Behavior Research Methods, Instruments, & Computers
1996, 28 (2), 305-310

COMPUTER-AIDED METHODS AND STATISTICAL ANALYSIS

Chaired by Russell M. Church, *Brown University*

Correction of errors in scientific research

RUSSELL M. CHURCH and JONATHON D. CRYSTAL
Brown University, Providence, Rhode Island

and

CHARLES E. COLLYER
University of Rhode Island, Kingston, Rhode Island

Four ways to reduce scientific errors are by tests of equipment and programs, examination of results, peer review, and replication. This article describes various types of errors that may occur and procedures available for the prevention and correction of both unintentional and intentional errors in experiments that use computer programs to generate the stimuli, record the responses, or analyze the data. We describe a case study of a particular experiment that produced a result that has been found to be erroneous. The case study provides additional evidence of the essential importance of replication for the identification and elimination of scientific error.

When computers were introduced into the psychological laboratory, initially for data analysis and later for experimental control and recording of data, they were thought to be devices that would serve to reduce human error. But it was soon recognized that errors can occur even with a computer, and there are some errors that occur only because a computer is being used.

Types of Errors

Computers are now used at all stages of research projects (Church, 1993), and errors may occur at any of these stages. The stages include a search of the literature, the design of the experiment, the control of the experiment, data acquisition, data retention, data analysis, the development of theory, the drafting of figures, and the preparation of the manuscript. Examples of errors include the following: (1) An incomplete literature search may lead to a failure to cite a relevant reference. (2) Errors in experimental design can lead to a confounded variable being present. (3) Errors in experimental control can result in the intended procedure not being the one that is actually used. (4) Errors in data acquisition can produce data that are not what actually occurred. (5) Errors in data retention can alter the data so that the data analyzed are not the data that were recorded. (6) Errors in data analysis may lead to the intended analysis procedure not being the one that is actually used. (7) Errors in the development of a theory may result in a simulation of a model that is not the one described. (8) Errors in the preparation of figures can produce points that are not accurately located. (9) Typographical errors can occur in the preparation of a manuscript. The procedures for correcting errors at each of these stages are somewhat different. This article will focus on two stages of the research process: data acquisition and data analysis.

Several different types of errors may occur in experimental data acquisition and data analysis programs: The intentions of the investigator may not be accurately translated into specifications for a procedure; these specifications may not be accurately translated into a program; and the program may not be appropriately used to implement the procedure. This article will focus on validation of programs; that is, determining that the program correctly represents the specifications for the program. In addition to unintentional errors that often may be prevented or corrected by good software engineering practices (Sommerville, 1996), there are intentional errors that may be prevented or corrected by software security practices (Icove, Seger, & Von Storch, 1995; National Research Council, 1991). The possible motives for the deliberate introduction of such errors are essentially unlimited. They include personal ambition, a desire to please, revenge, self-

We are indebted to Lorraine Allan for attempting to replicate our early results, and informing us of her failure to do so. We also thank Hilary Broadbent for her thorough analysis of data, and Matthew Kleiderman for preparing the program used in our replication experiment. Correspondence should be addressed to R. M. Church, Department of Psychology, Box 1853, Brown University, Providence, RI 02912 (e-mail: russell_church@brown.edu).

protection, service to others, playfulness, lack of self-control, and various unconscious motivations.

Prevention of Errors

A standard strategy for improving security is to attempt to limit access to the computer or to some of the resources of the computer. A less often used security measure is to attempt to verify that there have been no alterations in a program or database. A precaution with possible security benefits is to attempt to limit the consequences of any corruption of computer programs and data by, for example, making backups and storing copies of the backups in different places. Certain management practices can reduce the opportunity for errors being introduced, and good record-keeping practices can prevent many errors. Investigators should establish procedures, tailored to the needs of the laboratory, in which reasonable precautions are taken to prevent or reveal errors. The merit of such a practice is that it may detect some errors and prevent the introduction of intentional errors.

In general, logically simple, well-documented programs are less likely to contain errors than are more complex programs, and any errors that they may contain are easier to identify. Commercial software used by a large number of individuals is less likely to contain errors than custom programs of equal complexity. The use of higher level languages prevents many unintentional errors because modern compilers contain many well-tested facilities and built-in automatic checks against a large number of common programming errors, but such use also creates opportunities for the introduction of intentional errors, such as the alteration of libraries of standard procedures or runtime packages.

