

The Need for Experimental Controls in Research: Guidelines for Beginning Research Students

I. Overview

Science is a profession, an approach, and an attitude. In the actual day-to-day doing of science, *truth & progress* are far more likely to result by adopting a **skeptical** attitude. They most certainly will not result from being passive or unquestioning. Therefore, students pursuing careers in science need to be taught the healthy, balanced skepticism and good questioning skills associated with sound scientific experiments.

In this spirit, we offer this article as an introduction to the importance of doubt and skepticism in science and discuss how the use of *controlled* experiments that include standards, control samples, blanks, and dummy analyses can help remove bias and increase confidence in our measurements and observations. It is intended primarily for beginning research students, but also for students in laboratory courses and those who are making measurements aimed at answering chemical questions. We recognize that we cannot be comprehensive in our coverage or examples; rather, we intend that this article encourage discussions between faculty and students.

Doubt and How to Systematically Minimize It

In order to be successful at science, you *must* anticipate potential sources of errors in your measurements and perform analyses that safeguard against them. You must recognize that what you are doing might well be erroneous and make efforts to prove otherwise. This requires subjecting every measurement, hypothesis, and working explanation to a critical analysis to examine what might be wrong with them. It also requires considering alternative hypotheses that are equally capable of explaining your observations. If alternative explanations or potential flaws are discovered during this analysis, then further experiments must be conducted to address those concerns by isolating and correcting problems and by trying to disprove the alternative hypotheses. Herein lies the foundation of modern science as described by Karl Popper (1-3). We create theories to explain how nature works. We accept those theories if they pass all tests to refute them and all other *reasonable* explanations have been disproved (*ref. 3, chapter 9*). This syllogism applies equally to individual measurements, complex experiments, and the broad, overarching conclusions resulting from the accumulation of experimental results. While it is widely recognized that it is the predictive power of successful theories that make them useful, predictions emanating from theories also provides a means of testing them. It is by this iterative process of experiment, analysis, testing of theories, and refinement of experiments that science progresses towards a more complete understanding of the physical world.

We recognize that Popper's philosophy of science, while quite influential, is not universally accepted. References 1 through 8 address other philosophical approaches to science and knowledge, as well as the sheer enjoyment that can be derived through scientific pursuits.

To improve the probability that a measurement you have made is valid, you must **design** experiments to include the proper checks, consisting of dummy analyses, blanks, controls, and standards. If you do not use these techniques, there is no point in doing the work, since they are included to reveal the biases, drifts, and other problems that can arise during your measurements.

We want to be clear that the doubt we are discussing should not be associated with an equally important concept, namely, the statistical uncertainty inherent in measurements. We focus mainly on how to check for **systematic bias** in measurements. Statistical uncertainty, (e.g.

the standard deviation of a set of measurements) deals with the reproducibility of a measurement. Even if you repeat an analysis and get nearly the same result each time (i.e. have good reproducibility), it may still be quite wrong if there is a bias or determinate error in your measurement methods. Thus, reproducibility and bias are distinct concepts, but they are equally important in making measurements. A discussion of the issues surrounding reproducibility lies outside the scope of the present article. For reviews of these topics, the reader is directed to references 9 through 16.

II. How are individual measurements related to an experiment?

Neither a single measurement nor a collection of measurements is an experiment, no matter how fancy the apparatus used to gather the data. Thus, collecting a spectrum on the world's newest, most expensive 900 MHz NMR spectrometer is not necessarily an experiment. In its most elementary form, an experiment is set up to test the response of a system to a systematic change in a single variable. A series of such experiments in which the variable is changed over a wide range and other potentially influential variables are also tested leads to a set of mutually supportive measurements or observations from which a theory may be constructed. To test this new theory, every attempt to rule out all reasonable alternative hypotheses as to the explanation of the observation(s) must be made, thereby leaving in the best case only a single, reasonable explanation and thus "prove" the preferred hypothesis. In order to design a good experiment, one must think ahead several steps in an attempt to understand what can go wrong with each measurement or each hypothesis and thus lead to an incorrect conclusion. *Therefore, it is very important to design into the experiment significant controls that validate the measurements.* Confusion lies in not considering the possibility of other explanations for the observation, blindly forging ahead, and doing additional measurements based on your conclusions without doing the *essential* checks.

