see, for example, Ref. [47]). In most rare decay searches, the above criteria are generally satisfied, and we recommend that this blind analysis method always then be used. Rare decay searches are distinguished from counting experiments where a sizable signal is present, because even if a small signal is observed, it is generally insufficient to constrain or verify the expected signal characteristics. Thus the expected signal must be characterized by simulation in any case, and a blind analysis causes few extra difficulties. The dividing line between a search for a rare process and a branching fraction or cross-section measurement is a judgment for each experiment. The important consideration is whether or not the signal events themselves must be used to ensure that the experimental efficiency and background rejection are well

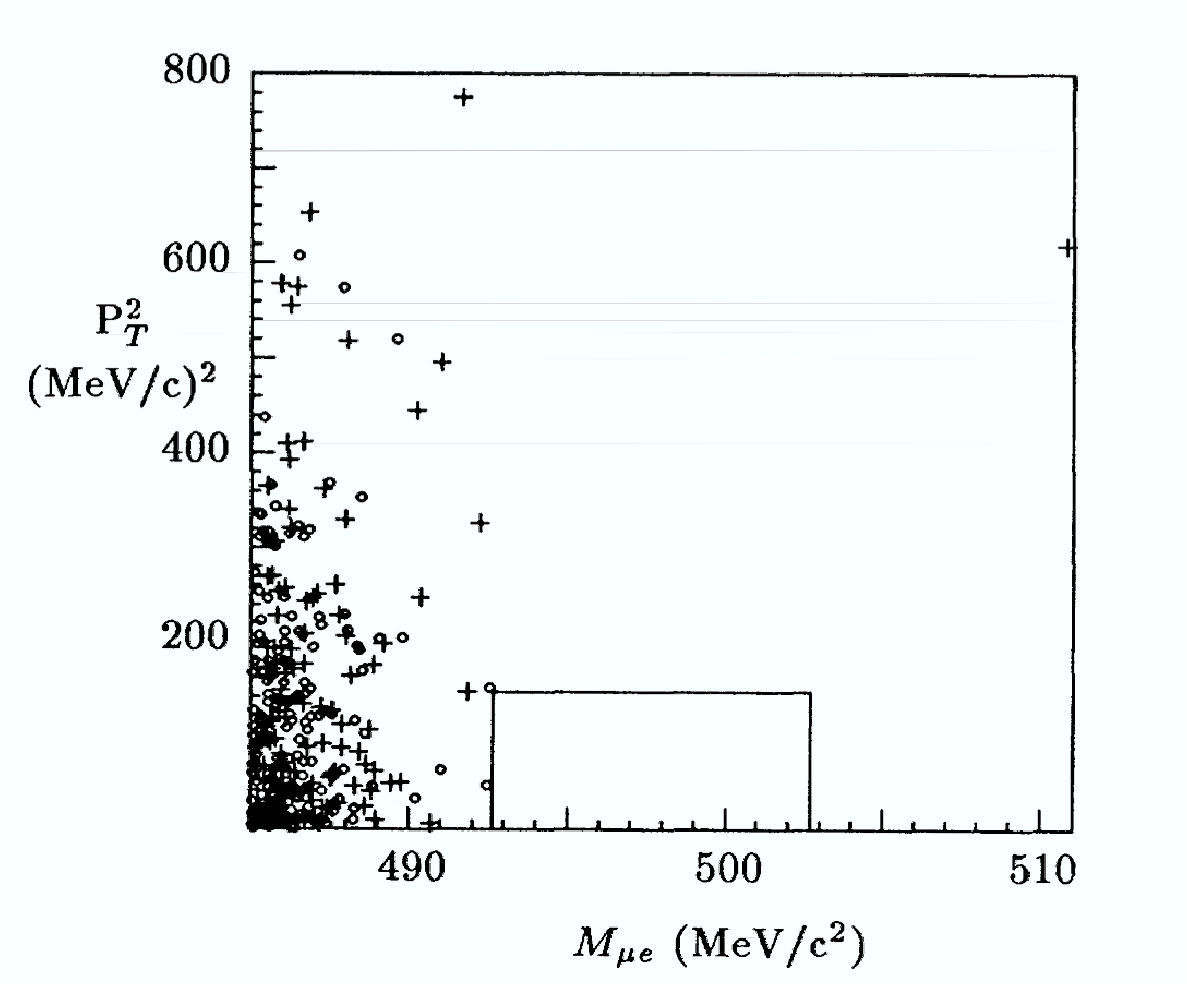

Figure 4: The P_T^2 vs. $M_{\mu e}$ distribution of events from the E791 experiment[46]. Also shown is the hidden signal box used for the blind search.

understood; if so then the hidden signal box method is no longer appropriate, and one of the alternative methods described in Sections 3.4 may be used instead.

There are several other considerations when using the hidden signal box technique. First, the hidden box should be chosen somewhat larger than the anticipated signal region. Since the final signal box may be made smaller than the hidden box, but not larger, this allows the exact size of the signal box to be optimized during the course of the blind analysis, to maximize the signal to background ratio, for instance. Next, in general some background may be in the signal box, and the number of background events should be predicted from data sideband regions in advance of opening the box.

