chapter 7

Introspection II: Access to the Causes of Behavior

Why did you major in psychology? Why do you think Dr. Smith is a good teacher? Why do you like Bill better than Ben? Why did you go to movie X, instead of movie Y? When you are asked to explain why you made a particular judgment, decision, or behavior, you can usually give some plausible explanation. On the surface, it would seem that people use introspection—looking within their own minds—to determine why they like what they like, and do what they do. However, some psychologists have come to doubt that people really have introspective access to the causes of their thoughts and actions.

Most of the empirical evidence for nonconscious information processing has come from research on *lower-order* mental processes, including perception, memory retrieval, and control of habitual actions. (I reviewed some of the evidence in the last chapter.) In a famous article titled "Telling More than We Can Know," Richard Nisbett and Timothy Wilson (1977) supported the radical proposal that *higher-order* mental processes—processes involved in judgments and decisions leading to voluntary actions—are also nonconscious. The belief in conscious awareness of higher-order thought processes is an illusion, in their view.

The research discussed by Nisbett and Wilson is related to an important topic in social psychology, the *attribution problem*, which concerns people's attempts to explain the causes of human behavior. Attribution research is concerned with factors that influence people's tendencies to attribute other people's behavior (such as unusually friendly or hostile behavior)

to internal causes (such as personality traits) versus external (situational) causes. Nisbett and Wilson were interested in the *self-attribution* problem, concerning people's attempts to explain their own behavior, attitudes, and emotional reactions. They concluded that, when people attempt to explain their own behavior, they do not have introspective access to the causes or thought processes that produce their behavior. Rather, they said, when people try to give introspective reports on the causes of their behavior, what they are really doing is making reasonable *inferences* about what the causes must have been.

Nisbett and Wilson's conclusion is contrary to the common-sense belief that we can look into our own minds to find the causes or reasons for our actions. Of course, Nisbett and Wilson were not the first to make such a proposal. Around 1900 Sigmund Freud, the father of psychoanalysis, argued that people's decisions and actions are influenced by desires and memories that are unconscious—in the sense that we cannot retrieve them to consciousness even if we try. In the psychoanalytic view, unconscious desires and memories are unconscious because they are repressed; that is, they are actively blocked from consciousness because to make them conscious would cause great anxiety. Nisbett and Wilson's proposal was broader than Freud's, since they suggested that a lack of introspective access to the causes of our behavior is commonplace, and is not limited to cases of repression of knowledge that might produce personal anxiety.

Nisbett and Wilson's arguments deserve to be considered carefully because, if their conclusions are correct, they are profoundly important, for two reasons. First, their conclusions imply that introspection is worthless as a psychological method for studying higher-order thought processes. In particular, the validity of interpretive introspection—attempts to look into our own minds to find the causes and reasons for our thoughts, feelings, and actions—is in question. Second, their conclusions have implications for our views of human nature and the limits of self control and moral responsibility. If we cannot truly know the causes of our decisions and actions as they happen, then how can we be expected to control them and take moral responsibility for them?

I will describe the main points of Nisbett and Wilson's (1977) arguments against introspective access to higher-level mental processes and the causes of behavior, and I will describe some of the supporting research. Then I will present some of the counter-arguments offered by other psychologists, and some newer research that suggests a more moderate conclusion.

NISBETT AND WILSON'S ANTI-INTROSPECTIONIST THEORY

Behavior controlled by higher-order thought processes may be distinguished from mere reflexive or habitual responses to stimuli. Control of behavior by thought processes implies the ability to respond flexibly and adaptively to changing situations. To do this, a variety of pertinent stimulus information from the immediate environment must be combined with information from memory, through a series of flexible judgment and decision processes. The external stimuli that enter into thought processes are among the causal de-

terminants of behavior. Nisbett and Wilson (1977) assumed that if people have introspective access to the thought processes that control their behavior, then they should be able to report the external stimuli that entered into those thought processes, and also to report how the stimuli affected their behavior.