III. Blanks, Dummy Analyses, Standards, Controls, and System Suitability Testing

Dummy analyses, controls, blanks, and standards form the heart and soul of making reliable measurements by providing the essential checks that maximize confidence. Often, however, the differences between standards, controls, and blanks are misunderstood. Thus, we attempt to define each of these and distinguish their use in validating a chemical measurement.

In this context, we narrow our focus to the simple, straightforward, specific examples of elucidating the structure or measuring the concentration of an analyte, rather than the testing of an elaborate theory. These measurements are important since complete research projects often consist of numerous determinations of structures or concentrations of an analyte as a function of some perturbation (e.g. the concentration of a drug metabolite in the body as a function of time or the effect of a series of catalysts on an synthetic method). When you set out to make a measurement, especially if you have never run that type of sample or are using an instrument that is new to you (i.e. whenever you are not on "tried and true" territory), no result should be believed until a dummy analysis, a blank, a set of standards (calibrants), and a control sample are run. Each of these, and the subtle differences between them, is discussed below.

Blanks: A blank is a sample that includes everything that will be present in the sample *except* the analyte of interest (i.e. a blank is the sample matrix). Ideally, the blank should be identical to the sample except that it contains none of the actual sought-for analyte. If the blank signal is simply compared to the signal obtained in the presence of analyte, it is a control showing that the

sample is indeed producing a signal. Additionally, running the proper blanks will prevent wasting time trying to interpret signals that are unrelated to the analyte. The blank will also reveal signals that can potentially overlap or interfere with analyte signals.

The signal from the blank must be entirely negligible, or small enough compared to the signal from the real sample, that it can be subtracted from the real sample without violating statistical common sense (don't ever rely on a small difference between two big measurements without very carefully checking the reproducibility). For example, many solvents such as methanol can be used for spectroscopic work in the visible and UV range. However, below about 210 nm methanol absorbs very strongly. A spectrophotometric determination in this spectral region using solvent mixtures is subject to errors caused by small variations in solvent composition. It is wise to avoid using wavelengths where the matrix also absorbs, if possible. One last note about blank signals: if they are subtracted from the sample signal, the result should be referred to as a "corrected value" (17).

If the blank reveals signals which cannot be rationally accounted for, then a "dummy analysis" should be run.

Dummy Analyses: In a dummy analysis, nothing (or as near to nothing as possible) is added to the instrument (no sample, no matrix, etc.). It is meant to test the instrument's ability to progress through the scheduled sequence of events and to check if there are any signals that are attributable to those events. For example, in gas chromatography, a dummy analysis consists of making an injection with only air in the injection needle, and then continuing on with the temperature program as if something had really been injected. When using spectrophotometers, one could record the spectrum of air in the absence of the sample, salt plates, cuvettes, etc. In each case, one would hope to see no signal arising from this dummy analysis. Sometimes, however, one can observe signals from the dummy analysis. In gas chromatography, these could arise from compounds from a previous analysis that had adsorbed to the inside of the injection needle – a common phenomenon called "carry-over". Without doing the dummy analysis, these compounds would have been injected with one of your samples, making it appear as if they were actually present in your sample. This would cause unexpected and unexplainable peaks to appear, or worse, cause interference with your peaks of interest. Thus, the purpose of the dummy analysis is to gain confidence that all of the signals you see in subsequent analyses have come from your samples and are not artifacts of the instrumentation.

Dummy analyses are different from blanks (see above) in that as near to nothing as possible is added to the instrument. To distinguish a dummy from a blank, consider the scenario in which a blank in a gas chromatographic analysis has just been run by injecting 'pure' solvent. If unexplainable peaks are observed, the question becomes "are the peaks due to impurities in the solvent or some other unforeseen source (carry-over, an instrumental "glitch", etc.)?" At this point, you would run a dummy analysis and inject air instead of solvent. If the same peaks were present in the dummy analysis as in the blank, then you would conclude that they did not arise from the solvent, but rather from something else (assuming that you had rigorously eliminated all of the solvent from the syringe before injecting the air). If the peaks are not present in the dummy analysis, then they are likely coming from the solvent.

One final note about dummy analyses: They are subtly different than backgrounds. A background spectrum or chromatogram is also collected without adding any sample or matrix components to the instrument. It is typically the first set of signals recorded and subtracted from all subsequent measurements. In this case, a dummy analysis would be collected immediately

following the collection of the background, the subtraction performed, and the resulting signal analyzed. Hopefully, the subtraction was perfect and the signal is zero, within the noise limit, across the spectrum or chromatogram. This is sometimes not the case, however. (seems awkward to end a sentence with a conjunction). For example, in FTIR analyses, fluctuating levels of carbon dioxide can lead to signals that do not perfectly cancel between the dummy and background spectra. In scanning spectrophotometers, artifacts that do not perfectly cancel can arise at the wavelength at which a mirror in the instrument is flipped to switch from a UV to a visible lamp. In these cases, the dummy helps identify the origin of signals that may continually fluctuate throughout the course of the sample measurements.

