Experimental design heuristics for scientific discovery: the use of “baseline” and “known standard” controls

LISA M. BAKER AND KEVIN DUNBAR

Department of Psychology, McGill University, 1205 Docteur Penfield Avenue, Montreal, Quebec H3A 1B1, Canada. email: lbaker@psych.mcgill.ca.

(Received 8 March 2000, accepted 10 May 2000)

What type of heuristics do scientists use when they design experiments? In this paper, we analysed the ways biological scientists designed complex experiments at their weekly laboratory meetings. In two studies, we found that one of the key components of experimental design is that specific types of control conditions are used in the service of different goals that are important in scientific discovery. “Baseline” control conditions are identical to the experimental manipulation, except that a key feature, such as a reagent, is absent from the control condition and present in the experimental condition. “Known standard” control conditions involve performing the experimental technique on materials where the expected result is already well known; if the expected result is obtained, the scientist can have confidence that the procedure is working. In Study 1, which analysed transcripts of real-world biology laboratory meetings, we found that scientists used baseline controls when testing hypotheses and known standard controls when focusing on possible error. In Study 2, undergraduate science students were asked to address the goals of hypothesis testing and dealing with potential error as they designed experiments. Like the real-world scientists, science majors proposed baseline controls to test hypotheses and known standard controls to deal with potential error. We argue that baseline control conditions play an important role in hypothesis testing: unexpected results obtained on baseline control conditions can alert scientists that their hypotheses are incorrect, and hence should encourage the scientists to reformulate their hypotheses. We further argue that use of known standard controls is a heuristic that enables scientists to solve an important problem in real-world science: when to trust their data. Both of these heuristics can be incorporated into experimental design programs, thus making it more likely that scientific discoveries will be made.

© 2000 Academic Press

KEYWORDS: ☐ ☐ ☐

Computational modelling has uncovered many important features of scientific reasoning and discovery. Using historical data, interviews with scientists, and experiments on scientific thinking, many of the heuristics that scientists use in their research have been discovered and incorporated into computational systems that can themselves participate in the discovery process. Most computational models of scientific reasoning have addressed the issue of how hypotheses or theories are generated from data (Langley, Simon, Bradshaw & Zytkow, 1987; Valdés-Pérez, 1996; Darden, 1997; Valdés-Pérez & Sleeman, 1997). That is, the models have primarily dealt with the process of induction. For example, computer systems have been used to identify genes and their functions.
(Brutlag, 2000) and to determine pathways underlying chemical reactions (Valdés-Pérez, 1994).

Over the past decade, we have been pursuing a complementary, but different, approach to building models of scientific discovery: investigating scientists as they think, reason, and make discoveries in their own laboratories. We have videotaped and audiotaped scientists reasoning at their weekly laboratory meetings and have uncovered some of the heuristics that they use, such as specific types of analogies (Dunbar, 1997, 2000a), use of unexpected findings (Baker, 1994; Dunbar, 1995, 1997, 1999), and distributed reasoning (Dama & Dunbar, 1999; Dunbar, 1997, 2000b). In this article, we turn to another important aspect of the scientific discovery process—how experiments are designed. Using transcripts of laboratory meetings, we analyse the type of heuristics that scientists use when they design experiments.

The design of experiments has received much less attention in the artificial intelligence literature than the process of scientific induction, even though experimental design has long been recognized as an important part of the scientific discovery process (Wason, 1960; Mynatt, Doherty & Tweney, 1977; Klahr & Dunbar, 1988; Le Grand, 1990; Schraagen, 1993; Schunn & Anderson, 2000). Part of the reason for the lack of attention to experimental design is that experimental design in modern-day science laboratories is extremely complex: for any one experiment, scientists must choose among different techniques and types of experimental apparatus, select various experimental materials and reagents, and specify numerous steps in the experimental protocol. Furthermore, some aspects of the experimental design process are specific to particular areas of sciences, such as the actual techniques and reagents used. Thus, modelling experimental design in specific domains, such as high-energy physics, may not reveal any more general experimental design heuristics. However, certain aspects of the experimental design process may be common to many scientific disciplines and can be incorporated into programs that can design experiments in many different fields.