Research and Conclusions

Nisbett and Wilson described a series of experiments designed to determine whether people can report the causes of their behavior. The general plan involved these steps: (1) Subjects were tested under two or more different stimulus conditions that were known to produce measurable differences in behavior, in a controlled situation where they had to make inference-based responses (that is, nonhabitual responses based on higher-order thought processes). (2) Afterward, the subjects were asked to report why they responded the way they did. (3) The reports were analyzed to determine whether the subjects reported the stimuli that the experimenters manipulated and that the experimenters knew to be true causal influences on the subjects' responses.

Nisbett and Wilson argued that the research supported these three major conclusions: (1) People do not have introspective access to the causal relationship between stimuli and their responses. That is, they cannot accurately report from introspection *which* stimuli affected their responses, and/or they cannot report *how* the stimuli affected their responses. (2) Reports of effects of stimuli on responses are based not on introspection, but on *a priori theories* (prior beliefs) about the causal connections between the stimuli and responses. (3) When people's reports on stimulus-response relationships are correct, it is because their a priori theories happen to be correct, not because of correct introspection. I will explain the rationale for these conclusions in more detail.

No introspective access to stimulus-response relationships. People often cannot report accurately on the effects of stimuli that influenced their responses. Sometimes they cannot report that the critical stimuli occurred. Sometimes they cannot report that the critical *response* occurred. Sometimes they are aware of the stimuli and the responses but they cannot report how the stimuli affected the responses. Sometimes people report incorrectly that certain stimuli affected their responses. Sometimes when people are informed which stimuli really affected their responses, they deny that this was the case.

For example, consider the "bystander apathy effect." Latane and Darley (1970) showed in several experiments that subjects were less likely to help a stranger in distress—such as someone who was apparently sick or injured—when there were several other people around than when they (the subject) were the only person available to help. When the subjects were asked why they had failed to lend assistance they said nothing about being influenced by the fact that other people were present. Furthermore, when the experimenter suggested that their behavior had been influenced by the presence of other people, the subjects denied that this was the case. Thus, the

subjects were unaware of the influence of a critical factor (presence of others) that had, in fact, affected their behavior according to the experimenter's objective analysis.

In a study of the relation between symptom attribution and pain tolerance, Nisbett and Schacter (1966) asked subjects to take a series of electric shocks that would increase steadily in intensity until the subjects said that they had reached the limit of their pain tolerance. Prior to the shocks series, half of the subjects were given a placebo pill (a substance with no physiological effects), and they were told that the pill would produce symptoms such as increased heartbeat rate, breathing irregularities, hand tremor, and butterflies in the stomach. These are the symptoms of anxiety that are normally produced by electric shock. As it turned out, the placebo group tolerated four times as much shock intensity as the control group. The experimenter's interpretation was that the placebo group attributed their anxiety symptoms to the pill, whereas the control group attributed their symptoms to the shock. Thus, when the anxiety symptoms became unpleasant the control group wanted to quit taking shocks, whereas the placebo group was willing to continue the shocks because they thought the anxiety symptoms were caused mainly be the pill. But when the placebo subjects were asked why they were willing to endure so much electric shock, they did not report that the pill had anything to do with it. Even when the experimenter explained that the pill was a placebo and that they had probably attributed their anxiety symptoms to the pill, the subjects denied that the pill had affected their responses to the shock.

A priori theories about the causes of behavior. Nisbett and Wilson argued that when people are asked to explain the causes of their own behavior, they do not do so by directly introspecting on the stimuli and thought processes that produced their responses. Instead, they make *post hoc* (after-the-fact) inferences about the causes of their responses. An inference is a conclusion based on some combination of reasoning, observations, and prior knowledge or beliefs, rather than on observations alone. Nisbett and Wilson emphasized the role of a priori theories, that is, the subjects' beliefs about causal connections between stimuli and responses—beliefs that existed prior to, and independently of, a particular attempt to introspect on the thought processes that led to a particular response.