Standards/Calibrants: A standard is used to calibrate the test equipment. Sometimes, standards are physical devices such as masses, resistors, voltage sources, etc. whose values are known exactly and are traceable to a primary reference maintained by a standards agency such as the National Institute of Standards and Technology (NIST) in the United States. Quite often, standards are solutions prepared from a pure or certified source of the species of interest in a very simple matrix (e.g. pure water or pure solvent). It might be a primary standard substance such as potassium hydrogen phthalate (KHP) used to standardize a solution of base for acid/base titrations. When determining concentrations, standards are used to prepare a calibration curve that defines the mathematical relationship between the signal and the concentration of analyte. In the context of structure elucidation, standards are used to calibrate wavelengths of spectrophotometers and mass-to-charge ratios of mass spectrometers.

More often than not, the sample matrix influences the sensitivity of an analytical method and can therefore introduce bias into measurements. For example, the presence of phosphate decreases the efficiency of atomization of calcium in a flame atomic absorption determination. In many cases the mechanism for the interference is difficult to ascertain or predict except that it is associated with the sample matrix. We can test to see if a **matrix effect** is present by applying the **method of standard additions**. When using the methods of standard additions, several portions of the same real sample are prepared for analysis. One is used without further treatment while the others are each spiked with known increasing amounts of authentic pure analyte. The signals for these samples are plotted against the increase in analyte concentration. The slope of this line is compared to the slope of the calibration curve prepared from pure standards in a simple matrix (i.e. water, buffer, or pure solvent). A matrix effect is likely present if the slopes of these two lines are significantly different. Consequently, determinations of analyte concentrations calculated by comparing the sample to the standard calibration curve in the simple matrix will be inaccurate. The analyst now has a choice: either prepare standards in a matrix that effectively mimics that of the sample or use the method of standard additions to calculate the analyte concentration. The latter approach is usually easier and is discussed extensively in articles in this Journal (18) and in texts on quantitative analysis (17, 19-21).

From the standpoint of collecting reliable data, it is important to appreciate the difference between *standards*, which establish the calibration or the concentration/response relationship of an instrument, and *controls*, which serve as independent checks to make sure your analysis method is generating accurate answers (see below). They serve different functions and thus achieve different goals in the practice of science.

Controls: A control is a sample whose structure or concentration is known and which is as similar to the real unknown as possible (i.e. similar matrix composition). It is subjected to the

analysis method periodically throughout the course of the experiment. It must be run in exactly the same way as the sample in a time frame as nearly identical as possible, both **before** and **after** the real unknown. It is important to run the control before and after the actual sample since if you get the expected value both times, it is unlikely that the instrument was operating in a different manner when you ran your sample. If you only run the control prior to running your sample, then you cannot be sure that the instrument was behaving the same way during your sample run as it was during the control run. Surely many more people will accept your answer on an unknown if you get the right answer on a control. Getting the "right" answer on a standard used to standardize or calibrate the system is clearly not supportive since the standard usually does not contain anything but the analyte species and solvent and was used to 'train' the system. Furthermore, standard solutions are not good controls since they will not detect biases that are introduced by the sample matrix. However, repeatedly running a standard or calibrant throughout a long run of analyses is not totally without merit since if you get the wrong answer on a standard/calibrant you surely will get the wrong answer on a real sample.

A control is much more than a simple duplicate or triplicate analysis, that is, a sample done in replicate in quick succession. One purpose of running the control *throughout* the time-course of the experiment is to detect slow changes (drift) in the performance of your analytical system. Drift can take many forms and arise from many sources. For example, it may be related to a gradual change in the temperature of the room throughout the day as lights and body heat warm the environment, a change in the dispersive properties of a spectrophotometer grating as the light from the source warms it, a change in the pump performance in chromatography as the seals wear out, or any slow change such as a reagent decomposition that would not be caught over a short time period by back-to-back, successive replicates. Minimally, in a sequence of measurements that takes more than a couple of hours, one should also choose one actual sample to be repetitively analyzed at the beginning, in the middle, and at the end of the study. If only a very few samples are run, then one sample should be repeated by running it first and last. References 17 and 18 provide further discussion focused on the use of controls.