One of the most extensive artificial intelligence treatments of experimental design to date is that of Friedland and Iwasaki (1985), which modelled the control structure of experimental design. In the Friedland and Iwasaki model, developed as part of the MOLGEN expert system, a set of experimental techniques (paradigms) is identified that could be used to meet the experimental goal of the scientist. Each technique that could be used to meet the goal is initially retrieved as an abstracted or “skeletal” plan, and the system proceeds to instantiate each step, or slot, in the skeletal plan. Three criteria are applied successively to choose among possible slot values: (1) whether the proposed value serves to meet the experimental goal, (2) whether it will work in the experimental environment and (3) whether it is optimal in terms of time, cost, convenience, etc., compared to other possibilities. After each skeletal plan has been refined in this manner, the scientist is presented with possible experiments to choose among. In our initial investigations of experimental design in the field of immunology (Baker & Dunbar, 1996), we found that the scientists tended to use a process similar to that identified by Friedland and Iwasaki.

Scientists design experiments that have both experimental conditions and control conditions (Baker & Dunbar, 1996; Dunbar, 1995, 1997). The experimental conditions usually consist of a manipulation of the variable that the scientist is interested in. For example, if the scientist thinks that a particular protein is responsible for a cell having
a particular function, the scientist might add this protein to cells and to see if the cells gain this function. In addition to the experimental condition, the scientists also include numerous control conditions in most experiments. In one experiment, we observed a graduate student had hypothesized that the gene trz is expressed (that is, transcribed into trz RNA) in rats with a disease called RLE.† This experiment contained an experimental condition and three control conditions. In each condition, the experimenter used a specific biological probe to detect whether the gene trz was being expressed or not. The experimental condition used the trz probe in a rat that had the disease RLE. If the hypothesis that trz is involved in RLE was correct, trz expression would be detected in the sick rat. One control condition tested for trz in a rat that did not have RLE disease. If the hypothesis was correct, trz expression should not be found in the healthy rat.

For the other two control conditions, the scientists used cells that were either infected or not infected with a different disease called LAA. LAA was a very different disease from RLE, but it had an important feature: scientists already knew that trz was expressed during the course of the LAA disease. In the two LAA control conditions, a trz probe was used on cells induced with LAA and on cells not induced with LAA. Because it was known that trz was expressed during LAA, it was expected that the LAA-induced cells would be positive for trz expression and the LAA-uninduced cells would be negative for trz expression. If the expected results were obtained for these two control conditions, the scientists would have confidence that their probe and experimental system were capable of detecting trz expression when and only when it occurred.

The experiments we observed appeared to contain two types of control conditions, which we labelled “baseline controls” and “known standard controls”. Baseline controls consist of a condition in which something is taken away from the experiment or not added to an experiment (relative to an experimental condition). The control of testing for trz expression in a mouse without the RLE disease would be called a baseline control condition, because relative to the experimental condition the RLE disease was absent. Known standard controls consist of conditions that have previously been used and validated and are standardized. The control conditions using trz probe with LAA-induced and LAA-uninduced cells might be termed known standard controls. They involve using an experimental technique on materials for which the expected result is well known (trz is expressed and not expressed, respectively). The results obtained on the known standard control condition form a standard against which findings on experimental conditions can be judged.

We hypothesized that the two types of control conditions, baseline controls and known standard controls, might be used to meet very different goals. Baseline controls appeared to be directed toward hypothesis testing; they isolated the factors specific to the hypothesis. It was more difficult to determine the purpose of known standard controls, which have not been addressed in previous studies of experimental design. However, our qualitative analyses of known standard controls indicated that this was a strategy the scientists used to respond to possible errors in the experimental system.

To take an everyday example, imagine a cook whose experimental soufflé recipe fails. The cook might try using a known standard control condition the next time she or he

†Some identifying features of the experiment have been changed.
used the recipe. That is, they would use a standard soufflé recipe, using the same oven, baking temperature and other baking conditions, in order to determine whether the soufflé failed because of the new recipe or because of the baking conditions. We hypothesized that the scientists use known standard control conditions, such as running a probe on tissue where the trz gene is known to be expressed, for similar reasons. Results obtained on this control condition might indicate whether the probe was working, the experimental technique for detection was working, the conditions were correct for DNA expression, and so on; in other words, whether error was present in the experimental system. The use of known standard controls appeared to be a way of making it possible to isolate potential problems and errors in an experiment before or as the problems arise.