Nisbett and Wilson suggested several possible origins for the a priori theories: (1) There are socially learned rules that govern much of our behavior. For example, one rule is "Stop at red lights." Thus, in explaining your behavior of stopping at an intersection, you could refer to the rule, without actually introspecting on the effect of the stimulus (red light) on your response. (2) The culture or subculture provides theories about causes of behavior and feelings—what some writers call "folk psychology" theories. For example, one folk-psychology theory is "Personal failures make people feel depressed." In explaining your feeling of depression, you might make use of the theory and say "I feel depressed today because I got a 'D' on a history exam." Your explanation would not be based on introspection, and you might entirely miss the fact that your depression was caused by physical factors. (3) A causal theory may be based on personal observation of *covariation*

between types of stimuli and types of responses. For example, you might have noticed in the past that you are usually grouchy after you play golf if you fail to break 100. If you are asked why you feel grouchy today, you might attribute it to your failure to break 100. In fact, other factors—either physical or psychological—might well be more important causes of your grouchiness. (Perhaps excessive fatigue caused both your poor play and your grouchiness.) (4) People may generate causal hypotheses based on judgments of connotative similarity between the stimuli and responses. That is, the stimuli and responses seem to go together, such that one seems to imply the other. For example, "I feel happy today. It must be because I was with Jane, who is always in a happy mood."

In general, the methods that people use to make inferences about causality are analogous to the *availability* and *representativeness* heuristics in category judgments (Kahneman & Tversky 1973). Availability means simply that people base their judgments or inferences more on highly salient or readily available information than on less obvious information that might really be more relevant to the problem at hand. For example, you might attribute your headache to loud rock music coming from the next room, whereas you might overlook less obvious factors that are really the cause of your headache (such as tension from worrying about a personal problem; or caffeine withdrawal, if you are a coffee or cola addict and you haven't had a "fix" recently).

In judgment research, according to the representativeness heuristic, people estimate the probability that a particular object is in category X by comparing the object with known objects that they believe to be typical or representative of category X. For example, if you believe that librarians are typically female, quiet, orderly, and love books, and you observe that Ruth is quiet, orderly, and loves books, then you might conclude that it is likely that Ruth is a librarian. You would be ignoring the more important fact that most quiet, orderly, book-loving females are *not* librarians.

Similarly, according to Nisbett and Wilson, we infer that a particular stimulus caused our behavior, not through introspection of our actual mental processes, but by judging whether the stimulus is representative of the category of stimuli that usually cause that sort of behavior. For example, anger is an emotional response to certain stimuli. According to (a priori) folk-psychology theory, frustrations and insults are representative of the type of stimuli that produce anger. If you feel irritable or angry, you might search your environment and memory for a recent case of frustration or insult. Failing to find an obvious case, you might interpret someone's ambiguous remark as an insult. (Professor to student: "You seem to be having a lot of trouble with this course." Student's thought: "That turkey! He thinks I'm stupid.") You might entirely overlook the influence of other factors on your feelings of irritability or anger (such as fatigue, sickness, hangover, food allergy, sexual frustration, or physiological cycles).

Implications. Nisbett and Wilson took the strong position that people have little or no introspective access to mental processes, and that their verbal reports on the causes of their behavior are based entirely on a priori the-

ories. This position carries two important implications: (1) People may sometimes report accurately on the causes of their behavior. But since their reports are based on a priori theories, their reports will be correct only if their a priori theories are correct. (2) People's reports on the causes of their own behavior will be no more accurate than the reports of observers who use the same a priori theories.

