3

Intentional Error

Intentional error production on the part of the experimenter is probably as relatively rare an event in the psychological experiment as it is in the sciences generally (Wilson, 1952; Shapiro, 1959; Turner, 1961b). Nevertheless, any serious attempt at understanding the social psychology of psychological research must consider the occurrence, nature, and control of this type of experimenter effect.

The Physical Sciences

Blondlot's N-rays have already been discussed as a fascinating example of observer effect. Rostand (1960) has raised the question, however, whether their original "discovery" might not have been the result of overzealousness on the part of one of Blondlot's research assistants. Were that the case then we could learn from this example how observer or interpreter effects may derive from intentional error even when the observers are not the perpetrators of the intentional error. This certainly seemed to be the case with the famous Piltdown man, that peculiar anthropological find which so puzzled anthropologists until it was discovered to be a planted fraud (Beck, 1957).

A geologist some two centuries ago, Johann Beringer, uncovered some remarkable fossils including Hebraic letters. "The[se] letters led him to interpret earth forms literally as the elements of a second Divine Book" (Williams, 1963, p. 1083). Beringer published his findings and their important implications. A short time after the book's publication a "fossil" turned up with his name inscribed upon it. Beringer tried to buy back copies of the book which were by now circulating, but the damage to his reputation had been done. The standard story had been that it was Beringer's students who had perpetrated the hoax. Now there is evidence that the hoax was no schoolboy prank but an effort on the part of two colleagues to discredit him

(Jahn & Woolf, 1963). Here again is a case where interpreter effects on the part of one scientist could be in large part attributed to the intentional error of others.

A more recent episode in the history of archaeological research, and one far more difficult to evaluate, has been reported on the pages of *The Sunday Observer*. Professor L. R. Palmer, a comparative philologist at Oxford, has called into question Sir Arthur Evans' reconstruction of the excavations at Knossos (Crete). These reconstructions were reported in 1904 and then again in 1921. The succession of floor levels, each yielding its own distinctive type of pottery, was called by Palmer a "complete figment of Evans' imagination." Palmer's evidence came from letters that contradicted Evans' reconstruction—letters written by Evans' assistant, Duncan Mackenzie, who was in charge of the actual on-site digging. These letters were written after Evans had reported his reconstruction to the scientific public. Evans did not retract his findings but rather in 1921 he reissued his earlier (1904) drawing. Palmer felt that the implications of these events for our understanding of Greece, Europe, and the Near East were "incalculable" (Palmer, 1962). In subsequent issues of *The Observer* Evans had his defenders. Most archaeologists (e.g., Boardman, Hood) felt that Palmer had little reason to attack Evans' character and question his motives, though, if they are right, questions about Duncan Mackenzie's might be implied. The Knossos affair serves as a good example of a possible intentional error which could conceivably turn out to have been simply an interpreter effect—a difference between an investigator and his assistant. One thing is clear, however: whatever did happen those several decades ago, the current debate in *The Observer* clearly illustrates interpreter differences.

C. P. Snow, scientist and best-selling novelist, has a high opinion of the average scientist's integrity (1961). Yet he refers to at least those few cases known to scientists in which, for example, data for the doctoral dissertation were fabricated. In one of his novels, *The Affair,* he deals extensively with the scientific, social, and personal consequences of an intentional error in scientific research (1960). Other references to intentional error, all somewhat more pessimistic in tone than was C. P. Snow, have been made by Beck (1957), George, (1938), and Noltingk (1959).

The Biological Sciences

When, two chapters ago, observer effects were under discussion the assumption was made that intentional error was not at issue. Over the long run this assumption seems safely tenable. However, for any given instance it is very difficult to feel certain. We must recall: (1) Fisher's (1936) suspicion that Mendel's assistant may have deceived him about the results of the plant breeding experiments; (2) Bean's (1953) suspicion

that Leser's assistant may have tried too hard to present him with nearly perfect correlations between harmless skin markings and cancer; (3) Binet's suspicion over his own assistant's erring so regularly in the desired direction in the taking of cephalometric measurements (Wolf, 1961).