If no events lie in the signal box, as in E791, the conclusion is straightforward. The presence of events in the box does require additional effort, however. If there is a significant signal, or if the events in the signal box are consistent with the estimated background, the results follow directly. Nevertheless, several ambiguous situations may occur. It is possible that the events in the box are clearly due to backgrounds, and remained in the signal box due to the unfortunate omission of certain cuts. Two options are available in this circumstance: to retain the events, treating them as background, and deriving a limit accordingly; or to remove them by applying the omitted cut, but also assessing the statistical impact of the entire procedure on the result. Ideally, experiments will consider in advance which procedure to use. The next quandary may occur if there are more events in the signal box than expected from backgrounds, but the events are very inconsistent with the expected signal properties. The use of a blind analysis does not require that events in the signal box must absolutely be treated as signal; only that the cuts may not be adjusted to reject, or include, individual events. The excess events may be interpreted as background, although *a posteriori* estimation of the signal probability for individual events is fraught with difficulty, and needs to be done rather carefully. Some experiments use a hidden signal box in conjunction with an unbinned maximum likelihood fit to the number of signal and background events; such fits incorporate directly the signal or background probability for each event and hence avoid this issue.

3.2 Hidden Answer Methods

Measurements where all, or the majority, of the data analysis can be separated from the numerical value of the result are most amenable to the *hidden answer* blind analysis technique. The e/m measurement of Dunnington[2] illustrates this technique. None of the data analysis, evaluations of corrections, or other features of the e/m measurement were dependent on the unknown angle in his spectrometer, only the final result. The separation between a narrowly constructed hidden feature and the bulk of the measurement permits a blind analysis with little risk or difficulty.

In general the hidden answer method works best for experiments measuring a single precise parameter, when that parameter does not depend directly on the number of observed events.

3.2.1 Hidden Detector Parameters

Most modern nuclear and particle physics experiments are too complex to keep their physics results hidden as Dunnington did. A single detector parameter is unlikely to be enough to keep the physics answer hidden, and very often these parameters need to be measured through calibration runs which cannot wait until the analysis is complete.

One area of fundamental physics in which a hidden detector parameter approach does work is the laboratory study of gravity. These experiments are often still performed in the kind of laboratory environment in which *ex situ* measurements of the apparatus can determine the final answer. Experimental tests of the gravitational inverse-square law [48], like those done at the University of California at Irvine's Laboratory for Gravitation Research in 1985, employ this kind of hidden parameter approach to hide the physics answer.

Like most laboratory measurements of gravity, the Irvine group measured the torque exerted on a torsion balance by a set of test masses. One pair of 7.3-kg masses (the 'far masses') was positioned 105 cm away, while a 43-g 'near mass' was positioned 5 cm away. They chose the masses and positions so that when they were moved to opposing positions, the change in torque predicted by Newtonian gravity was nearly zero. The Irvine group aimed for a precision which would allow them to test deviations from Newtonian gravity as small as one part in 10^4.

Their measurement relied on precise knowledge of many different detector parameters—the dimensions of the torsion balance and test masses, the positions of the test masses, and of course the masses of all test components. To prevent themselves from selecting data in a biased way, or from (in their words) "slackening of analysis effort" when their answer began to meet their expectations (what we have called a stopping bias), they kept the value of their near mass known only to 1%—the exact mass known only to someone outside their collaboration. They used the true value of the mass only when they had completed the analysis and were ready to report their initial results. Subsequent improvements to the analysis were made and later published, but they nevertheless published the measurement made before these improvements were made.

3.2.2 Hidden Offset

The *hidden offset* method inserts an unknown numerical offset into the data analysis, so that the true measured value is hidden from the experimenters. This method was first used in the measurement of the direct CP violation parameter ϵ'/ϵ by the KTeV collaboration.[49]

Direct CP violation is measured in neutral Kaon decays using the double ratio of decay rates for K_S and K_L into charged $\pi^+\pi^-$ and $\pi^0\pi^0$ final states, according to the expression

$$\frac{\Gamma(\mathrm{K}_L \to \pi^0\pi^0)/\Gamma(\mathrm{K}_S \to \pi^0\pi^0)}{\Gamma(\mathrm{K}_L \to \pi^+\pi^-)/\Gamma(\mathrm{K}_S \to \pi^+\pi^-)} = 1 - 6Re(\epsilon'/\epsilon) \tag{2}$$

In practice, KTeV fit its data, in kaon energy bins, to extract a value for ϵ'/ϵ, as well as other parameters relevant to the experiment. The aim for KTeV was to determine ϵ'/ϵ with a precision of $1 - 2 \times 10^{-4}$. This required both a very large sample of kaon decays, including of order six million $\mathrm{K}_L \to \pi^0\pi^0$ events, as well as exquisite control over systematic uncertainties. For instance, KTeV made an acceptance correction to the ratio of observed K_L to K_S events, derived from simulation, of roughly 10% which had to be understood at the 1×10^{-3} level or better.

In addition, two prior experiments had measured ϵ'/ϵ to a precision of around $\sigma_{\epsilon'/\epsilon} \sim 7\times10^{-4}$, but differed by roughly 2.5σ. Theoretical estimates ranged from a few 10^{-4} to perhaps 15×10^{-4}. Therefore, KTeV used a blind analysis to prevent any experimenter's bias in what is a difficult and systematically sensitive measurement.

KTeV used a hidden offset directly in its ϵ'/ϵ fit. Instead of fitting for the value of ϵ'/ϵ, the fit used

$$\epsilon'/\epsilon\,(\text{Hidden}) = \left\{ \begin{array}{c} 1 \\ -1 \end{array} \right\} \times \epsilon'/\epsilon + C \tag{3}$$

where C was a hidden random constant, and the choice of 1 or -1 was also hidden and random. KTeV relied on extensive comparisons of data and simulation to make its event selections, detector simulations, acceptance corrections, and background subtractions. None of these were affected or impeded by the hidden offset in the fit to ϵ'/ϵ. In addition, direct but separate comparisons were made between the distributions and number of events in data and simulation for $\mathrm{K} \to \pi^0\pi^0$ and $\mathrm{K} \to \pi^+\pi^-$. The one comparison that could not be made was to form the double ratio that appears in the expression for ϵ'/ϵ. Fortunately, KTeV's method for ϵ'/ϵ almost completely separated the charged and neutral analyses, so that there was no impact from this restriction.