The second point needs elaboration because it provides the rationale behind several experiments—some of them to be described here—on introspective access to the causes of behavior. The basic idea is that if your explanations of the causes of your behavior are based not on introspection but on a priori theories, then another person—an "observer"—who obviously does not have introspective access to your mental processes, should be able to provide the same explanations of your behavior. This assumes that the observer is the same gender as you, about the same age, and a member of the same culture as you, so that they would have the opportunity to learn the same a priori theories about the causes of human behavior, and also that they know the important details of the situation in which you made the responses in question. Since—according to Nisbett and Wilson—your explanation of your behavior did not benefit from introspective access to your own mental processes, it should not make any difference that the observer has no direct knowledge of your mental processes. The important point is that you would share the same a priori theories, so this person would explain your behavior in the same way that you would, and equally accurately or inaccurately. It should be noted that Nisbett and Wilson did not deny that people have introspective access to some of their conscious experiences or contents. But they denied that people use this introspective knowledge to explain their own behavior.

Though Nisbett and Wilson's argument may seem to go against your intuition and your everyday experience, there is some compelling experimental evidence to show that they are at least partly right, if not entirely so.

Criterion for introspective access. From the argument described above, Nisbett and Wilson (1977) derived a criterion for introspective access to the causes of behavior. They said that introspective access would be demonstrated if it were shown that real subjects who respond in a particular experimental situation can subsequently make verbal reports on the causes of their responses that are more accurate than the reports obtained from observers. The observers' task is to report the causes of the real subjects' responses, in cases where the observers are provided with a general description of the experimental situation, including the stimuli and responses in question, but where the observers do not make their own responses to the experimental situation.

Nisbett and Wilson (1977; Wilson & Nisbett 1978) described several experiments intended to test for introspective access to the causes of behavior, according to the criterion described above. None of the studies found clear evidence for introspective access. For example, Nisbett and Bellows (1977) compared the reports of observers with those of real subjects in a study of stimulus factors that influence judgments of people. (Henceforth, I will fol-

low the jargon of this research literature and call the real subjects "actors," in the sense of one who takes action or responds to a situation.) Different actors read different descriptions of a job applicant that varied on several stimulus factors (such as the applicant's attractiveness, academic credentials, and so forth). Then they judged the applicant on several person dimensions (likability, sympathy, flexibility, intelligence). Finally, the actors tried to report the causes of their responses by estimating how much each of the stimulus factors had influenced each of their person judgments. For comparison, observer subjects were given only a brief description of each of the stimulus factors, and were asked to estimate how much each of them would affect the actors' judgments of job applicants on the person dimensions. Two aspects of the results were important: (1) The actors and observers agreed closely in their estimates of how much each stimulus factor would affect each of the person judgments. (2) These estimates were not accurate, according to the researcher's data analysis that showed how the person judgments had actually varied in relation to the stimulus factors. Thus, Nisbett and Wilson concluded that the actors and observers had used the same a priori theories to explain the actors' responses. Furthermore, both groups were wrong, with one exception. Actors and observers were both correct in saying that the description of the job applicant's academic credentials influenced the actors' judgments of the applicant's intelligence. In this case, the a priori theory happened to be correct.

In another example, in research on the bystander apathy effect (described earlier), observers have sometimes been asked to explain why actors failed to lend assistance to a person in distress. Like the actors, observers almost always failed to mention the fact that other people were present, even though the research has objectively shown the presence or absence of other people is a critical factor influencing the likelihood of helping behavior. Thus, from these and other studies, Nisbett and Wilson concluded that people do not have introspective access to the causes of their behavior. Rather, they base their reports on a priori theories, that may or may not be correct.

Reasons for belief in introspective access. If we do not truly have introspective access to the causes of our own behavior, why do we continue to believe that we usually are aware of the causes of our behavior? Nisbett and Wilson suggested that people have an illusion of introspective access because they confuse awareness of mental *contents* with awareness of mental *processes*. Mental processes make various computations and judgments on pertinent information to produce particular decisions, which in turn guide behavior. Nisbett and Wilson argued that we do not have introspective access to these mental processes. Rather, we have introspective access to the *results* of those processes, which are mental contents, such as decision outcomes (not the decision process itself). We also may have access to intermediate results at various stages of the mental processes that lead to decisions. (This is analogous to intermediate results that you get during computation of a long mathematical formula.) Thus, we may think we are introspecting mental processes when we are really introspecting mental contents.