System Suitability Check: A system suitability check is much less rigorous than a real control. A suitability check of the apparatus is run by analyzing some "test mix" which might not resemble the unknown to make sure that the system is behaving as it did in the past. For example, a simple system suitability check for an FT-IR instrument would be to collect the spectrum of the same piece of polystyrene film each time the instrument is used. If the spectrum collected on any given day is different than those collected previously, whether in terms of the wavelengths at which the absorbances are observed or their intensities, then something has changed, either with the film or the instrument. Whichever it is, the change needs to be tracked down and explained before we can have confidence in the results obtained with the instrument that day. Keep in mind, however, that this is not a true control. It merely indicates that the apparatus is in good repair. Certainly, if the system fails a "suitability check" it must not be used to measure an unknown sample. This is a very good thing to do before collecting any data but does not circumvent the need to run a true control sample. In doing studies over the course of several days, suitability checks guard against biases caused by instrument drift such as mobile phase composition changes, wavelength calibration changes, and loss of temperature control, and are therefore valuable to incorporate into the experimental protocol.

Comparison to Established Methods: Another form of a control experiment which would help convince a skeptic to accept a new analytical method (or indeed any new approach) would be to use an older, very well established ("proven") analytical method on the actual real unknown and show that one gets the same answer by both methods. Use of a carefully calibrated thermometer to check an instrument readout temperature (not to 'restandardize' it), and measurement of flow rate volumetrically or gravimetrically are forms of control measurements with established ("incontrovertible") methods.

In some cases it is possible to obtain a **reference sample**. NIST has carefully prepared and analyzed many common analytes in specific matrices (e.g. trace metals in orchard leaves). If you are lucky enough to be determining an analyte in a sample with a matrix **similar** to one that is available from NIST, obtaining the same analyte concentration by your method as the value certified by NIST would greatly increase everyone's confidence in your method. Of course, if your method does not give the value obtained by NIST you may have a very significant problem on your hands. Certainly it is possible that there is nothing in your sample's matrix that interferes but how can you be sure?

Summary of dummy analyses, controls, blanks, and standards

In summary, dummy analyses check for signals arising from effects not related to your sample matrix or analyte. Blanks are used to find signals, if any, that arise from the matrix itself rather than from the analyte. Spiked samples are useful for checking matrix effects on signal sensitivity. Thus, dummy analyses and blanks highlight potential interferences that could overlap or mask the signal of the analyte. Standards and calibrants are used to train the system and measure the perturbation/response behavior of the instrument. Controls are critically important in that they can detect instances in which an analytical method does not generate accurate results, or they can provide confidence that the analytical method is working well. System suitability checks make sure that the apparatus you are using is behaving consistently over time and providing accurate answers for known test mixtures. All of these techniques must be *rationaly* built into experimental protocols to generate reliable measurements. Examples of this are presented below in the form of case studies.

We strongly encourage the incorporation of these (concepts??) in undergraduate laboratory experiments, especially in analytical chemistry courses, both for their pedagogical value in developing the sense of skepticism for which we argue here, and also to increase a student's confidence in his/her ability to obtain "correct" results. A few examples of laboratory experiences which have appeared in this *Journal* and which focus on the quality control of measurements are provided in the references (22-25).

IV. How the use of controls impacts experimental design: Case studies

These case studies are meant to hint at the design of experimental protocols that incorporate the techniques discussed above. An exhaustive discussion of experimental design would be too extensive for this article. The reader is directed to references 12 and 13 and 26 through 29 for reviews of experimental design. We begin by expanding a common place (homely?) scenario presented by Silberberg (30) to illustrate that the same familiar thought processes that we employ in every-day, non-science problem solving can be applied in science.

Case 1: The malfunctioning stereo system.