Our qualitative analyses suggested that the use of known standard controls was related to potential errors in experiments (Baker & Dunbar, 1996). We found that error in experimental feedback is a salient issue in real-world science laboratories. Experimental results are not always what they appear; for example, a negative result could be due to an equipment malfunction or materials degradation rather than the true absence of a hypothesized phenomenon. Despite its prevalence in real-world science, the issue of error also does not arise in most psychology research using “science-like tasks”. However, a few psychology studies have been conducted that address the issue of error in experimental feedback.

Gorman (1986) investigated the effect of telling participants that there might be error in the feedback they received on a science-like task. He found that people often attributed feedback inconsistent with their hypotheses to error and they replicated experiments on which they obtained surprising results. Recently, Penner and Klahr (1996) investigated the effect of error using a variation of the 2–4–6 task in which people were given experimental feedback that was erroneous. Again, people’s reaction to the error was to replicate the experiments. In each of these error studies, the types of tasks used limited the responses of participants to the possibility of error. Error was assumed to occur, or did occur, randomly, and participants had no way of knowing or discovering when error might be introduced. In terms of experimental design strategies, participants’ only recourse was to replicate experiments when they suspected results were in error. It is possible that real-world scientists, who work in a more complex experimentation environment than study participants, develop specific strategies for dealing with error. We hypothesize that the use of known standard control conditions is just such a strategy.

We used a two-pronged strategy to investigate the use of controls. First, we collected new data on the way that control conditions are used by scientists when they design experiments at their lab meetings. Second, we brought science students into our laboratory and asked them to design experiments. We predicted that both the scientists and the students would use baseline controls when designing experiments to test hypotheses. We further predicted that they would use known standard controls when responding to possible experimental error.

1. Study 1: scientists’ use of controls when designing experiments

In Study 1 we investigated experimental design in two real-world biology laboratories at a major Canadian university.
1.1. METHOD

Data: The data for this study were drawn from four laboratory meetings, two meetings each at two different immunology laboratories. At these laboratory meetings, members of the laboratory took turns presenting their current research. The laboratory meetings were tape-recorded and transcribed, and two segments of each meeting were selected for analysis because they involved discussions of experimental designs. The segments were chosen so that four different laboratory members, two from each laboratory, were represented in two segments each. That is, we analysed the design of eight experiments.

Coding scheme: Experimental designs discussed in each segment were coded for the following:

“Experimental conditions”, which were experimental manipulations that directly tested the claim of the current hypothesis.

“Baseline control conditions”, which were the same as the experimental manipulation, except that one or more aspects of the manipulation were missing. Baseline controls might be lacking one piece of the experimental manipulation, such as not using a blocking agent when the experimental condition used a blocking agent. Baseline controls might also be biological, such as using a healthy organism when the experimental condition used a diseased organism.

“Known standard control conditions”, which were conditions where the scientists knew, based on previous work or from the scientific literature, what the result should be. Known standard controls used the same techniques as the experimental manipulation but with materials unrelated to the hypothesis. For example, if the experimental condition uses an antibody to see whether a certain molecule is present on an experimental cell, a known standard control might verify that the antibody could detect the molecule on a cell known to express it.

In addition, the goal of each experimental segment was coded into one of two categories: (1) testing a hypothesis or (2) dealing with potential error or getting an experimental system to work (Turnbull & Stokes, 1990).

1.2. RESULTS AND DISCUSSION

Table 1 shows the number of experimental conditions, baseline controls and known standard controls planned during each experimental design segment. In eight segments, a total of 20 control conditions were planned, nine baseline and 11 known standard, and a total of eight experimental conditions were planned. The number of experimental conditions in each experiment (1 or 2) was the same as the number of variables considered (1 or 2). Multiple variable were never “crossed” in these experiments.

Both baseline and known standard controls were used in these experiments. To investigate when each type of control was used, segments were grouped into two categories according to the major goal of the segment. Segments 5–7 fell into the test hypothesis group, and segments 1–4, and 8 fell into the possible error group. In Table 2, the 20 control conditions listed in Table 1 are grouped according to the goal of the segment in which they occurred. For example, control conditions proposed during segment 1 are placed in the second row (“Possible Error”), because the scientists were primarily concerned with responding to possible error during segment 1. As can be seen in Table 2, most baseline controls were proposed during segments where hypothesis
TABLE 1
Experimental and control conditions proposed by real-world scientists

<table>
<thead>
<tr>
<th>Segment number</th>
<th>Lab number scientist number</th>
<th>Experimental conditions</th>
<th>Baseline controls</th>
<th>Known standard controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 1</td>
<td>2 4</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2 1</td>
<td>0 2</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>3 1</td>
<td>0 1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4 1</td>
<td>0 1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5 2</td>
<td>1 1</td>
<td>1</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>6 2</td>
<td>1 2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>7 2</td>
<td>1 1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>8 2</td>
<td>2 1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