Although investigators should take reasonable measures to prevent errors, such practices cannot prevent all possible errors (Thompson, 1984). The next line of defense is to correct the error, and this should occur at the earliest possible stage of a research project. Errors can be identified and corrected with tests of the equipment and programs, examination of results, peer review, and replications.

Tests of Equipment and Programs

Tests should first be conducted at the level of individual modules and then at the level of complete programs. A standard way to test a data acquisition program is to compare the data recorded by the program to some known input data. The main limitation of this procedure is that it is practical to do this for only a subset of possible input conditions.

Tests of a data analysis program are essential. A standard way to test a data analysis program is, for some particular input data, to compare the output of the data analysis program with the output of a hand calculation. Many types of errors can be identified by such comparisons, including some intentional ones. In any experimental undertaking there will be certain conventions and practices that can be targets for intentional errors. For example, if the name of a data file is never changed, it can be used to seed a random number generator to produce infrequent but deterministic errors. The main limitation of hand calculations is that it is practical to do them only for small data sets that may not be representative of the data that must be analyzed. An alternative is to compare the output of two different data analysis programs with the same input data (Roberts, 1980). The programs should be written by different individuals, preferably in very different computer languages, so that the errors are more likely to be independent. The output of such a pair of programs can be compared on data that are considered to be typical, well specified, or extreme, but there is no guarantee that the program is correct for all possible data.

Examination of the Results

Inspection of the results for plausibility can sometimes identify errors at earlier stages. It may reveal some impossible or unlikely values, or some impossible or unlikely combinations of values. Sometimes the distribution of a value seems implausible, and this can lead to the identification of an error. For example, Gaffan (1992) applied a statistical test that identifies lower than expected binomial variance and may have revealed errors in a procedure or analysis.

Peer Review

If an error has not been prevented, and if it has not been detected during the tests that were conducted by the investigator or by examination of results, the next line of defense is peer review, either formal or informal. A strength of peer review is that experts in the field attempt to evaluate the procedures, results, and interpretations in a critical manner, with an emphasis on identifying any problems. A weakness of peer review is that the experts are not provided with sufficient information to detect most errors. An additional weakness of peer review is that it is difficult to publish experimental replications, particularly failures to replicate. As a consequence, published errors may remain uncorrected.

Replication

Replication is the standard test of the reliability of a result. In the culinary arts one has faith in a recipe that can be used in different kitchens to make very similar cakes; in science one has faith in a method that can be used in different laboratories to produce very similar results. Replications within a laboratory are useful, but replication of research in different laboratories is essential for confidence in a result. A problem with the former is that features other than the published statement of methods may have been common to replications within a laboratory. For example, both experiments may use the same flawed program.

Failures to replicate a result are also useful. Sometimes they may be resolved by the identification of significant differences in the methods used, and these must be shown to be relevant in a further experiment. Sometimes they may lead to the identification of an error in one of the experiments, and it is much easier to find a flaw if one is expected to exist. Occasionally, a failure to replicate cannot be explained as a particular difference in procedure or an error in one of the experiments. In this case, no confidence can be placed in either result until more research is conducted.

A CASE STUDY

Two experiments were recently published that described a small, systematic bias function in continuation tapping, a function that was called "the oscillator signature" (Collyer, Broadbent, & Church, 1992, 1994). We asked whether this function was a specifically motor phenomenon, or reflected biases in the perception of time intervals. In an attempt to determine whether or not there was an oscillator signature in time perception, we developed a simple duration-estimation procedure: a procedure in which a tone was presented for some duration, and the subject moved a mouse to adjust a pointer to a location on a line to indicate the perceived duration of the tone. The subject was instructed to point to the left end of the line for the shortest tone, and the right end of the line for the longest tone, and for other tones, to point to the location on the line in proportion to its perceived duration.

In the first experiment 4 subjects were tested, and two ranges were used: 100–900 and 200–1,000 msec, each with 81 different durations spaced at 10-msec intervals.