Imagine you have a stereo system consisting of a receiver, a CD player, and a pair of speakers. If, while listening to the radio, the sound becomes garbled, what checks and controls would you need to do to identify the source of the problem and be confident in your diagnosis? First you might suspect that the problem resides with the particular station that you are listening to, so you would change to a different station. If the problem persists, you might suspect that you have generally poor reception for all radio stations, or perhaps you may begin to suspect that you have a problem with your speakers. After checking a few more stations (all garbled), you might try playing a CD. If the sound is okay, what conclusions would you draw? You would probably take this to mean that the speakers are working and that the problem resides somewhere within the receiver. But suppose that the sound is garbled while playing the CD. You might then try using headphones. If the sound is clear in the headphones, then you might begin to suspect that the problem is likely in the speakers. However, the problem could still reside in the receiver, but at least you have isolated it to the circuitry that controls the speakers as opposed to the circuitry involving the headphones (i.e. the problem likely resides in the sound output circuitry, rather than in the input or conversion steps). How could you isolate the problem to the speakers as opposed to the receiver? If you hooked up your speakers to a friend's receiver that you knew was working, and you got clear sound for the same radio station and the same CD, then your speakers must be working and the problem seemingly resides in your receiver. For a further test, you might hook up your questionable receiver to your friend's working speakers. If the radio and CD were both garbled, you would again conclude that there is something wrong with your receiver. This, then, is the component that you would take to get fixed. Or, if it were under warranty, you would want it replaced. Note, though, that you would not insist on a new receiver if you were not sure that the receiver was causing the garbled sound – you might end up being told that nothing is wrong with it and that you should go home and check your speakers. Doing all of the above tests, however, gives you confidence that you are justified in asking for a new receiver. Through all of the above, you have tested and re-tested (i.e. controlled for) all other reasonable explanations as to why the sound was garbled. You have formed testable hypotheses (e.g. the speakers are faulty), and systematically verified or denied each one (test with the radio station and a CD). The same logic and types of tests/controls must be employed in science to reach sound, reasonable conclusions. Clearly there is nothing extraordinarily complex in designing controls into an experiment. We do it every day.

Case 2: Is chloride present in solution?

Consider the following situation that is somewhat analogous to the example above: You have an aqueous solution that you suspect contains chloride ions at roughly 1 mM concentration. How can you test if your suspicion is correct? In its simplest form a test for chloride is to add silver ions and thereby form a white AgCl precipitate. You go into the laboratory find no solid silver nitrate but do locate an old bottle of solution labeled "silver nitrate" that you did not prepare. If you add it to the unknown solution and no precipitate forms, does this unequivocally mean that the sample does not contain chloride? Conversely, if you see a precipitate, does it absolutely mean that chloride is present? There are many questions you would need to ask, and a number of controls you would need to run to answer those questions. For instance, you would need to verify that the solution really contains silver nitrate – after all, it is old and may have changed with time, and you did not make it so you do not know if it was properly labeled in the first place. This could be done by making a sodium chloride solution and adding some of the silver nitrate solution to it. If silver ions are present, a precipitate will form (does the formation

of a precipitate *unequivocally* prove that silver is present?). If the solution is found to contain silver ions (or at least something which precipitates with chloride), how can you be sure that you would see the precipitate if your unknown solution truly does have 1 mM chloride in it? What if the sample is less than 1 mM? You would need to make a series of solutions with decreasing chloride concentrations and test each one with the silver nitrate solution to determine the lowest concentration can you detect. What if a precipitate forms – can you be sure that it is due to chloride? You would also need to consider the possibility that bromide or iodide ions, which also precipitate with silver ions, are present. How could you differentiate bromide and iodide from chloride?

Hopefully, the similarities between this example and the one involving the stereo speakers are clear and serve to illustrate that the same logic we use to test every-day hypotheses is required to test scientific hypotheses. It would be embarrassing to call a technician to fix your speakers and have him/her find that the problem was the fault of the radio station you were listening to and there is nothing wrong with your speakers (i.e. you did not do the necessary checks to isolate the problem). Likewise, it would be embarrassing in science to publish a result, or have companies make decisions based on results, without conducting the obvious and reasonable tests and controls on the measurements you are making.

Case 3: A series of pH measurements.

Consider something as seemingly simple as making a long series of pH measurements of a number of samples. Perhaps one is trying to determine if an environmental spill is changing the pH of a nearby stream, and how far downstream this effect may be observed. Samples are taken every tenth of a mile for two miles downstream, resulting in a set of twenty samples. Certainly, you calibrate the meter and electrodes at the beginning of your measurements. But how do you know that the measuring system (temperature, liquid junction, etc.) does not drift during the series of measurements? One gains much more confidence in the whole data set when one periodically runs one of the first samples (not a standard used to calibrate the meter) repeatedly throughout the set of measurements. Even more confidence would be gained if you were to also run a solution whose pH was known (say a third “standard” buffer i.e. one **not used to calibrate the meter**) periodically throughout the series of measurements. The issue of quality control of pH measurements has been addressed in this *Journal* by Stapanian and Metcalf (31).