TABLE 2
Control conditions proposed by real-world scientists for different goals

<table>
<thead>
<tr>
<th>Goal</th>
<th>Baseline</th>
<th>Known standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test hypothesis</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Possible error</td>
<td>2</td>
<td>7</td>
</tr>
</tbody>
</table>

testing was the primary goal, and most known standard controls were proposed during segments where dealing with possible error was the goal, \( \chi^2(1, N = 20) = 3.43, p = 0.09. \)

Overall, the results of this study indicate that real-world scientists were more likely to use baseline controls when testing a hypothesis and to use known standard controls when error was suspected in experimental feedback. Study 2 was designed to manipulate experimental goals and determine the effect of goals on type of control conditions used.

2. Study 2: goals and type of control condition used

In Study 2, undergraduate non-science and science majors were asked to design experiments. They were initially given a hypothesis-testing goal and were then switched to a goal of dealing with possible error. Again it was predicted that participants would use baseline controls when given a hypothesis-testing goal and known standard controls when dealing with possible error.

2.1. METHOD

Participants: There were three groups of paid participants in this study. The first group was 30 undergraduates recruited from non-science courses, none of whom were science majors.

\( \dagger \) Segment 8, which had a goal of solving a technical problem, resulted in one experimental condition and no control conditions being proposed. In that segment, the researcher’s response to technical problems was to repeat the experiment with a modified experimental protocol.
majors nor had significant scientific or statistical training. The second group was 30 undergraduates recruited from an upper-level developmental biology course. The third group was 30 undergraduates recruited from an upper-level immunology course. All students in the second and third groups were science majors. Of the 60 science participants, 42 had taken upper-level laboratory courses. (Participants were in the equivalent of year three of a four-year US university program.)

Materials: We consulted with experts in developmental biology and immunology to construct experimental design problems in those domains. Two different developmental biology problems and two different immunology problems were used, and non-science and science students were given slightly different versions reflecting their different levels of background knowledge.

Each non-science student was given one randomly chosen problem. Each science student was given two problems, a developmental biology problem and an immunology problem. Order of presentation and combination of problems was counterbalanced.

The Appendix contains the text of all problems. The first part of each problem (the test hypothesis prompt) directed participants to design an experiment to test a given hypothesis. After they completed the first part of the problem, participants responded to a possible error prompt.

Coding scheme: Experimental and control conditions were coded as in Study 1. For example, for the first developmental biology problem (see the appendix for the full text), the experimental conditions proposed by participants typically included some manipulation of \( \text{Lmdv} \) (e.g. a procedure to detect \( \text{Lmdv} \) expression). Participants were coded as proposing a baseline control condition if they included a condition that was the same as the experimental condition except that it lacked the \( \text{Lmdv} \) manipulation.

Participants were coded as proposing a known standard control condition if they tested the experimental technique in a situation where they knew it should work (e.g. tested for \( \text{Lmdv} \) expression on tissue known to express \( \text{Lmdv} \)).

Reliability of coding scheme: To assess reliability of the coding scheme, 20% of the data collected from science students in Study 2 was coded by two coders. Inter-coder percentage of agreement was 0.94 for baseline control conditions and 0.94 for known standard control conditions. Cohen’s \( \kappa \) was 0.63 for baseline controls and 0.73 for known standard controls (Bakeman & Gottman, 1986). This is exceptionally good agreement for an abstract coding scheme such as that employed in this study. This level of agreement indicates that confidence can be placed in the objectivity of the coding scheme. Findings reported in this paper are based on the codings of one of the two coders involved in the reliability assessment.

2.2. RESULTS AND DISCUSSION

Each response included at least one experimental condition. Some of the responses given by science students included multiple experimental conditions because the participant proposed more than one technique for testing the hypothesis; in these cases, an experimental condition was proposed for each technique.