The top panel of Figure 1 plots the duration estimates as a function of actual duration. The location on the line, from 0 to 100, increased as a function of the duration of the stimulus, both for the series that ranged from 100 to 900 msec and for the series that ranged from 200 to 1,000 msec. The results indicated that the subject could point on average to the approximately correct location on the line.

To examine local changes in the slope of the psychophysical function, the bottom panel plots first differences (smoothed with a 5-point running mean) as a function of duration. A *first difference* is the difference between the mean location pointed to at each duration minus the mean location pointed to at a duration 10 msec shorter. If the slope were approximately linear, these first differences would have been approximately constant. The systematic pattern of first differences was similar to that of the systematic residuals that had previously been observed in the finger-tapping experiments. The systematic pattern of first differences was found to be a function of the physical duration presented, not the location on the line to which the subject had pointed. These data suggested multiple, systematic departures from linearity in the psychophysical function for time. Similar results were obtained in additional experiments. As we will see, this result was not correct, and the way in which the error was found may be instructive.

Peer Review

Because this result was not published, it did not receive extensive peer review. It was, however, presented at several meetings (e.g., Crystal, Broadbent, Maksik, Collyer, & Church, 1995), and a manuscript was prepared. Those hearing the talks or reading the manuscript did not have sufficient information to detect programming errors.

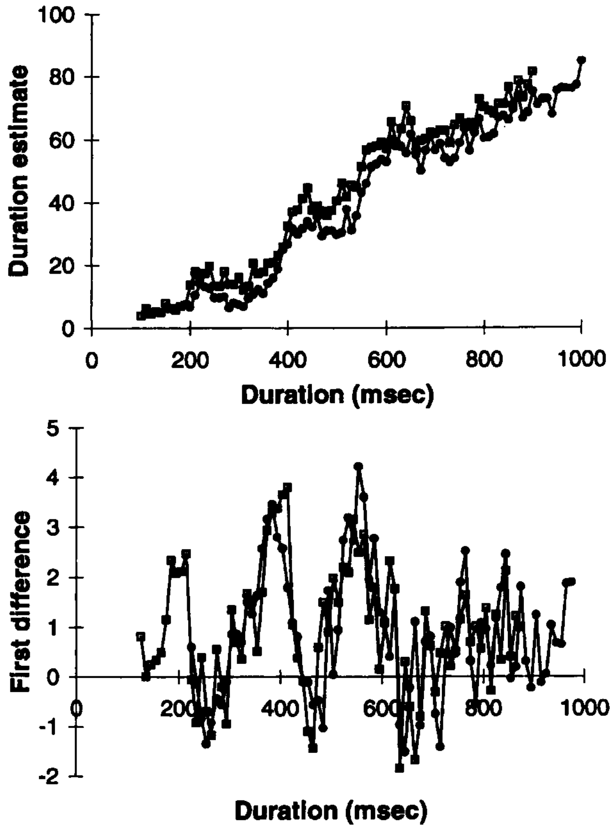

Figure 1. Data from first duration estimation experiment. The top panel plots mean duration estimates as a function of stimulus duration. The bottom panel plots mean first differences (smoothed with a 5-point running mean) as a function of stimulus duration. This first experiment produced systematic departures from a linear psychophysical function for time.

Examination of Results

A result is more likely to be accepted as correct if it confirms prior expectations. This bias, studied by Rosenthal (1966) and others, has been called an "experimenter expectancy effect." According to Bayes's theorem, there is a rational basis for such a bias, but it is one that makes an investigator resist changing his or her mind. Thus, unexpected results will be carefully scrutinized for possible sources of error, and expected results may not receive the same critical attention. A datum near the mean is more likely to be accepted as correct than an extreme outlier; a pattern of results predicted from theory is more likely to be accepted as correct than one that was not expected.

Because of the importance of experiment expectation effects, it is relevant to consider the background of the experiment that produced the erroneous result. It provides a relevant context for the error. As noted, a small, systematic bias function had been observed in continuation tapping. A multiple-oscillator, connectionist model of timing (Church & Broadbent, 1990) had been developed to account for three major facts of interval timing. Subsequent simulations showed that this model produced a small, systematic bias function that was similar to the one observed in continuation tapping. The observation of systematic bias functions was of general interest because of these simulations.

Replication

The essential features of the results shown in Figure 1 were replicated in six experiments in our laboratory conducted by different investigators using different modifications in the basic data acquisition program.