Case 4: Baseline drift in a liquid chromatographic analysis.

Consider a situation where the baseline of a chromatogram becomes rather noisy. Suppose that you swap the pump, injector, and column with those from a fully functioning neighboring instrument, but connect them to the UV-visible detector on the malfunctioning LC. If the baseline remains bad, this likely eliminates the pump, injector, and column as sources of the problem, and suggests that the problem resides in the detector. The problem could be due to an air bubble, faulty electronics, etc., but most likely it is due to the lamp because the lamp indicator is giving a warning (this does not prove it is the lamp as it is possible that the device that measures the lamp strength has gone bad). A good control is to take a lamp from a functioning LC and put it in the bad detector. Suppose that the instrument now works. As a double check, take the suspect bad lamp and put it in the LC that was working. If that LC now gives a bad baseline you can be certain that a bad lamp caused the initial problem. It is a little more work, but in the long run it builds confidence in the conclusion.

Case 5. Pharmaceutical stability tests.

Pharmaceutical companies routinely perform what are known as ‘stability tests’ to measure the shelf life of their products so as to provide expiration dates for their medications. Stability tests consist of determining the concentration of the drug over a period of many days or weeks, often using liquid chromatography to separate and quantify the active ingredients. Imagine that over the course of several weeks, the concentration of the drug, as determined by the area under the peak, is found to be essentially constant. An unquestioning interpretation would be that the drug is stable for the time period studied. But can the pharmaceutical company truly be certain of this interpretation? Consider the fact that the structures of the degradation products might be similar to the structure of the drug itself. In this case, the degradation products would behave quite similarly to the actual drug in terms of the response they generate during the measurements, making it *seem* like the drug is not degrading when in fact it is. Because of this possibility, pharmaceutical companies must do what are called ‘forced degradation studies’ in which the drug of interest is subjected to conditions known to cause the breakdown of most organic compounds (extreme high and low pH, high temperatures, high humidity, etc.). The forced degradation sample is then analyzed by the same method as the original drug sample. If the degradation products elute at different times than the drug, one is more certain that the initial analyses that showed no degradation are valid. If the degraded sample has a peak at the same or nearly the same time as the drug, then the initial analyses are invalid since it cannot be determined if the signal was arising from the drug or its degradation products. In this case, the weeks spent collecting the original data were wasted because one of the logical controls was not performed early in the measurement process, and more time must now be spent finding a method that is capable of discriminating the drug from its degradation products.

Case 6. The measurement of the enthalpy of transfer by headspace gas chromatography.

In headspace gas chromatography, a gas phase and liquid phase are contained in a vessel sealed with a septum and allowed to reach equilibrium. Typically, the gas phase (headspace) is sampled by piercing the septum with a syringe. In this example, assume that the bulk liquid phase is water to which a very small amount of a nonpolar organic analyte has been added. At equilibrium, some of the analyte molecules will reside in the gas phase and some in the liquid phase. The amount of analyte in the gas phase is determined by sampling the headspace with a syringe and injecting it into a gas chromatograph. The peak area, in combination with a set of standards, yields the gas phase concentration of the analyte. Since water is the bulk solvent, and not very volatile, the little bit of water vapor withdrawn from the vessel will not change the total amount of water in the vessel by any appreciable amount. But what happens to the analyte concentration after repeated samplings, especially if the molecule were very hydrophobic and thus preferentially partitioned into the gas phase? Before we think about that question, we need to finish describing the experiment. As stated in the case study title, the goal is to determine the enthalpy of transfer of the analyte, not just the amount of analyte in the gas phase at a single temperature. This is done by measuring the amount of analyte at equilibrium in the gas phase as a function of temperature. So the entire process for this single determination of the enthalpy of transfer requires a whole set of measurements. Specifically, the experiment is carried out by running three replicates at each temperature and then raising the temperature by 2 °C. Each temperature takes about three hours and there are about 30 temperatures. Thus an entire series of runs involves more than 90 hours of data collection. Furthermore, there are 90 total samplings,

each of which removes some analyte from the gas phase. This amount is hopefully negligible, but one must be skeptical and design the proper controls into the experiment to check for this.