Both the hidden offset C and the sign choice were made by a pseudo-random number generator, with a seed chosen by the experimenters. The generator picked a value of C with a Gaussian distribution, centered at zero, with a width of around 60×10^{-4}. The $+1$ or -1 in the hidden value served to hide the direction ϵ'/ϵ changed as different corrections or selections were applied[50]. In practice, KTeV had to remove the sign choice at an earlier stage to permit a full evaluation of systematic errors. Nevertheless, the first KTeV ϵ'/ϵ result was unblinded only one week before the result was made public.

The pseudo-random distribution used for a hidden offset must be chosen with a little care.

A smooth Gaussian distribution with a width large enough to cover prior measurements and predictions is often a good choice. The addition of an unknown sign also hides the direction the result has moved with changes to the analysis. The hidden offset technique also permits multiple analyses within an experiment. In this case, the competing groups should begin with different seeds for the pseudo-random generation of the hidden offset C and the hidden sign. In this way the analyses are blind with respect to the final result and also with respect to each other. Since the motivation for multiple analyses is to provide a strong internal cross-check, this arrangement prevents the groups from comparing their results directly. To make comparisons, both analyses can switch to a common seed. At this stage, any problems with the internal cross-check may be addressed, without unblinding the final result.

3.2.3 Hidden Offset and Hidden Asymmetry

For many measurements, hiding the answer would be an appropriate approach to prevent a biased result, but is not sufficient for a blind analysis. In particular, the numerical result may be evident in certain experimental distributions that display the data. For example, the value of a lifetime may be inferred from the decay time distribution. A blind analysis is still possible, but more care is required in constructing it.

A good example of such a measurement is the observation of a *CP*-violating asymmetry in B-meson decays by the *BABAR* experiment at the PEP-II asymmetric-energy B factory. The *CP*-violating parameter $\sin 2\beta$ is measured by comparing the decay-time distribution for B^0 and $\overline{B}^0$ decays into *CP*-eigenstates, such as $J/\psi\, K^0_S$. The B flavor, B^0 or $\overline{B}^0$, is constrained by the flavor specific decay (or flavor tag) of the other neutral B-meson in the event. Before *CP*-violation had been observed, *BABAR* adopted a blind analysis to avoid the possibility of bias, especially with respect to the prior expectations around $\sin 2\beta \equiv 0.7$ from other weak-interaction measurements and the unitarity of the CKM matrix.[3]

The value for the *CP* asymmetry is determined in a complex unbinned maximum likelihood fit to the decay time, Δt, along with information about the flavor tag and the kinematics of the B^0 decay.[51]. As such, the hidden offset method, as described above, can easily be used to hide the value of $\sin 2\beta$. The hidden offset method by itself is not enough, however: one of the distributions that must be examined during the course of the analysis is the decay time itself. For example, to ensure that the maximum likelihood fit is done correctly, it is crucial that the probability density function (PDF) used to describe the decay-time is a good match to the data. In this case, the PDFs are determined using much larger samples of other exclusively reconstructed B decays, and simulation may be used to verify that the PDFs will apply for the rarer *CP* eigenstates. Nevertheless, the decay time distribution for the *CP* sample must still be examined. The problem for a blind analysis is that the decay time distribution shown separately for B^0 and $\overline{B}^0$ flavor tags, as in Fig. 5a, uncovers the asymmetry.

To solve this problem, *BABAR* used two extra restrictions in its blind analysis. The first restriction, used in the initial $\sin 2\beta$ measurements[51], was to hide the asymmetry in the time variable used in plots. The asymmetry is evident in two ways in the time distribution:

as a difference between B^0 and $\overline{B}^0$ flavor tags and as an asymmetry around $\Delta t = 0$. Both visible asymmetries can be obscured by using a hidden Δt variable, defined as:

$$\Delta t\,(\text{Hidden}) = \left\{ \begin{array}{c} 1 \\ -1 \end{array} \right\} \times s_{\text{Tag}} \times \Delta t + \text{Offset} \tag{4}$$

The variable s_{Tag} is equal to 1 or -1 for B^0 or $\overline{B}^0$ flavor tags. Since the asymmetry is nearly equal and opposite for the different B flavors, the asymmetry is hidden by flipping one of the distributions. The asymmetry of the individual Δt distributions around zero is hidden by the unknown offset. The result is shown in Figure 5b, where the remaining difference in curves is due to the charm lifetime not *CP* violation. While the asymmetric shape of the distribution is still visible in these curves, in an actual sample, limited by statistics, the asymmetry is effectively hidden by the statistical uncertainty of the mean.

Of course, in the actual fits to the data, the true Δt is used, not the hidden Δt. This technique allowed *BABAR* to look at the Δt distribution, but remain blind to any *CP* asymmetry. In addition, there was one extra restriction: the resulting Δt distribution from the fit could not be overlaid directly the data, since the smooth PDF would effectively unblind the asymmetry. Instead the residuals between data and the smooth PDF were used to assess the quality of the fit.