§A third prompt that appeared between the two others directed participants to focus on alternate hypotheses. Results relating to this prompt are not relevant to this study.
Control conditions proposed in response to each prompt are shown in Table 3. A total of 78 control conditions were proposed in 120 responses by science students. Science students were more likely to propose baseline controls in response to the test hypothesis prompt, and they were more likely to propose known standard controls in response to the possible error prompt, \( \chi^2(1, N = 78) = 45.97, p < 0.0001 \). There were no differences between the developmental biology and immunology students in the types of controls conducted, nor were there differences between controls proposed for developmental biology and immunology problems. Turning now to non-science students, a different pattern emerged. A total of 15 control conditions were proposed in 30 responses by non-science students. Non-science students proposed baseline controls exclusively in response to the test hypothesis prompt. They proposed a small number of known standard controls and these were split evenly between the two prompts.

Thus, science students used baseline controls to test hypotheses and known standard controls when confronted with possible error, independent of scientific domain. Non-science students used baseline controls to test hypotheses. However, non-science students did not propose known standard controls frequently or systematically, suggesting that the greater scientific knowledge or training of science students facilitated their use of this strategy. Known standard controls supplemented but did not replace experimental conditions and baseline controls in the experimental designs of science students.

The major difference between findings of Study 1 and Study 2 was that the real-world scientists used proportionately more control conditions in each experiment than the science undergraduates. The scientists included 20 control conditions in eight experiments (a mean of 2.5 controls per experiment), whereas the science students included 78 control conditions in 120 experiments (a mean of 0.65 control conditions per experiment). However, both groups used the same types of control conditions to meet the same goals.

### Table 3
Control conditions proposed by non-science and science undergraduates for different goals

<table>
<thead>
<tr>
<th>Goal</th>
<th>Baseline</th>
<th>Known standard</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Non-science students</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test hypothesis</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>Possible error</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td><strong>Science students</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test hypothesis</td>
<td>31</td>
<td>4</td>
</tr>
<tr>
<td>Possible error</td>
<td>5</td>
<td>38</td>
</tr>
</tbody>
</table>

### 3. General discussion

The results of these two studies reveal that different types of control conditions are used in the same experiment to meet two different goals. Baseline controls are used for ruling out alternate hypotheses, and known standard controls for dealing with error. Baseline
controls are similar to the standard “control group” used in psychology studies, when the manipulation in the experimental condition is absent. This type of control condition plays an important role in hypothesis testing. When the experiment proceeds as expected, results obtained on baseline control conditions, taken together with the results obtained on experimental conditions, provide convincing evidence for the scientist’s hypothesis. Using the example presented in the introduction, the scientist predicts $trz$ expression in the mouse that has RLE disease (experimental condition) but not in the healthy mouse (baseline control). If the scientist obtains the predicted results, a convincing argument can be made that $trz$ plays a role in RLE. Unexpected results obtained on a baseline control condition can also provide important information, indicating that the current hypothesis is not correct. In the example above, if $trz$ expression is observed in the healthy mouse as well as the sick mouse, the scientist might conclude that $trz$ expression is not a unique factor in RLE. Systematic use of baseline control conditions is thus an important heuristic for the hypothesis-testing process in scientific discovery. This finding suggests that discovery programs that design experiments should generate baseline controls.

The other type of control condition investigated in this paper was the known standard control condition. It was found that real-world scientists and science undergraduates used known standard control conditions when dealing with possible error. The use of known standard control conditions is an experimental design strategy that is apparently common in the sciences, yet it has not been identified in previous studies of scientific reasoning. Previous work with science-like tasks had found that undergraduates either replicated experiments in the face of error or else discounted surprising findings (Gorman, 1986, 1989; Penner & Klahr, 1996). These psychology studies presented participants with randomly occurring error, and assessed the effect of error on participants’ data interpretation and induction strategies. Participants’ possible strategies to change experimental designs to deal with error were limited by the randomly occurring nature of the error. Random error does occur in some real-world science settings, but there are many sources of error that have an identifiable cause (Gibbings, 1986). As Sleeman (1990) has argued, an important challenge for scientists is to distinguish between results due to error and results of theoretical importance; it may also be difficult to distinguish between different types of error that can occur. What we have found is that scientists and science students use known standard control conditions to anticipate different types of potential error, to determine whether error is occurring in the experimental system, and to identify sources of error.

The use of known standard controls makes it possible for scientists to determine the cause of an unexpected finding. Scientists frequently obtain unexpected findings (Dunbar 1995, 1997, 1999), and an important problem in dealing with unexpected findings is deciding what to do with the findings. By comparing the results of an unexpected finding to the results for known standard controls, the scientists can quickly determine whether the unexpected finding was due to some sort of methodological problem. The use of known standard controls in interpreting unexpected findings is important, as scientists’ first reaction to an unexpected finding is that the finding was due to a methodological problem (Dunbar, 1999). This then allows the scientists to fix their experiment without spending an inordinate amount of time trying to uncover the source of problem. If the known standard controls suggest that there is no methodological problem, then the
scientists can begin to propose new models and hypotheses to account for the unexpected results.