Allan (in press) reported a failure to replicate this result. On the basis of personal communication and examination of data, we concluded that the two laboratories had used essentially the same procedure, but had gotten different results. A replication at the University of Rhode Island, using different equipment and software, verified Allan's results, so we strongly suspected there was an error in our original data collection program. Figure 2 shows the results from the University of Rhode Island replication with 4 subjects. Multiple, systematic departures from linearity were not observed in the psychophysical function for time.

Further Tests

Further tests of our original program were conducted to assess the cause of the replication failure. Oscilloscope measurements verified that the duration of the tones presented corresponded to the durations intended, and these durations were correctly entered into the data file. The subroutine to record the location to which the mouse pointed was a standard one that was correct at both extremes. To verify that the routine that recorded the location on the line to which the subject pointed was correctly recorded when it was included as part of the entire program, a manual record of responses during tests was maintained. An acetate sheet with a ruled line labeled with 100 equally spaced tick marks was taped on the computer monitor. The distance between successive tick marks was approximately 4 pixels. On each trial the tester closed one eye and lined up the reflection of the open eye on the computer monitor at the to-be-selected location to minimize variability in measurements. The tester pointed to a known response location and kept a written record of each response. These tests provided a record of the location of response inputs to compare with the output of the data collection system. The expected difference between these two records of response location is zero, plus random error.

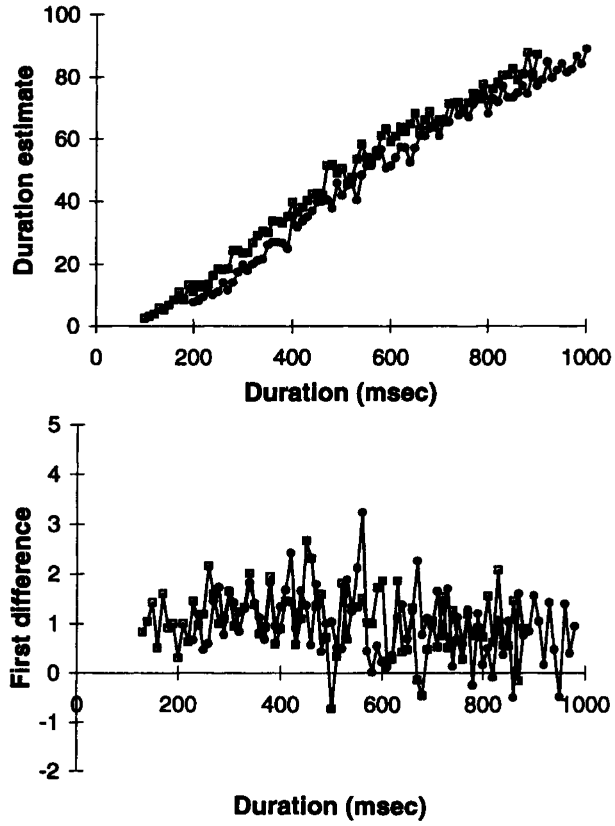

Figure 2. Data from a replication experiment. The top panel plots mean duration estimates as a function of stimulus duration. The bottom panel plots mean first differences (smoothed with a 5-point running mean) as a function of stimulus duration. The replication experiment did not produce systematic departures from a linear psychophysical function for time.