The *BABAR* experiment has regularly updated its measurement of sin2β as more data has been collected. By the third public result, a simpler method of hiding the visual asymmetry was adopted. Instead of using the hidden Δt variable, the only Δt distribution used was for the combination of B^0 and $\overline{B}^0$ flavor tags. The asymmetry completely vanishes if no distinction is made between the *CP* eigenstates. With the experience accumulated over the course of repeating this analysis, it was clear that the combined Δt distribution was adequate for checking the maximum likelihood fit prior to unblinding.

3.2.4 Divided Analyses

As discussed in Section 2, the analysis of data by independent groups is a very powerful tool for uncovering errors, for encouraging creativity in the analysis process, and for instilling a healthy sense of competition to produce answers in a timely way. In this Review, we do not include this approach as a method of blind analysis, if each group is able to calculate a physically meaningful answer based on their own work.

The one exception is the case where the independent groups cannot by themselves calculate a physics answer—only the combination of the two (or more) pieces of the analysis can do so, and this combination is never made until the individual analyses have been completed.

A very nice example of this approach was the measurement of the anomalous magnetic moment of the muon, done by the BNL $g-2$ Collaboration [52]. The determination of the anomalous magnetic moment a_μ relies on two completely independent measurements: the angular frequency ω_a of the difference in the muon's spin precession frequency and the cyclotron frequency, and the free proton NMR frequency ω_p which yields a precise measurement of the magnetic field B. Two independent groups were charged with the analysis—one which

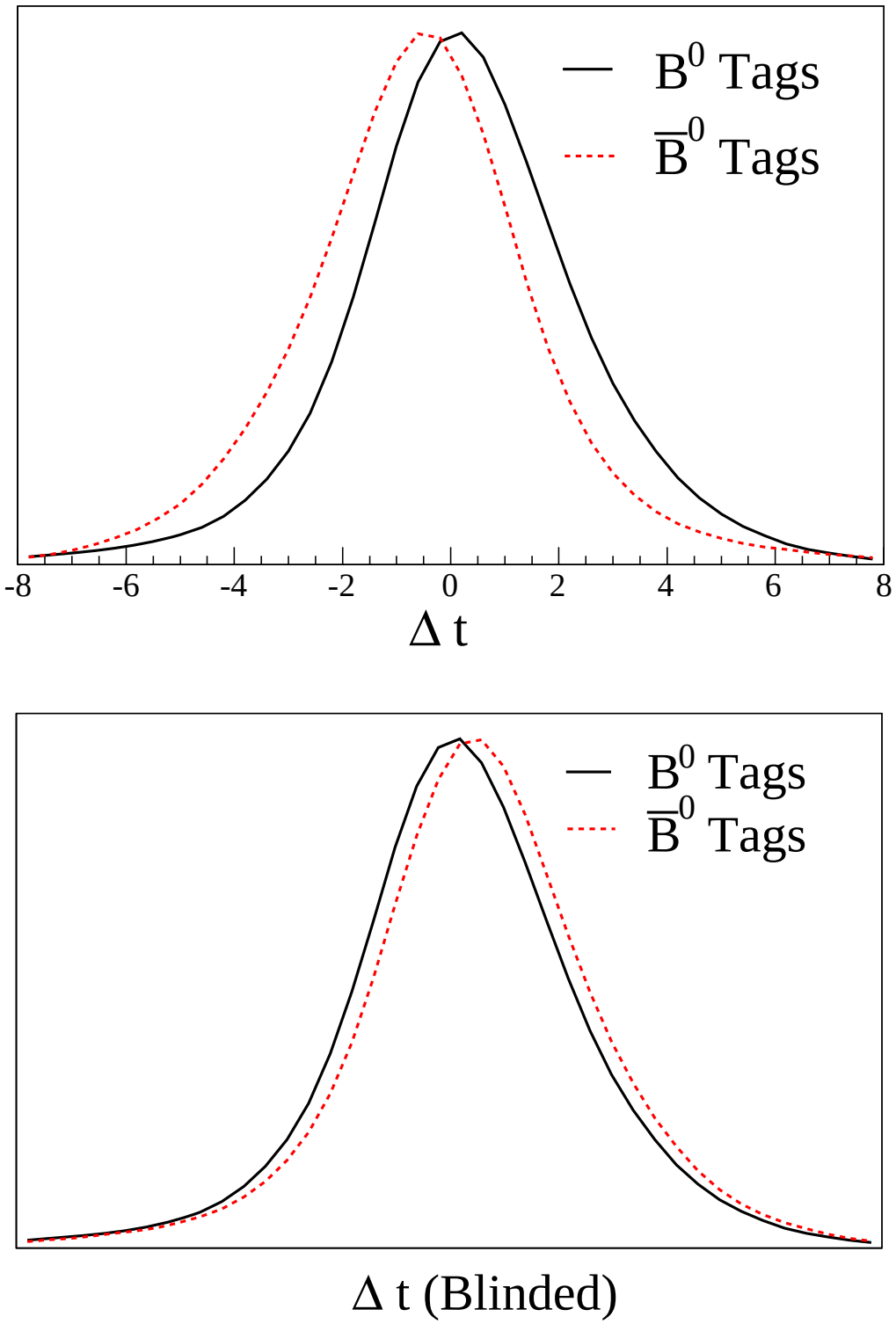

Figure 5: The Δt distributions for B decays into CP eigenstates, for $\sin 2\beta = 0.75$ and with the B^0 flavor tagging and vertex resolution which are typical for the *BABAR* experiment. a) The number of B^0 (solid line) and $\overline{B}^0$ (dashed line) decays into CP eigenstates as a function of Δt. b) The hidden Δt distributions for B^0 (solid) and $\overline{B}^0$ (dashed).

measured only ω_a and one which measured ω_p. To discuss results between the two groups, each group had its own hidden offset which it applied to its measurement. Only when both analyses were complete were the two results combined to provide the final measurement.