One of the best known and one of the most tragic cases in the history of intentional error in the biological sciences is the Kammerer case. Kammerer was engaged in experiments on the inheritance of acquired characteristics in the toad. The characteristic acquired was a black thumb pad, and it was reported that the offspring also showed a black thumb pad. Here was apparent evidence for the Lamarckian hypothesis. A suspicious investigator gained access to one of the specimens, and it was shown that the thumb pad of the offspring toad had been blackened, not by the inherited pigment, but by India ink (MacDougall, 1940). There cannot, of course, be any question in this case that an intentional error had been perpetrated, and Kammerer recognized that prior to his suicide. To this day, however, it cannot be said with certainty that the intentional error was of his own doing or that of an assistant. A good illustration of the operation of interpreter effects is provided by Zirkle (1954) who noted that scientists were still citing Kammerer's data, and in reputable journals, without mentioning its fraudulent basis. More recently, two cases of possible data fabrication in the biological sciences came to light. One case ended in a public exposé before the scientific community (Editorial Board, 1961); the other ended in an indictment by an agency of the federal government (Editorial Board, 1964).

The Behavioral Sciences

The problem of the intentional error in the behavioral sciences may not differ from the problem in the sciences generally. It has been said, however, that at least in the physical sciences, error of either intentional or unintentional origin is more quickly checked by replication. In the behavioral sciences replication leads so often to uninterpretable differences in data obtained that it seems difficult to establish whether "error" has occurred at all, or whether the conditions of the experiment differed sufficiently by chance to account for the difference in outcome. In the behavioral sciences it is difficult to specify as explicitly as in the physical sciences just how an experiment should be replicated and how "exact" a replication is sufficient. There is the additional problem that replications are carried out on a different sample of human or animal subjects which we know may differ vary markedly from the original sample of subjects. The steel balls rolled down inclined planes to demonstrate the laws of motion are more dependably similar to one another than are the human subjects who by their verbalizations are to demonstrate the laws of learning.

In survey research the "cheater problem" among field interviewers

is of sufficient importance to have occasioned a panel discussion of the problem in the *International Journal of Attitude and Opinion Research* (1947). Such workers as Blankenship, Connelly, Reed, Platten, and Trescott seem to agree that, though statistically infrequent, the cheating interviewer can affect the results of survey research, especially if the dishonest interviewer is responsible for a large segment of the data collected. A systematic attempt to assess the frequency and degree of interviewer cheating has been reported by Hyman, Cobb, Feldman, Hart, and Stember (1954). Cheating was defined as data fabrication, as when the interviewer recorded a response to a question that was never asked of the respondent. Fifteen interviewers were employed to conduct a survey, and unknown to them, each interviewed one or more "planted" respondents. One of the "planted" interviewees was described as a "punctilious liberal" who qualified all his responses so that no clear coding of responses could be undertaken. Another of the planted respondents played the role of a "hostile bigot." Uncooperative, suspicious, and unpleasant, the bigot tried to avoid committing himself to any answer at all on many of the questions. Interviews with the planted respondents were tape recorded without the interviewers' knowledge. It was in the interview with the hostile bigot that most cheating errors occurred. Four of the interviewers fabricated a great deal of the interview data they reported, and these interviewers tended also to cheat more on interviews with the punctilious liberal, although, in general, there was less cheating in that interview. Frequency of cheating, then, bore some relation to the specific data-collection situation and was at least to some extent predictable from one situation to another.

In science generally, the assumption of predictability of intentional erring is made and is manifested by the distrust of data reported by an investigator who has been known, with varying degrees of certainty, to have erred intentionally on some other occasion. In science, a worker can contribute to the common data pool a bit of intentionally erring data only once. We should not, of course, equate the survey research interviewer with the laboratory scientist or his assistants. The interviewer in survey research is often a part-time employee, less well educated, less intelligent, and less interested in the scientific implications of the data collected than are the scientist, his students, and his assistants. The survey research interviewer has rarely made any identification with a scientific career role with its very strong taboos against data fabrication or other intentional errors, and its strong positive sanctions for the collection of accurate, "uncontaminated" data. Indeed, in the study of interviewers' intentional errors just described, the subjects were less experienced than many survey interviewers, and this lack of experience could have played its part in the production of such a high proportion of intentional errors. In that study, too, it must be remembered, the design was such as to increase the incidence of all kinds of interviewer effects by supplying unusually difficult situations for inex-

perienced interviewers to deal with. However, even if these factors increased the incidence of intentional error production by 400 percent, enough remains to make intentional erring a fairly serious problem for the survey researcher (Cahalan, Tamulonis, & Verner, 1947; Crespi, 1945–46; Mahalanobis, 1946).