Use of known standard control conditions may be expected to arise when experimental techniques are very complex and experimental findings (e.g. a number of an image on a gel) are many steps removed from the phenomena under investigation. When techniques call for complex protocols with many steps, error can occur at many points in the process of running the experiment. Without a strategy like use of known standard control conditions, the scientist may not be aware that error has occurred. To date, most psychological and computational models of experimental design have not addressed the issue of error. However, the potential for error is a constant issue in scientific laboratory work, an issue that must be dealt with before scientists can trust data enough to base theories on it. The finding that non-science students rarely used the known standard controls heuristic again suggests that scientists’ familiarity with complex experimental designs and their potential for error is a key factor in the use of these types of controls.

The findings presented in this paper are based on our codings of the scientists’ goals. We were also interested in whether the scientists stated their reasons for using certain types of control conditions. However, we found that the scientists we investigated rarely articulated explicitly their reasons for using various control conditions; instead control conditions appeared to be included in their experimental designs as a matter of course. An exception is a passage in one laboratory meeting in which the scientists clearly expressed why they were using certain known standard control conditions. The quotations that follow (with terminology modified to preserve the anonymity of the laboratory), refer to the RLE example given in the introduction. The known standard control conditions in this example involved using trz probe in cells that were induced or not induced with the disease LAA. In the quotations, the laboratory’s principal investigator is speaking:

I think at the very very least, and we’re really I’m talking about the bottom before we start discussing the experiment properly, there ought to be LAA-induced versus uninduced. Which you know already gives you a signal versus no signal. I think that’s got to be there minimum, and then we can interpret from there.

Later in the meeting, the principal investigator said:

If the probe lights up uninduced LAA it raises questions about its specificity or about some aspect of the protocol. Because you’ve told us before that uninduced LAA don’t light up with this probe. So it’s uh... It’s a it’s a nice control.

The principal investigator argues that including the known standard control conditions (LAA-induced and LAA-uninduced) is necessary to be able to “interpret” results on the experimental condition, and that in particular an unexpected result on the control condition using uninduced cells would indicate an error involving the experimental materials or protocol. The scientist’s comments about why the LAA-induced and LAA-uninduced control conditions were used are consistent with our argument that known standard control conditions are intended to detect and identify sources of error in the experimental system.

The distinction we have made between baseline and known standard control conditions is, to our knowledge, a new one. Scientists do not appear to explicitly refer to the
same distinction. However, scientists do appear to make this distinction in practice, using these different types of control conditions to meet different goals. We believe it is important for those who study science to make the distinction between these types of control conditions because they are so closely tied to experimental goals: baseline controls to hypothesis testing and known standard controls to identifying possible error. Only by making this distinction can appropriate control conditions be included in computational models of experimentation and in psychological accounts of scientific reasoning.

To the extent that the scientists we studied articulated a distinction between different types of control conditions, the distinction most often made was between “positive” and “negative” control conditions. In a positive control condition (e.g. the LAA-induced cells), the result (trz expression) is predicted to be the same as the result on the experimental condition. In contrast, the result predicted for the experimental control condition is expected to be absent for the negative control condition (e.g. LAA-uninduced cells). The categories of negative and positive control conditions are not the same as those we have identified, baseline and known standard controls. For instance, both the positive and negative controls in this example are known standard control conditions. In our data, most but not all baseline control conditions could be classified as negative controls; known standard controls were mixed between positive and negative. Grinnell (1992) argues that negative controls “determine nonspecific background” and positive controls “determine the level of full activity in the system under investigation” (p. 34). In our analyses, distinguishing between positive and negative control conditions was not as fruitful as distinguishing between baseline and known standard controls, in terms of identifying scientists’ experimental goals and strategies. However, this distinction may yet prove to be an important one.