3.3 Adding or Removing Events

Measurements of cross sections, branching ratios, or fluxes are typically based on a count of the number of events passing all analysis cuts. As discussed in Section 3.2, a hidden answer method can be difficult to use in such cases, because there is no simple offset which can hide the number of events. While a hidden signal box approach like that described in Section 3.1 can be used for many of these measurements, it prevents the experimenter from being able to examine the characteristics of the signal, and hence carries the kind of risk discussed in Section 2: a large and obvious signal can be missed while the experimenters examine background details which later turn out to have little impact upon the measurement. In addition, a hidden signal box approach assumes that the characteristics of the backgrounds are known well enough that nothing unexpected will be discovered when the signal box is opened.

A very general approach to blind analysis which is appropriate for counting experiments is to spoil the event count itself in an unknown way. The spoiling can be done by adding an unknown set of false signal events, by removing a small unknown number of all events from the data set, or by doing both.

3.3.1 Adding Unknown Numbers of Events

If an unknown number of false signal events can be added to a data sample, an analysis can examine an entire data set while still remaining blind to the physical measurement being made. In such an approach, the experimenters tag the false events in some way and the tag is then used to remove them only when the analysis has been completed. To ensure that the number of added events remains unknown, it is critical that the false events mimic signal events as closely as possible. In some cases, a Monte Carlo simulation which can produce realistic-looking data—with all the associated noise and instrumental effects—can be used to provide a false signal event sample. Very often, however, simulations are not so realistic, and a better approach may be to add true detector events from a sample which looks nearly identical to the signal.

One example of such an approach was the Sudbury Neutrino Observatory (SNO) Collaboration's second direct measurement of the total active solar ^{8}B neutrino flux [41]. The measurement was made in SNO's second phase of operation, in which the sensitivity to neutrons produced via neutral current (NC) neutrino interactions with heavy water ($\nu + d \rightarrow n + p + \nu$) was enhanced by the addition of $\sim$ 2 tons of NaCl. The goal of the blind analysis in the second phase of the experiment was to ensure that the measurement would be independent of the previously published first phase measurement.

The primary detected signal from the NC reaction is the capture of a neutron on the dissolved Cl in the heavy water. To hide the answer, SNO added an unknown number of tagged neutrons which were not the result of the NC process by solar neutrinos. Cosmic ray interactions within the SNO detector provided just such a tagged neutron sample, and except for the preceding muon, they looked nearly identical to the signal neutrons. Figure 6

illustrates the similarity of the reconstructed energy distribution of these 'muon follower' events compared to data from a deployment of a ^{252}Cf neutron source and simulated neutral current solar neutrino events [53].

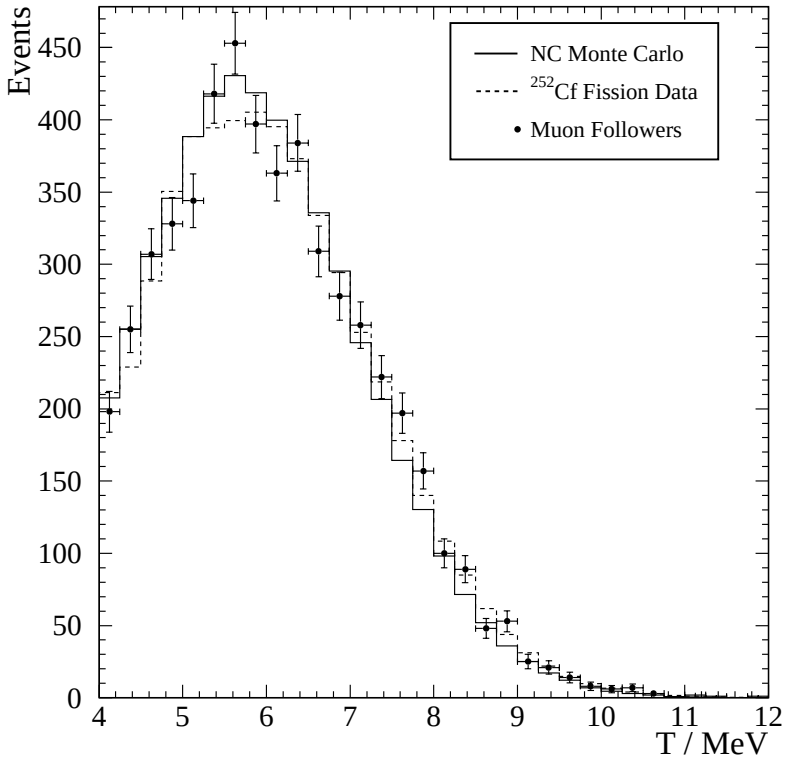

Figure 6: Comparison of muon-follower neutron events in the Sudbury Neutrino Observatory to neutron calibration source data and Monte Carlo simulated ^{8}B solar neutrino neutral current events.