Critique of Nisbett and Wilson's Theory

Nisbett and Wilson's anti-introspectionist theory may seem contrary to our personal intuitions and everyday experiences, but that in itself is not sufficient proof against it. Arguments based on logic and evidence from controlled research are more pertinent to establishing the validity of a scientific theory.

Problems of logic. Several critics (Smith & Miller 1978; P. White 1980) have argued that Nisbett and Wilson did not clearly distinguish mental processes from mental contents. They did not provide definitions that were clear enough to enable one reliably to judge whether subjects were describing mental processes or mental contents. Nisbett and Wilson themselves were prepared to accept actors' reports of pertinent stimuli as evidence for introspective access to mental processes (if their reports were more accurate than those of the observers), though it could be argued that the remembered stimuli are mental contents and not mental processes.

Also, Smith and Miller (1978) suggested that any mental process can be described at several different levels, ranging from high-level strategies or rules for problem solving and social behavior, through intermediate-level computation and judgment stages, to the lowest level of activities in various brain circuits that actually perform the computations and control the behavior. People may be able to make introspective reports on the higher-level mental processes, but not on the intermediate or lower-level processes.

In more recent work it has been acknowledged that it is difficult to make a clear distinction between mental processes and mental contents, and perhaps this distinction should be deemphasized (Wilson and Stone 1985). What is more important is the question of whether people can report which stimuli influenced their behavior, and how.

On this point, Kenneth Bowers (1984) argued that while people may sometimes have introspective access to the stimuli that cause their behavior, it is logically impossible to have introspective access to the causal connection between the stimuli and their behavioral effects. Rather, understanding of the causal connections between events is always and necessarily a theoretical inference. Psychologists devise formal theories to explain people's behavior, based on controlled research, and ordinary people devise informal theories based on informal observations. But in no case can the causal connection between events be directly observed; it has to be inferred. For example, you might observe that almost every time that you feel tired you also are irritable, and so you infer that tiredness causes irritability. But you do not directly observe the process that links tiredness with irritability. (And in fact, tiredness might not cause irritability; they might both be caused by another factor that happens sometimes to be correlated with both of them.)

Thus, regarding the illusion of introspective access to mental processes, what is more important than the confusion between mental processes and mental contents is people's failure to distinguish between their introspective observations and their inferences. People may introspectively (or retrospectively) observe certain stimuli and responses, then infer a causal

connection between them, and subsequently report (wrongly) that they introspectively observed that a particular stimulus caused a particular response. What they really did was to infer that the stimulus caused the response.

The implication is that we can still ask whether people have introspective access to the stimuli that influence their behavior, but we should understand that their reports on *how* the stimuli affects their behavior are based on inferences, not introspective observations. People's inferences about causal connections may be based on a priori theories that are shared by observers from the same culture, as Nisbett and Wilson argued. Or alternatively, some of their inferences may be based on privileged knowledge that is not available to observers.

Finally, you should note that Nisbett and Wilson's criterion for introspective access—more accurate reports on causal stimuli by actors than observers—does not work in reverse. That is, failure of actors' reports to be more accurate than those of observers does not in itself prove lack of introspective access. In some cases, actors might have introspective access to pertinent stimuli, but make inaccurate reports because they made incorrect inferences about the relationships between the stimuli and their responses. In other cases, actors and observers might both be very accurate, with the actors having introspective access to information similar to that assumed by observers on the basis of a priori theories. In such cases, actors are actually using introspective information, but it is hard to prove it by objective criteria.

Problems of research method. Several psychologists have criticized some of the research that Nisbett and Wilson used to support their arguments. One problem was that actors' reports on causal influences on their behavior were all made retrospectively, several minutes or hours after the critical behavior occurred. Thus, they may have forgotten pertinent information that they would have known if they had been questioned sooner.