A situation somewhere between that of collecting data as part of a part-time job and collecting data for scientific purposes exists in those undergraduate science courses in which students conduct laboratory exercises. These students have usually not yet identified to a great extent with the scientific values of their instructors, nor do they regard their laboratory work as simply a way to earn extra money. Data fabrication in these circumstances is commonplace and well-known to instructors of courses in physics and psychology alike. Students' motivation for cheating is not, of course, to hoax their instructors or to earn more money in less time but rather to hand in a "better report," where better is defined in terms of the expected data. Sometimes the need for better data arises from students' lateness, carelessness, or laziness, but sometimes it arises from fear that a poor grade will be the result of an accurately observed and recorded event which does not conform to the expected event. Such deviations may be due to faulty equipment or faulty procedure, but sometimes these deviations should be expected simply on the basis of sampling error. One is reminded of the Berkson, Magath, and Hurn (1940) findings which showed that laboratory technicians were consistently reporting blood counts that agreed with each other too well, so well that they could hardly have been accurately made. We shall have occasion to return to the topic of intentional erring in laboratory course work when we consider the control of intentional errors. For the moment we may simply document that in two experiments examined for intentional erring by students in a laboratory course in animal learning, one showed a clear instance of data fabrication (Rosenthal & Lawson, 1964), and the other, while showing some deviations from the prescribed procedure, did not show any evidence of outright intentional erring (Rosenthal & Fode, 1963a). In these two experiments, the incidence of intentional erring may have been reduced by the students' belief that their data were collected not simply for their own edification but also for use by others for serious scientific purposes. Such error reduction may be postulated if we can assume that data collected only for laboratory learning are less "sacred" than those collected for scientific purposes.

Student experimenters are often employed as data collectors for scientific purposes. In one such study Verplanck (1955) concluded that following certain reinforcement procedures the content of conversation could be altered. Again employing student experimenters Azrin, Holz, Ulrich, and Goldiamond (1961) obtained similar results. However, an informal post-experimental check revealed that data had been fabricated by their student

experimenters. When very advanced graduate student experimenters were employed, they discovered that the programmed procedure for controlling the content of conversation simply did not work.

Although it seems reasonable to assume that more-advanced graduate students are generally less likely to err intentionally, few data are at hand for documenting that assumption. We do know, of course, that sometimes even very advanced students commit intentional errors. Dr. Ralph Kolstoe has related an instance in which a graduate student working for a well-known psychologist fabricated his data over a period of some time. Finally, the psychologist, who had become suspicious, was forced to use an entrapment procedure which was successful and led to the student's immediate expulsion.

What has been said of very advanced graduate students applies as well to fully professional scientific workers. It would appear that the incidence of intentional errors is very low among them, but, again, few data are available to document either that assumption or its opposite. Most of the cases of "generally known" intentional error are imperfectly documented and perhaps apocryphal.

In the last chapter there was occasion to discuss those types of interpreter effects which serve to keep certain data off the market either literally or for all practical purposes. It was mentioned that sometimes data were kept out of the common exchange system because no one knew quite what to say about them. Sometimes, though, data are kept off the market because the investigator knows all too well what will be said of them. Such intentional suppression of data damaging to one's own theoretical position must be regarded as an instance of intentional error only a little different from the fabrication of data. What difference there is seems due to the "either-or-ness" of the latter and the "shades of grayness" of the former. A set of data may be viewed as fabricated or not. A set of legitimate data damaging to a theory may be withheld for a variety of motives, only some of which seem clearly self-serving. The scientist may honestly feel that the data were badly collected or contaminated in some way and may therefore hold them off the market. He may feel that while damaging to his theory their implications might be damaging to the general welfare of mankind. These and other reasons, not at all self-serving, may account for the suppression of damaging data. Recently a number of workers have called attention to the problem of data suppression, all more or less stressing the self-serving motives (Beck, 1957; Garrett, 1960; Maier, 1960). One of these writers (Garrett) has emphasized a fear motive operating to suppress certain data. He suggests that young scientists fear reprisal should they report data that seem to weaken the theory of racial equality.

Sometimes the suppression of data proceeds, not by withholding data already obtained, but by insuring that unwanted data will not be collected. In some cases we are hard put to decide whether we have an instance of

intentional error or an instance of incompetence so magnificent that one is reduced to laughter. Consider, for example, (1) an investigator interested in showing the widespread prevalence of psychosis who chooses his sample entirely from the back wards of a mental hospital; (2) an investigator interested in showing the widespread prevalence of blindness who chooses his sample entirely from a list of students enrolled in a school for the rehabilitation of the blind; (3) an investigator interested in showing that the aged are very well off financially who chooses his sample entirely from a list of white, noninstitutionalized persons who are not on relief. The first two examples are fictional, the third, according to the pages of *Science,* unfortunately, is not. (One sociologist participating in that all too real "data"-collecting enterprise was told to avoid apartment dwellers.) A spokesman for a political group which made use of these data noted helpfully that the survey was supported by an organization having a "conservative outlook" (*Science,* 1960). The issue, of course, is not whether an organization having a "liberal outlook" would have made similar errors either of incompetence or of intent but rather that such errors do occur and may have social as well as scientific implications.