These 'follower' events are normally removed from the final data sample by applying an offline veto to all events falling no more than 20 seconds after a muon event. With NaCl in the detector, the time between the liberation of a neutron from a deuteron in the heavy water and its subsequent capture on Cl is roughly 5 ms. To accept a small number of these neutrons into the final sample, a small window was cut out from the 20 second veto. The window was selected to be near enough to the 5 ms neutron capture time that it would add a number of neutrons that was on the order of the number expected from the neutrino neutral current interactions. Both the location of the window and its width were kept hidden from the Collaboration, and so the number of added neutrons was not known until the analysis was complete and the events removed from the data set. The first phase measurement by SNO of the total active ^{8}B neutrino flux was $5.09^{+0.44}_{-0.43}(\text{stat.})^{+0.46}_{-0.43}(\text{syst.})\ \times 10^6\text{cm}^{-2}\text{s}^{-2}$, and the result of the blind analysis in the second phase yielded $4.90 \pm 0.24(\text{stat.})^{+0.29}_{-0.27}(\text{syst.})\ \times 10^6\text{cm}^{-2}\text{s}^{-2}$, in excellent agreement.

3.3.2 Removing Unknown Numbers of Events

When there is no tagged sample of signal-like events to add to a data set, the complementary approach—*removing* an unknown number of events—can be almost as effective. Unlike a 'data prescaling' method in which the majority of the data set is kept blind (see Section 3.4), the fraction of removed events in this method need not be large, as long as its intentional uncertainty is large enough to prevent the experimenters from deriving a meaningful answer.

The method can be particularly powerful when used in conjunction with the approach described in the previous section. Adding an unknown number of events still allows an experimenter to have some information about the number of true signal events in the data set—they can never be larger than the number measured before the additional events are removed. There is thus a small possibility for bias—if the current count of events is 'too small' given that events have been added to the data set by hand, then searches for problems in acceptance measurements may be given more weight than (say) searches for additional sources of background. Removing an unknown fraction of events, however, gets around the problem—one does not know which effect is bigger and so the value of the measurement before the blind criteria are removed holds no information at all (and thus there is no possibility of bias). An example of this approach was the neutrino flux measurement by SNO described in the previous section: in addition to adding an unknown number of neutrons, the Collaboration removed 20-40% (the exact fraction of course being unknown) of all the events from the data set. The events were then added back to the data sets, and the fits to extract the fluxes re-done, shortly before publication.

The method can also be used effectively on its own, and is simple and general enough to work in almost any experiment. The one potential problem is handling time-correlated data, which is often present in non-accelerator experiments. To deal with this, the data needs to be divided into blocks, and rather than removing individual events from the data set, a fraction of the blocks are removed.

3.4 Data Prescaling

Perhaps the most direct blind analysis is one in which the entire analysis chain—cuts, calibrations, acceptance calculations, normalizations—are developed without any reference to the physics data set at all. Such a scenario could arise if the complete analysis could be based upon a Monte Carlo simulation, for example, and then applied without change to the data. A second case might be an experiment which has already analyzed a first run of data, and feels confident in applying the identical analysis to a new run.

Neither of these two cases is often practical—rarely is the Monte Carlo simulation trusted enough to use in the creation of an entire analysis chain, nor is a new run of data likely to be so identical to the first that one is willing to blindly (and blithely) apply an older analysis. The same benefits as this completely blind approach can be achieved, however, if the analysis is developed on a prescaled fraction of the data set and then applied to the remainder. Unlike the small removal of events described in the previous section, the prescaling fraction in this

case is known, since what is blind here is the majority of the data set.

By itself, data prescaling only avoids statistical bias—the tuning of cuts to enhance statistical fluctuations in the data. For an experiment of limited lifetime or low statistics, this can be a potentially damaging source of bias as there is no way to determine after the fact whether there has been any unintentional tuning—there is only one instance of the data set to study.

Data prescaling relies on the fact that the set of cuts being applied is asymptotically unbiased—if applied to an infinite data set, they have no preference for accepting non-signal events which happen to lie in the signal region. Imagine, for example, that it is found that by cutting harder on a reconstructed track χ^2, the significance of an invariant mass peak grows slightly. The growth may simply be due to the fact that by moving the cut, the number of events in the peak fluctuates high, as depicted in Fig. 3. If the cut is then fixed and applied to a much larger data set, the cut will have almost no preference for events in the peak region if there are no real signal events.

Several assumptions are inherent in any data prescaling scheme:

1. The prescaling is done in an unbiased way
2. Any data sample is the same as any other, *or* time-dependent variations have characteristic scales which can be properly sampled or contained within a subset of the data which spans those characteristic times
3. The statistics of the prescaled sample are large enough to identify backgrounds, but small enough that they will not bias the result for the entire data set

The first assumption almost always holds, unless one picks a prescaling scheme which is based upon some criteria within the data set itself. A poor choice might be to look only at data when the trigger rates are high, for example. The safest scheme randomly decides whether an event or set of events falls within the prescale sample or not. For some experiments—particularly non-accelerator experiments—time correlations between events can be important, and thus blocks of data need to be prescaled rather than individual events.

The second assumption is not necessarily always true. It is rare that any sample of data is exactly the same as any other—data can be taken during times when detector sub-systems are offline, or at different beam intensities, or even different times of day. Any external variation which can change the background levels or the detector sensitivity or acceptance can make a prescaling scheme fail if the variations are not reasonably sampled. Here 'reasonably sampled' means that the sampling frequency—how often an event or block of data is selected to be in the prescaled sample—is higher than the rate of known variations in detector or beam conditions.