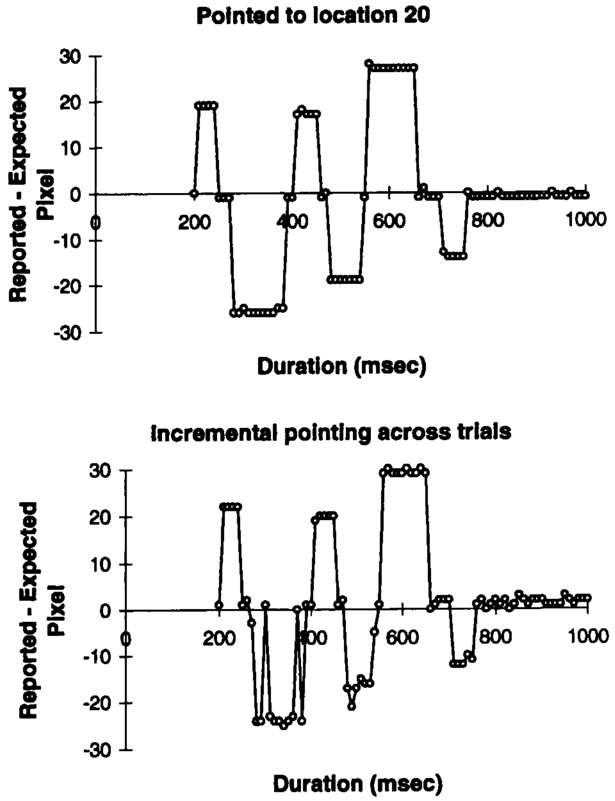

Figure 3. Mean error (reported–expected pixel) in reported response location as a function of stimulus duration with the data collection program used in the first experiment. Responses to a fixed location (top panel) and a random location with respect to stimulus duration (bottom panel) revealed systematic errors in reported response locations as a function of stimulus duration.

In one test the tester pointed to the same location (20 on the ruled line) regardless of the duration of the tone. The top panel of Figure 3 shows that the difference between the pixel location recorded in the file and the pixel location expected on the basis of the hand recording depended on the duration of the stimulus.

In an additional test, the tester pointed to incremental locations on the ruled line regardless of the duration of the tone. Thus, the tester pointed at locations 0, 1, 2, and so on, following successive stimuli that were randomly ordered with respect to duration. Thus the location pointed to was randomly associated with the actual tone duration. The bottom panel of Figure 3 shows that the difference between the pixel location recorded in the file and the pixel location expected on the basis of the hand recording again depended on the duration of the stimulus.

The same two tests were conducted on a second flawed program, and neither of these problems was present. The results of a constant input are shown in panel A of Figure 4, and the results of an incremental input are shown in panel B of Figure 4. In both cases the difference between the pixel location recorded in the file and the pixel location expected on the basis of the hand recording was independent of the duration of the stimulus, just as it should have been. (The small, approximately constant, difference from zero was presumably due to the difficulty of pointing precisely at the intended pixel.)

On a third test, the tester pointed to the location on the ruled line that corresponded to the subjective duration of the tone that was presented. Thus, the tester acted like a subject would act in the experiment, although he also made a hand record of the location of each response. panel C of Figure 4 shows that the difference between pixel location recorded in the file and the pixel location expected on the basis of the hand recording depended on the duration of the stimulus.

On a fourth test, the tester used incremental input for the first half of the session (the first 41 stimulus durations from a random ordering of the 81 stimulus durations) and responses based on subjective duration for the second half of the session. Thus, the first half of the session used an input that would not produce erroneous results (as shown in panel B), and the second half of the session used an input that would produce erroneous results (as shown in panel C). The results for the combined input are shown in panel D of Figure 4: The difference between the pixel location recorded in the file and the pixel location expected on the basis of hand-recorded data again depended on the duration of the stimulus when the last half of the responses was based on subjective durations. The errors were present in both the first and sec-