In our conversations with scientists and our surveys of scientific laboratory courses (Baker & Dunbar, 1999), we found little evidence of formal instruction about the use of control conditions. The scientists themselves do not articulate the same distinction we have made between baseline and known standard control conditions, and there is apparently no alternate formal logic of control conditions that is explicitly taught. Laboratory courses are the primary means by which undergraduates are instructed in experimentation, but in these courses the emphasis is on learning experimental techniques rather than designing the multi-condition experiments common in scientific laboratories. Graduate students learn about the use of control conditions on the job, rather than through formal instruction. Given the ubiquitous use of control conditions in real-world laboratories and their importance in interpreting experimental results, more formal training in the use of control conditions, at the undergraduate or graduate level, would be valuable. This training should emphasize the different goals that different types of control conditions can be used to meet.

Overall, the results of the two studies presented here demonstrate that the use of baseline and known standard controls are key heuristics in experimental design and scientific discovery. Scientists use both types of control conditions because they know that there are many processes that may be at work in the systems they are investigating. They also know that there are numerous potential technical problems that can be isolated by having the appropriate controls. Both of these types of control condition heuristics can easily be incorporated into scientific experimental design programs. The
use of these heuristics may speed the process of scientific discovery by making it easier to diagnose errors in a system using known standard controls, and by uncovering new phenomena when unexpected findings are obtained in baseline control conditions. These results demonstrate that models of discovery must focus both on experimental design and data interpretation to realize the true potential of computational discovery programs.

This research was originally conducted as part of the first author’s doctoral dissertation. Preparation of this article was supported in part by a McGill Major Fellowship to Lisa M. Baker and by NSERC Grant OGP0037356 and SSHRC Grant 410 95 1206 to Kevin Dunbar. We gratefully acknowledge the assistance of the following biology professors and graduate students who assisted in the design of the science problems: Mike Deyholos, Sandra Demaries, Vincent Dodelet, Pascal Lachance, Paul Lasko, and Trevor Owens. We also wish to thank David Klahr, Christian Schunn, and two anonymous reviewers for comments on earlier drafts of this manuscript.

References


Appendix: Problems for Study 2

"Test Hypothesis" Prompts, Problems for Science Students

First Developmental Biology Problem: You are an honors student in Professor Tremblay’s lab. Based on earlier research, Professor Tremblay hypothesizes that the gene lmdv, which is involved in limb development in mice, is only expressed in the cleavage and blastula stages of development. lmdv is one of a family of genes sharing a common sequence motif, all of which have been cloned and are available (including lmdv). You have been asked to design an experiment to test Professor Tremblay’s hypothesis. You should include all details of the experimental design that you can. Give reasons for your choices about the experimental design.

Second Developmental Biology Problem: You are honors student in Professor Tremblay’s lab. Based on earlier research, Professor Tremblay hypothesizes that the function of gene grw in Xenopus is to make the legs grow longer. grw is one of a family of genes sharing a common sequence motif, all of which have been cloned and are available...
Members of the lab have not been able to create a knockout or make a mutation to study the effect of the gene grw. You have been asked to design an experiment to test Professor Tremblay’s hypothesis, without modifying the genomic DNA. You should include all details of the experimental design that you can. Give reasons for your choices about the experimental design.

First Immunology Problem: You are an honors student in Professor Tremblay’s lab. There is a mouse in Professor Tremblay’s laboratory with a non-infectious illness. Based on earlier research, Professor Tremblay hypothesizes that the cause of this disease is a self-antigen protein called SP3. SP3 is one of a family of self-antigen proteins, and you have samples of all these proteins available (including SP3). You have been asked to design an experiment to determine whether the sick mouse has antibodies to one of the suspected antigen proteins in its blood serum. You should include all details of the experimental design that you can. Give reasons for your choices about the experimental design.

Second Immunology Problem: You are an honors student in Professor Tremblay’s lab. Professor Tremblay is interested in the pattern of surface cell activation in cytotoxic T-cells during the effector stage of target cell lysis. Based on earlier research, Professor Tremblay hypothesizes that the CD197 molecule on the surface of the cytotoxic T-cell acts as a marker during lysis of cells infected with the Antipop virus. CD197 is one of a family of closely related CD markers, and you have antibodies for all these markers available (including CD197). You have been asked to design an experiment to test Professor Tremblay’s hypothesis. You should include all details of the experimental design that you can. Give reasons for your choices about the experimental design.

“Possible Error” Prompt for All Problems. Science Students
You are worried that if your experiment doesn’t give clear results, Professor Tremblay will think you didn’t do it correctly. Without switching to a whole new experimental technique, is there anything you want to modify or add to your experimental design to show that the technique is working properly? Give reasons for your choices.