The third assumption, strictly speaking, is almost always false. Determining that the background in a pre-scaled sample is zero, for example, provides a very weak limit on the background levels inside the remainder of the data set, unless the prescaled sample is a large fraction of the full data set. In practice, one uses the smallest data sample possible that

makes the first two assumptions true, and then determines whether the upper limit on residual backgrounds in the full data set is acceptable. If one feels that the prescale fraction needs to be large in order to measure backgrounds—in excess of 30%—then in principle the prescaled fraction should be discarded in the calculations of the final results, as otherwise any statistical bias in the analysis of the prescaled sample will have a non-trivial effect on the final answer.

Many experiments have used data prescaling either alone or in combination with other techniques. Brookhaven E888 [54] used a hidden signal box technique in combination with a 10% prescaled sample of data within the signal box. The NOMAD experiment, which searched for $\nu_\mu \to \nu_\tau$ oscillations, had an 'effective' prescaling scheme in which they analyzed openly a first run of data which constituted 20% of the data set, and then analyzed a much higher statistics run using a hidden signal box technique [56]. The E787 experiment, which searched for the rare decay $K^+ \to \pi^+\nu\bar{\nu}$, used a hidden signal box as their primary blind approach, but then divided the events outside the signal box into a 1/3 and 2/3 sample. E787 then developed their analysis of backgrounds on the 1/3 sample and applied that analysis to the remaining 2/3 to provide the final background measurement [47]. The Sudbury Neutrino Observatory, in its first publication [57], used a prescale technique as a test of gross statistical bias—30% of the data was kept in reserve and the analysis developed on the other 70% applied to the hidden sample. The LIGO gravity wave experiment used a 10% sample[58] for analysis optimizations in their search for gravity wave bursts. This subsample was then discarded in the final results. The MiniBooNE [55] neutrino experiment currently uses a data prescaling scheme of just 0.5% for all data, in addition to a hidden signal box approach.

One question that can arise in treating the hidden sample as a blind data set is, how blind is blind? As the events are recorded, on-line event displays and other monitors are often viewed by collaborators. Does some blindness scheme need to be imposed upon these? As discussed in Section 2, a blindness scheme is not intended to keep experimenters from looking at the data, but to keep the results of the analysis from influencing the analysis itself. For most experiments, event displays and other monitors carry no information about the physics results, and therefore do not influence the analysis itself. The MiniBooNE experiment is one example of this—low-level event and electronic channel properties can be examined throughout the data set. The possible exceptions to this guideline are open-ended rare process searches, where one is looking for unique and unusual events. Noticing a few strange events by hand-scanning is likely to influence a later search for new physics—it is hard to design a new analysis which doesn't ensure that these interesting events make it into the final sample.

Finally, in general it may be difficult to use a blind analysis method in searches for new particles. However, there are also a number of cases in which new particles, observed as bumps in invariant mass distributions, were ultimately found to be experimental artifacts[59]. Often such statistical fluctuations do not have physical characteristics—the width of the bump, for example, is found to be inconsistent with the expected detector resolution. In addition, the evaluation of the statistical significance of the bump depends on the measure of the search space, or the number of places a bump could have appeared (the 'trials' problem).

Given the vagaries of both of these issues, searches for new particles, or bump hunting, would clearly benefit from a blind approach. Unfortunately, the hidden signal box technique cannot readily be applied, since the mass of the particles is obviously not known. Of the methods described, only the data division method may be readily used. Despite the limitations described above, this method for determining experimental selections and analysis techniques may help prevent some purely statistical artifacts. However, creative new approaches may be possible, and we urge experiments making new particle searches to consider new blind methods.

4 Conclusion

In his speech *Cargo Cult Science*[60], Richard Feynman warns that

> It's a thing that scientists are ashamed of - this history - because it's apparent that people did things like this: When they got a number that was too high above Millikan's, they thought something must be wrong - and they would look for and find a reason why something might be wrong. When they got a number closer to Millikan's value they didn't look so hard...
>
> The first principle is that you must not fool yourself - and you are the easiest person to fool.

Experimenter's bias represents one way to fool yourself, and blind analysis provides the solution. By describing the application of different blind analysis methods to a range of measurements, including both the motivations and potential problems, we hope that this review will be useful as a resource for experiments and will promote the current trend towards blind analysis in nuclear and particle physics.

Acknowledgments

The work of A.R. is supported by the Department of Energy under contract DE-AC02-76-SF00515. The work of J.K. is supported by the Department of Energy under grants DE-FG02-04ER41332 and DE-FG02-93ER40757, and by the Alfred P. Sloan Foundation. The authors would like to acknowledge useful discussions, both in the development of blind analysis methods and in the preparation of this review, with Art Snyder, Pat Burchat, William Molzon, Bruce Winstein, Peter Shawhan, Jack Ritchie, Robert Cousins, Huaizhang Deng, Richard Van de Water, Hirohisa Tanaka, and Stanton Goldman.

References

[1] De Angelis, et al., The New England Journal of Medicine, 351, pp. 1250-1251, (2004).

[2] Dunnington, Frank G., *Phys. Rev.* 43, pp 404-416, (1932)

[3] Roodman, A., eConf C030908:TUIT001, (2003)

[4] Birge, R.T., *Rep. Prog. Phys.* **8**, 90 (1941); Cohen, E.R. and DuMond, J.W.M., *Rev. Mod. Phys.* **37**, 537 (1965).

[5] Cohen, E.R. and Taylor, B.N., *J. Phys. Chem. Ref. Data* **2**, 663 (1973).

[6] Birge, R.T., *Nuovo Cim. Suppl.* **6** 39 (1957).