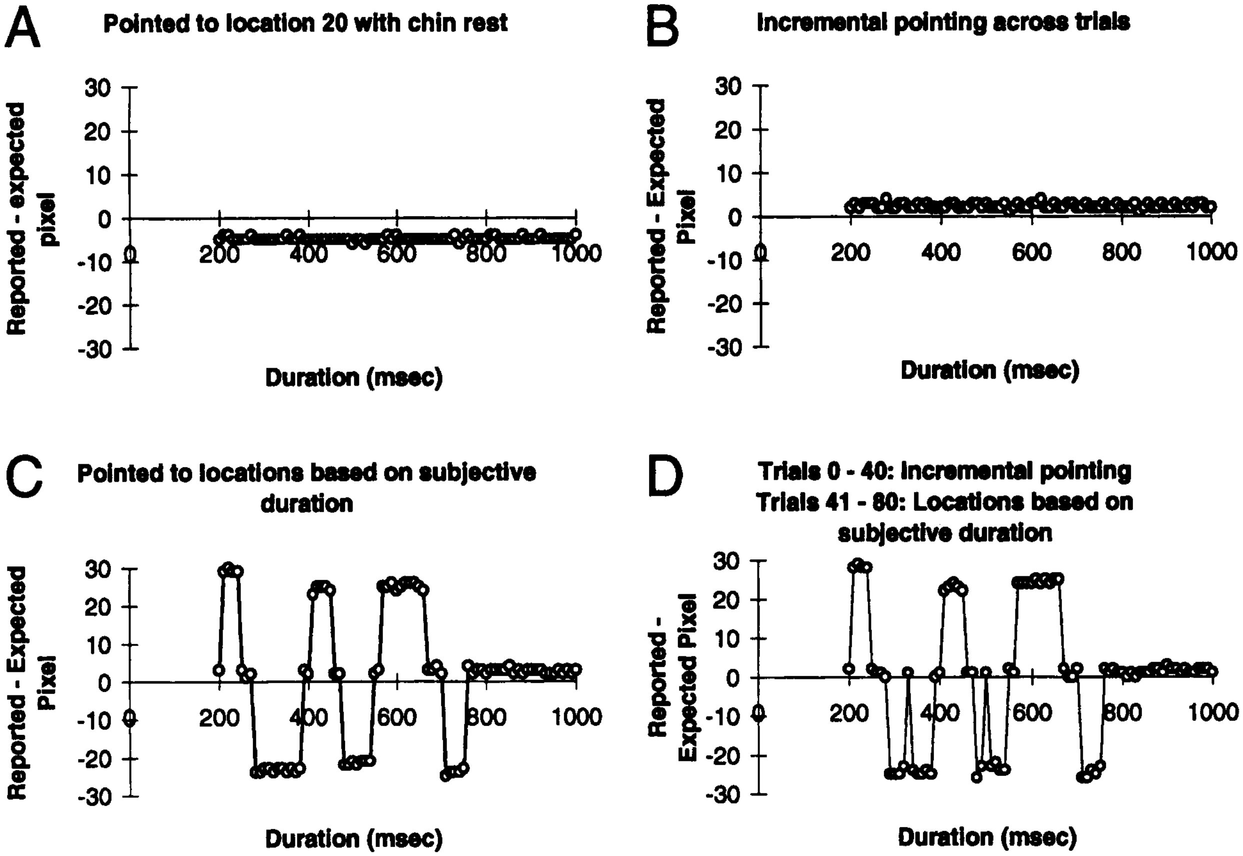

Figure 4. Mean error (reported–expected pixels) in response location as a function of stimulus duration in a second flawed program. Responses to a fixed location (panel A) and a random location with respect to stimulus duration (panel B) did not reveal systematic errors in reported response locations. Responses to locations based on the subjective duration of a stimulus (panel C) revealed systematic errors in reported response locations. Responses to a random location with respect to stimulus duration (as in panel B) in the first half of a test session and to locations based on the subjective duration of a stimulus (as in panel C) in the second half of the session revealed systematic errors in reported response locations throughout the test session (panel D).

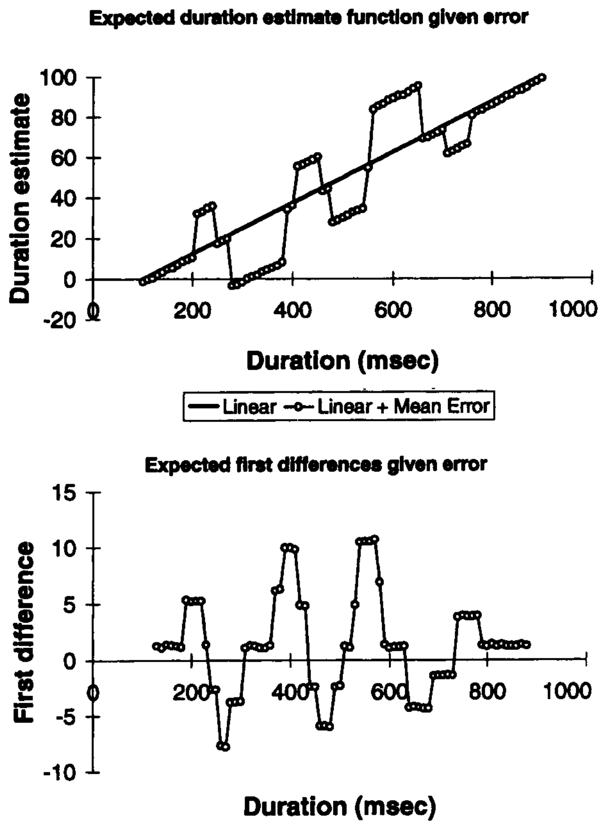

Figure 5. Expected data pattern given error (mean error from four observations at each stimulus duration) using data collection program from the original experiment. The top panel plots duration estimates as a function of stimulus duration. The bottom panel plots first differences (smoothed with a 5-point running mean) as a function of stimulus duration.

ond halves of the session. Thus the program was the source of the systematic bias.