A system suitability check is run at the start of each day's work, some time in the middle of the day, and again at night after the last run of the day. We fully expect that this will detect and correct for instrument drift during the day. However, what about the sample of interest *per se*? How can we be sure that it is not changing during these 90 hours of work (due to over-sampling, escape of the gas-phase analyte from the analysis vessel, analyte adsorption onto surfaces, analyte decomposition at higher temperatures, etc.)? If the sample changes during the week it takes to collect all of the data, then the data is meaningless. The best way to do this experiment is to build in some controls. Here is what is done. First, acquire the data at 26 °C, then lower the temperature to 0 °C and start the run. Next, include 26 °C as one of the intermediate temperatures. Make sure that you have agreement with the first run at 26 °C. If not, quit now because the experiment is no good and to continue is a waste of time. If everything is fine, continue the experiment to the upper temperature limit. Once you have reached the upper limit, reset the temperature to 26 °C and run again. If you have agreement, you have added tremendous confidence to your data. More importantly, if the control measurements at 26 °C were not done, the data would be unreliable and therefore unusable. You would have no way of knowing if the changes in concentration you observed were related to the temperature at all, or if they instead resulted from something else like adsorption onto the vessel, or over-sampling of the headspace (that is, removing more than a negligible amount of the total analyte). If something like over-sampling occurs, the measurements made later in the experiment are being done on a system that is different than the one at the start of the experiment, resulting in a worthless set of data. These controls would catch this problem. The necessary adjustments could be made (i.e. sample less gas phase in each analysis), and the experiment could be restarted without wasting a considerable amount of time. In the worst-case scenario, in the absence of controls, the experimenter would not even realize that the data are faulty and publish incorrect results and conclusions based upon them.

Case 7: Organic Reaction Mechanism Studies

The above case studies are largely related to issues in analytical chemistry. Naturally, however, all areas of chemistry benefit from healthy skepticism and the use of control experiments. Roald Hoffman provides an excellent example of this in his book, *The Same and Not the Same* (32). In chapter 29, he discusses studies aimed at investigating the mechanism of the photolysis of ethane, and the experiments that were conducted to eliminate postulated mechanisms. While there are a number of examples that can be provided related to organic chemistry, Hoffman's treatment is precisely in line with the material and intent of this article, so we direct the interested reader to his work.

A Few Recommendation for Further Immersion into Science

It is important to note that the cases clearly do not encompass all or nearly all of the issues involved in the overall design of sound experiments, much less the general process of science. Our goal here has been to emphasize the use of careful controls, and checks. Full experimental designs also need to consider issues related to sampling, reproducibility, system optimization, randomization, factorial design, and numerous other aspects. In this regard, we highly recommend that any who has not read E. Bright Wilson's classic *An Introduction to Scientific Research* do so (11) and we strongly recommend *Statistics for Experimenters* by Box,

Hunter, and Hunter (13), especially the chapters in both books which relate to experimental design, factorial analysis, data analysis, statistics, and the reporting of scientific results.

Of course, in emphasizing doubt, skepticism, the things that can go wrong, and all of the constraints that must be built into experiments, we risk ignoring or de-emphasizing the enjoyment derived from successful experiments. In this regard, we also highly recommend that students be directed to sources that describe scientific creativity (32), provide first-hand accounts of scientific advances (33), and unabashedly underscore the excitement generated by uncovering the secrets of nature (8).

Conclusion

Reliable scientific conclusions can only be reached by questioning results and continuously anticipating sources of bias and error in measurements. In this regard, the importance of dummy analyses, blanks, controls, and standards cannot be underestimated. The rational incorporation of controls into experimental protocols ultimately prevents the collection of ambiguous or even misleading data. Reliable measurements that contribute to scientific conclusions can only be made when these techniques are used in combination with a consideration of experimental uncertainty (i.e. reproducibility) and higher-level experimental design concerns. It is therefore imperative that students learn about these issues and practice incorporating these techniques into their laboratory work and research projects early in their scientific careers.

Acknowledgments

We are deeply indebted to our friends and colleagues --Professors Tony Borgerding, Colin Cairns, Chuck Lucy, Larry Potts, Steve Weber, for their contributions and stimulating insights; we do not agree with all of their comments nor are they in full agreement with the approach herein. Their comments and suggestions broadened our perspective and strengthened nearly every aspect of this manuscript.