[7] Christensen, C.J., et al., *Phys. Lett.* 26B, pp. 11-13, (1967)

[8] Christensen, C.J., et al., *Phys. Rev.* D5, pp. 1628-1640, (1972)

[9] Barkas, W.H. and Rosenfeld, A.H., *UCRL-8030*, (1958).

[10] M. Roos et al., *Rev. Mod. Phys.* 35, (1963).

[11] M. Roos et al., *Nucl. Phys.* 52, pp. 1-24, (1964).

[12] A.H. Rosenfeld et al., *Rev. Mod. Phys.* 36, pp. 977-1004, (1964).

[13] A.H. Rosenfeld et al., *Rev. Mod. Phys.* 37, pp. 633-651, (1965).

[14] A.H. Rosenfeld et al., *UCRL-8030-Rev* (1966).

[15] A.H. Rosenfeld et al., *13th International Conference on High-energy Physics* (1966).

[16] A.H. Rosenfeld et al., *Rev. Mod. Phys.* 39, pp. 1-51, (1967).

[17] A.H. Rosenfeld et al., *Rev. Mod. Phys.* 40, pp. 77-128, (1968).

[18] N. Barash-Schmidt et al., *UCRL-8030-Pt-1-Rev* (1968).

[19] N. Barash-Schmidt et al., *Rev. Mod. Phys.* 41 pp. 109-192 (1969).

[20] A. Barbaro-Galtieri et al., *Rev. Mod. Phys.* 42 pp. 87-200 (1970).

[21] M. Roos et al., et al., *Phys. Lett.* B33 pp. 1-127 (1970).

[22] A. Rittenberg et al., *Rev. Mod. Phys.* 43 pp. S1-S150 (1971).

[23] P. Söding et al., *Phys. Lett.* B39 pp. 1-145 (1972).

[24] T.A. Lasinski et al., *Rev. Mod. Phys.* 45 pp. S1-S175 (1973).

[25] V. Chaloupka et al., *Phys. Lett.* B50 pp. 1-198 (1974).

[26] T.G. Trippe et al., *Rev. Mod. Phys.* 48 pp. S1-S245 (1976).

[27] C. Bricman et al., *Phys. Lett.* B75 pp. 1-250 (1978).

[28] C. Bricman et al., *Rev. Mod. Phys.* 52 pp. S1-S286 (1980).

[29] M. Roos et al., *Phys. Lett.* B111 pp. 1-294 (1982).

[30] C.G. Wohl et al., *Rev. Mod. Phys.* 56 pp. S1-S304 (1984).

[31] M. Aguilar-Benítez et al., *Phys. Lett.* B170 pp. 1-350 (1986).

[32] G.P. Yost et al., *Phys. Lett.* B204 pp. 1-486 (1988).

[33] J.J. Hernández et al., *Phys. Lett.* B239 pp. 1-516 (1990).

[34] K. Hikasa et al., *Phys. Rev.* D45 pp. S1-S574 (1992).

[35] L. Montanet et al., *Phys. Rev.* D50 pp. 1173-1826 (1994).

[36] R.M. Barnett et al., *Phys. Rev.* D54 pp. 1-708 (1996).

[37] C. Caso et al., *Eur. Phys. J.* C3 pp. 1-794 (1998).

[38] D.E. Groom et al., *Eur. Phys. J.* C15 pp. 1-878 (2000).

[39] K. Hagiwara et al., *Phys. Rev.* D66 pp. 010001 (2002).

[40] S. Eidelman et al., *Phys. Lett.* B592 pp. 1-1109 (2004).

[41] Ahmed. S.N., et al, *Phys. Rev. Lett.* 92, 181301-1, (2004)

[42] Ritchie, Jack R., private communication.

[43] Galison, Peter, *How Experiments End*, The University of Chicago Press (1987).

[44] Richter, Burt, private communication.

[45] Blind Analysis Task Force [Babar Collaboration], Babar Analysis Document # 91, (2000).

[46] K. Arisaka *et al.*, Phys. Rev. Lett. **70**, 1049 (1993).

[47] Adler, S., et al., *Phys.Rev.* D70, p37102 (2004)

[48] Hoskins, J.K., Newman, R.D., Spero, R., and Schultz, J., *Physical Review* D12, pp. 3084-3096, (1985)

[49] A. Alavi-Harati *et al.* [KTeV Collaboration], Phys. Rev. Lett. **83**, 22 (1999).

[50] Shawhan P. S., *Observation of direct CP violation in K(S,L) → pi pi decays,* Ph.D. Thesis U of Chicago, (1999).

[51] B. Aubert *et al.* [BABAR Collaboration], Phys. Rev. Lett. **86**, 2515 (2001).

[52] Bennett, G.W., et al., *Phys. Rev. Lett.* **92**, 161802 (2004).

[53] We thank the SNO Collaboration and Neil Mccauley for this figure.

[54] Belz, J., et al., *Phys. Rev.* D53, 3487-3491, (1996)

[55] R.G. Van de Water and H.A. Tanaka, private communication, (2005).

[56] Astier, P. et al., *Phys. Lett.* B 453, 1690186, (1999)

[57] Ahmad, Q.R. et al, *Phys. Rev. Lett.* 87, 071301, (2001).

[58] Abbott, B. *et. al.* [LIGO Collaboration], *Phys. Rev. D* **69** 102001 (2004).

[59] Stone, S. *Pathological Science* Flavor Physics for the Millennium: TASI 2000 Proceedings. Edited by J. L. Rosner, Singapore, World Scientific, (2001).

[60] Feynman, R. *Surely You're Joking, Mr. Feynman!* W. W. Norton & Company (1985).