This error was sufficient to account for the data shown in Figure 1 and several other experiments that were conducted with these programs. It is the basis for the discrepancy in the results of the original experiment and the replication experiment. The top panel of Figure 5 shows the expected duration estimate with the known error pattern superimposed on a straight line. The bottom panel of Figure 5 shows the expected first differences. This is qualitatively very similar to our initial results with this flawed software.

It should be noted that these tests would not necessarily detect all errors that might have been present. Tests of a data acquisition system cannot be guaranteed to find all possible errors.

Summary

The purpose of the original experiment was to examine the shape of the psychophysical function of time. On the basis of the results shown in Figure 1, duration estimation increased as a function of duration, but there were local plateaus in the psychophysical function. The plateaus could reflect an important perceptual finding or a scientific error. The tests revealed that the plateaus were due to an error in the data collection program.

CONCLUSION

We have drawn three major conclusions from this example:

1. Investigators should take reasonable precautions to prevent errors, and they should perform reasonable tests to identify errors at all stages of a research project. Although no set of tests can be guaranteed to reveal all errors, tests that resemble the naturally occurring input are likely to be particularly useful.

2. Investigators should not develop scientific neophobia, a fear of new results. In the language of signal detection theory, scientific neophobia is an abnormally high criterion that reduces both hits and false alarms.

3. Investigators should have confidence in the process of replication, particularly replications carried out in different laboratories with different equipment, software, and personnel. Failures to replicate can lead to correction of errors and successful replications can increase confidence in the reliability of a result.

REFERENCES

Allan, L. G. (in press). Psychological time: Continuous or discrete? In C.-A. Possamai (Ed.), *Fechner Day 95*. Cassis, France: International Society for Psychophysics.

Church, R. M. (1993). Uses of computers in psychological research. In G. Keren & C. Lewis (Eds.) *A handbook for data analysis in the behavioral sciences: Statistical issues* (pp. 459-476). Hillsdale, NJ: Erlbaum.

Church, R. M., & Broadbent, H. A. (1990). Alternative representation of time, number, and rate. *Cognition*, **37**, 55-81.

Collyer, C. E., Broadbent, H. A., & Church, R. M. (1992). Categorical time production: Evidence for discrete timing in motor control. *Perception & Psychophysics*, **51**, 134-144.

Collyer, C. E., Broadbent, H. A., & Church, R. M. (1994). Preferred rates of repetitive tapping and categorical time production. *Perception & Psychophysics*, **55**, 443-453.

Crystal, J. D., Broadbent, H. A., Maksik, Y. A., Collyer, C. E., & Church, R. M. (1995, April). *A categorical representation of time in duration estimation.* Paper presented at the annual meeting of the Eastern Psychological Association, Boston.

Gaffan, E. A. (1992). Primacy, recency, and the variability of data in studies of animals' working memory. *Animal Learning & Behavior*, **20**, 240-252.

Icove, D., Seger, K., & Von Storch, W. (1995). *Computer crime: A crimefighter's handbook.* Sebastopol, CA: O'Reilly.

National Research Council. (1991). *Computers at risk.* Washington, DC: National Academy Press.

Roberts, S. (1980). How to check a computer program. *Behavior Research Methods & Instrumentation*, **12**, 155-156.

Rosenthal, R. (1966). *Experimenter effects in behavioral research.* New York: Appleton-Century-Crofts.

Sommerville, I. (1996). *Software engineering* (5th ed.). Reading, MA: Addison-Wesley.

Thompson, K. (1984). Reflections on trusting trust. *Communications of the ACM*, **27**, 761-764.

(Manuscript received December 29, 1995;
accepted for publication January 5, 1996.)