THE CONTROL OF INTENTIONAL ERROR

The scientific enterprise generally is characterized by an enormous degree of trust that data have been collected and reported in good faith, and by and large this general trust seems well justified. More than simply justified, the trust seems essential to the continued progress of the various sciences. It is difficult to imagine a field of science in which each worker feared that another might at any time contaminate the common data pool. Perhaps because of this great faith, science has a way of being very harsh with those who break the faith (e.g., Kammerer's suicide) and very unforgiving. A clearly established fraud by a scientist is not, nor can it be, overlooked. There are no second chances. The sanctions are severe not only because the faith is great but also because detection is so difficult. There is virtually no way a fraud can be detected as such in the normal course of events.

The charge of fraud is such a serious one that it is leveled only at the peril of the accuser, and suspicions of fraud are not sufficient bases to discount the data collected by a given laboratory. Sometimes such a suspicion is raised when investigators are unwilling to let others see their data or when the incidence of data-destroying fires exceeds the limits of credibility (Wolins, 1962). It would be a useful convention to have all scientists agree to an open-data-books policy. Only rarely, after all, is the question of fraud raised by him who wants to see another's data, although other types of errors do turn up on such occasions. But if there is to be an

open-books system, the borrower must make it convenient for the lender. A request to "send me all your data on verbal conditioning" made of a scientist who has for ten years been collecting data on that subject rightly winds up being ignored. If data are reasonably requested, the reason for the request given as an accompanying courtesy, they can be duplicated at the borrower's expense and then given to the borrower. Such a data-sharing system not only would serve to allay any doubts about the extent and type of errors in a set of data but would, of course, often reveal to the borrower something very useful to him though it was not useful to the original data collector.

The basic control for intentional errors in science, as for other types of error, is the tradition of replication of research findings. In the sciences generally this has sometimes led to the discovery of intentional errors. Perhaps, though, in the behavioral sciences this must be less true. The reason is that whereas all are agreed on the desirability or even necessity of replication, behavioral scientists have learned that unsuccessful replication is so common that we hardly know what it means when one's data don't confirm another's. Always there are sampling differences, different subjects, and different experimenters. Often there are procedural differences so trivial on the surface that no one would expect them to make a difference, yet, when the results are in, it is to these we turn in part to account for the different results. We require replication but can conclude too little from the failure to achieve confirming data. Still, replication has been used to suggest the occurrence of intentional error, as when Azrin's group (1961) suggested that Verplanck's (1955) data collectors had deceived him. In fact, it cannot be established that they did simply because Azrin's group had been deceived by their data collectors. Science, it is said, is self-correcting, but in the behavioral sciences especially, it corrects only very slowly.

It seems clear that the best control of intentional error is its prevention. In order to prevent these errors, however, we would have to know something about their causes. There seems to be agreement on that point but few clues as to what these causes might be. Sometimes in the history of science the causes have been so idiosyncratic that one despairs of making any general guesses about them, as when a scientist sought instant eminence or to embarrass another, or when an assistant deceived the investigator to please him. Crespi (1945–46) felt that poor morale was a cause of cheating among survey research interviewers. But what is the cause of poor morale? And what of the possibility that better morale might be associated with worsened performance, a possibility implied by the research of Kelley and Ring (1961)? Of course, we need to investigate the problem more systematically, but here the clarion call for "more research" is likely to go unheeded. Research on events so rare is no easy matter.

There is no evidence on the matter, but it seems reasonable to sup-

pose that scientists may be affected by the widespread data fabrication they encountered in laboratory courses when they were still undergraduates. The attitude of acceptance of intentional error under these circumstances might have a carry-over effect into at least some scientists' adult lives. Perhaps it would be useful to discuss with undergraduate students in the various sciences the different types of experimenter effects. They should, but often do not, know about observer effects, interpreter effects, and intentional effects, though they quickly learn of these latter effects. If instructors imposed more negative sanctions on data fabrication at this level of education, perhaps there would be less intentional erring at more advanced levels.