“Test Hypothesis” prompts, problems modified for non-science students
First Developmental Biology Problem: A friend of yours who works in Professor Tremblay’s biology laboratory has asked for your help. Your friend is an honors student in Professor Tremblay’s lab. Professor Tremblay has been researching a gene called lmdv in mice. It is known that lmdv is involved in limb development. Based on this earlier research, Professor Tremblay hypothesizes that lmdv is only switched on during certain stages of development of mouse embryos. He hypothesizes that lmdv is switched on when the fertilized egg starts to divide into multiple cells, and is switched off again when the cells start to differentiate into the different kinds of cells that make up parts of the mouse’s body. lmdv is one of a family of closely related genes. Your friend has been asked to design an experiment to test Professor Tremblay’s hypothesis. She knows there is a technique she can use that can usually detect the signal a gene sends when it is switched on. Your friend has asked if you can help her design the experiment. What experiment would you do? You should include all details of the experimental design that you can. Give reasons for your choices about the experimental design.
Second Developmental Biology Problem: A friend of yours who works in Professor Tremblay’s biology laboratory has asked for your help. Your friend is an honors student in Professor Tremblay’s lab. Professor Tremblay has been researching a gene called grw in a frog species called *Xenopus*. Professor Tremblay hypothesizes that the function of the gene grw is to make the legs of the frog grow longer. grw is one of a family of closely related genes. Your friend has been asked to design an experiment to test Professor Tremblay’s hypothesis. She knows that there is a technique she can use that can usually block the signal a gene sends to other parts of the body. Your friend has asked if you can help her design an experiment. What experiment would you do? You should include all details of the experimental design that you can. Give reasons for your choices about the experimental design.

First Immunology Problem: A friend of yours who works in Professor Tremblay’s biology laboratory has asked for your help. Your friend is an honors student in Professor Tremblay’s lab. There is a mouse in Professor Tremblay’s laboratory with an illness that does not transmit infectiously to other mice. Based on earlier research, Professor Tremblay hypothesizes that the cause of this disease is that the body is producing a protein, and that the immune system is mistakenly attacking this protein. Professor Tremblay thinks the protein involved in the disease is SP3. SP3 is one of a family of closely related proteins that are known to be involved in diseases like this. Your friend has been asked to design an experiment to determine whether the sick mouse has antibodies to one of the suspected proteins in its blood. She knows there is a technique she can use that can usually detect whether antibodies to a specific protein are present. Your friend has asked if you can help her design the experiment. What experiment would you do? You should include all details of the experimental design that you can. Give reasons for your choices about the experimental design.

Second Immunology Problem: A friend of yours who work in Professor Tremblay’s biology laboratory has asked for your help. Your friend is an honors student in Professor Tremblay’s lab. Professor Tremblay studies cells called T-cells that have the ability to lyse (break open and kill) other cells in the body that are infected with a virus. Professor Tremblay is interested in which molecules on the surface of the T-cells get “turned on” while the T-cell is in the process of lysing an infected cell. Based on earlier research, Professor Tremblay hypothesizes that the CD197 molecule on the surface of the T-cell acts as a marker during lysis of cells infected with the *Antiopan* virus. CD197 is one of a family of closely related CD markers. Your friend has been asked to design an experiment to test Professor Tremblay’s hypothesis. She knows there is a technique she can use that can usually detect whether a specific marker is present on a cell. Your friend has asked if you can help her design the experiment. What experiment would you do? You should include all details of the experimental design that you can. Give reasons for your choices about the experimental design.

“Possible Error” prompt for all problems, non-science students
Your friend is worried that if your experiment doesn’t give clear results, Professor Tremblay will think she didn’t do it correctly. Is there anything you want to modify or add to your experimental design to show that the technique is working properly? Give reasons for your choices.
### Queries and/or remarks

<table>
<thead>
<tr>
<th>Manuscript Page/line</th>
<th>Details required</th>
<th>Author’s response</th>
</tr>
</thead>
<tbody>
<tr>
<td>3/17</td>
<td>Baker (1994) - not in ref. list</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Brutlag, Baker &amp; Dunbar (1999)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dama &amp; Dunbar (1999)</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Dunbar (1999) - pl. update</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Dunbar (2000a), (2000b) - pl. update</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Schunn &amp; Anderson - pl update</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Simon et al. (1997) - pg. no. pl.</td>
<td></td>
</tr>
</tbody>
</table>