References

1. Popper, K. R. *Objective Knowledge: An Evolutionary Approach*; Oxford University Press: Oxford, 1972.
2. Popper, K. R. *Popper Selections*; Miller, D. (Ed.); Princeton University Press: Princeton, 1985; Parts I and II.
3. Popper, K. R. *Conjectures and Refutations: The Growth of Scientific Knowledge* 5th ed.; Routledge: London, 1989.
4. Russell, B. *Logic and Knowledge*; Marsh, R. C. (ed.); George Allen and Unwin Ltd.: London, 1956.
5. Eames, E. R. *Bertrand Russell's Theory of Knowledge*; George Allen and Unwin Ltd.: London, 1969.
6. Medawar, P. B. *Advice to a Young Scientist*; Harper Collins Publishers: United States of America, 1979; Chapters 9 and 11.
7. Chalmers, A. F. *What is This Thing Called Science?* 2nd ed.; Open University Press: Milton Keynes, 1986.
8. Feynman, R. P. *The Pleasure of Finding Things Out*; Robbins, J. (Ed.); Perseus Publishing: Cambridge, MA, 1999; Chapters 1, 4, 6, 8, 10, and 13.
9. Youden, W. J. *Experimentation and Measurement*; U.S. Department of Commerce NIST Special Publication 672, 1997.
10. Young, H.D. *Statistical Treatment of Experimental Data*; McGraw-Hill Book Company, Inc.: New York, 1962.
11. Wilson Jr., E. B. *An Introduction to Scientific Research*; McGraw-Hill Book Company, Inc.: New York, 1952; Chapters 3, 4, 6, 7, 8, 9, and 10.
12. Levin, I. P. *Relating Statistics and Experimental Design*; Sage Publications: Thousand Oaks, 1999.
13. Box, G. E. P.; Hunter, W. G.; Hunter, J. S., *Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building*; John Wiley and Sons: New York, 1978.
14. Thompson, M.; Brown, D. W.; Gardner, M. J.; Greenhow, E. J.; Howarth, R. J.; Miller, J. N.; Newman, E. J.; Ripley, B. D.; Williams, A.; Wood, R.; Wilson, J. J. *Analyst*, **1995**, *120*, 2303-2308.

15. Thompson, M.; Brown, D. W.; Evans, W. H.; Gardner, M. J.; Greenhow, E. J.; Howarth, R. J.; Miller, J. N.; Newman, E. J.; Ripley, B. D.; Swan, K. J.; Williams, A.; Wood, R.; Wilson, J. J. *Analyst*, **1995**, *120*, 29-34.
16. Thompson, M.; Wood, R. *Pure & Appl. Chem.* **1995**, *67*, 649-666.
17. Harris, D.C. *Quantitative Chemical Analysis*, 6th ed.; W. H. Freeman and Company: New York, 2003.
18. Bader, M., *J. Chem. Educ.*, **1980**, *57*, 703.
19. Harvey, D. *Modern Analytical Chemistry*; McGraw Hill: New York, 2000.
20. Skoog, D. A.; West, D. M.; Holler, F. J. *Fundamentals of Analytical Chemistry*, 7th ed.; Saunders College Publishing: New York, 1996.
21. Enke, C. G. *The Art and Science of Chemical Analysis*; John Wiley and Sons: New York, 2001.
22. Libes, S. M. *J. Chem. Educ.* **1999**, *76*, 1642-1648.
23. Bell, S.C.; Moore, J. *J. Chem. Educ.* **1998**, *75*, 874-877.
24. Marcos, J.; Rios, A.; Valcarcel, M. **1995**, *J. Chem. Educ.*, *72*, 947-949.
25. Laquer, F.C. *J. Chem. Educ.* **1990**, *67*, 900-902.
26. Fisher, R. A. *The Design of Experiments*; Hafner Publishing Company: New York, 1960.
27. Cobb, G. W. *Introduction to Design and Analysis of Experiments*; Springer Publishing: New York, 1998.
28. Clarke, G. M.; Kempson, R. E. *Introduction to the Design and Analysis of Experiments*; Arnold Publishing: London, 1997.
29. Deming, S. N.; Morgan, S. L. *Experimental Design: A Chemometric Approach* 2nd ed.; Elsevier Science: Amsterdam, 1993.
30. Silberberg, M.S. *Chemistry: The Molecular Nature of Matter and Change* 2nd ed.; McGraw Hill Companies, Inc.: New York, 2000.
31. Stapanian, M.A.; Metcalf, R.C. *J. Chem. Educ.* **1990**, *67*, 623-626.
32. Hoffman, R. *The Same and Not the Same*; Columbia University Press: New York, 1995.

33. Watson, J. D. *The Double Helix: A Personal Account of the Discovery of the Structure of DNA*; Simon and Schuster: New York, 1968.