Whereas most instructors of laboratory courses in various disciplines tend to be very conscious of experimental procedures, students tend to show more outcome-consciousness than procedure-consciousness. That is, they are more interested in the data they obtain than in what they did to obtain those data. Perhaps the current system of academic reward for obtaining the "proper" data reinforces this outcome-consciousness, and perhaps it could be changed somewhat. The selection of laboratory experiments might be such that interspersed with the usual, fairly obvious demonstrations there would be some simple procedures that demonstrate phenomena that are not well understood and are not highly reliable. Even for students who "read ahead" in their texts it would be difficult to determine what the "right" outcome should be. Academic emphasis for all the exercises should be on the procedures rather than on the results. What the student needs to learn is, not that learning curves descend, but how to set up a demonstration of learning phenomena, how to observe the events carefully, record them accurately, report them thoroughly, and interpret them sensibly and in some cases even creatively.

A general strategy might be to have all experiments performed before the topics they are designed to illustrate are taken up in class. The spirit, consistent with that endorsed by Bakan (1965), would be "What happens if we do thus-and-so" rather than "Now please demonstrate what has been shown to be true." The procedures would have to be spelled out very explicitly for students, and generally this is already done. Not having been told what to expect and not being graded for getting "good" data, students might be more carefully observant, attending to the phenomena before them without the single set which would restrict their perceptual field to those few events that illustrate a particular point. It is not inconceivable that under such less restrictive conditions, some students would observe phenomena that have not been observed before. That is unlikely, of course, if they record only that the rat turned right six times in ten trials. Observational skills may sharpen, and especially so if the instructor rewards with praise the careful observation and recording of the organism's response. The results of a laboratory demonstration experiment are not new or exciting to the in-

structor, but there is no reason why they cannot be for the student. The day may even come when classic demonstration experiments are not used at all in laboratory courses, and then it need not be dull even for the instructor. That the day may really come soon is suggested by the fact that so many excellent teachers are already requiring that at least one of the scheduled experiments be completely original with the student. That, of course, is more like Science, less like Science-Fair.

If we are seriously interested in shifting students' orientations from outcome-consciousness to procedure-consciousness there are some implications for us, their teachers, as well. One of these has to do with a change in policy regarding the evaluation of research. To evaluate research too much in terms of its results is to illustrate outcome-consciousness, and we do it very often. Doctoral committees too often send the candidate back to the laboratory to run another group of subjects because the experiment as originally designed (and approved by them) yielded negative results. Those universities show wisdom that protect the doctoral candidate from such outcome-consciousness by regarding the candidate's thesis proposal as a kind of contract, binding on both student and faculty.

The same problem occurs in our publication policies. One can always account for an unexpected, undesired, or negative result by referring to the specific procedures employed. That this occurs so often is testament to our outcome-consciousness. What we may need is a system for evaluating research based only on the procedures employed. If the procedures are judged appropriate, sensible, and sufficiently rigorous to permit conclusions from the results, the research cannot then be judged inconclusive on the basis of the results and rejected by the referees or editors. Whether the procedures were adequate would be judged independently of the outcome. To accomplish this might require that procedures only be submitted initially for editorial review or that only the result-less section be sent to a referee or, at least, that an evaluation of the procedures be set down before the referee or editor reads the results. This change in policy would serve to decrease the outcome-consciousness of editorial decisions, but it might lead to an increased demand for journal space. This practical problem could be met in part by an increased use of "brief reports" which summarize the research in the journal but promise the availability of full reports to interested workers. Journals such as the *Journal of Consulting Psychology* and *Science* are already making extensive use of briefer reports. If journal policies became less outcome-conscious, particularly in the matter of negative results, psychological researchers might not unwittingly be taught by these policies that negative results are useless and might as well be suppressed. In Part III negative results will be discussed further. Here, as long as the discussion has focused on editorial policies which are so crucial to the development of our scientific life styles and thinking modes, it should be mentioned that the practice of reading manuscripts for critical

review would be greatly improved if the authors' name and affiliation were routinely omitted before evaluation.[1] Author data, like experimental results, detract from the independent assessment of procedures.

[1] Both Gardner Lindzey and Kenneth MacCorquodale have advocated this procedure. The usual objection is that to know a man's name and affiliation provides very useful information about the quality of his work. Such information certainly seems relevant to the process of predicting what a man will do, and that is the task of the referee of a research proposal submitted to a research funding agency. When the work is not being proposed but rather reported as an accomplished fact, it seems difficult to justify the assessment of its merit by the reputation of its author.