Another problem frequently mentioned (Smith & Miller 1978) was that most of the experiments involved *between-subjects designs*, in which different subjects (or groups of subjects) received different stimuli or experimental treatments, and made different responses. In order to understand how a stimulus influences one's behavior, it is important to have an opportunity to experience *covariation* between the stimulus and the response. That is, you need to have a chance to notice that your response is different when the stimulus is different. In the between-subjects experiments, it might be unrealistic to expect actors to know that a stimulus influenced their behavior, when they experienced only one stimulus condition and made only one response. For example, in the bystander apathy effect, if you experienced only one condition—an apparently injured person, with a crowd of people standing around—and you failed to help the injured person, it might not occur to you that the crowd had an important influence on your decision not to help, and that you would have helped if you had been the only person on the scene.

In reply to the criticism of between-subjects experimental designs, Nisbett and Ross (1980) argued that such experiments are most realistic in comparison to people's everyday life judgments. Often we try to judge the effect of a stimulus on our behavior in situations where we have only experi-

enced that stimulus once, and we have not had a chance to observe covariation between the stimulus and our response. Other researchers have argued, however, that if we want to find out whether people *can* have introspective access to the causes of their behavior, then we should test people under conditions that maximize the opportunity for introspective access and accurate reporting. Thus, it has been recommended that experiments on introspective access should involve *within-subjects* designs, in which all subjects experience each of the different stimulus conditions on several occasions, so they have the opportunity to notice how their responses covary with the stimulus conditions. Another advantage of within-subjects designs is that it enables experimenters to evaluate the accuracy of individual subject's reports of stimulus-response relationships, rather than trying to draw conclusions from average group accuracy computed in between-subjects designs. In the next section I will describe some newer research, using within-subjects designs, and the conclusions that have followed from that research.

EVIDENCE FOR INTROSPECTIVE ACCESS

Since the original review by Nisbett and Wilson (1977), several studies have used the recommended within-subjects experimental designs to test for introspective access to the causes of behavior. In each of these studies, real subjects (actors) made judgments of each of several different stimuli (such as different people), in situations where the stimuli varied on several factors or attributes. Most of the studies found evidence for introspective access—or at least some sort of privileged information access—by showing that the actors' reports of the effects of different stimulus factors on their judgments were more accurate than the estimates made by observers. I will briefly describe some of these studies, going into more detail on the first one so you can better understand the rationale and method and results.

Experiments Using Within-Subjects Designs

Kraut and Lewis (1982) had some ninety-six subjects (actors) watch a film showing a series of short interviews of fifty international airline passengers conducted by U.S. Customs inspectors. The inspectors asked the passengers questions about things such as their age, home residence, length and purpose of trip, occupation, and purchases. The passengers varied widely in age, social class, and behavior. The subjects judged each of the passengers (target stimuli) for three person characteristics: intelligence, friendliness, and deceptiveness. The person judgments were made on six-point rating scales, ranging from "1, below-average" to "6, above average," compared to the other passengers.

After the subjects had made person judgments for all fifty passengers, the subjects were asked to estimate the impact of each of eighteen stimulus factors on their judgments. The stimulus factors—characteristics of the passengers—included things such as age, sex, occupation prestige, attractiveness, frequency of smiles, eye contact, formality of dress, formality of speech, response latency, and volunteering of information. The subjective impact

estimates were made on thirteen-point scales, ranging from negative influence through no influence to positive influence. (For example, a subject might estimate that passengers who smiled more frequently had been judged as more friendly or that smiles had had no influence on judgments of intelligence.) Also, observer subjects—who did not see the film and did not make person judgments—used the same thirteen-point scales to estimate how they thought each of the eighteen stimulus factors would influence each of their person judgments (friendliness, intelligence, deceptiveness) if they were actually to make such judgments. The ratings of stimulus-factor impacts by the actor subjects are termed *actor estimates*, whereas those made by the observers are termed *observer estimates*. Finally, for each actor subject, the *actual impact* of each of the eighteen stimulus factors on each person judgment was determined by computing the correlation of each stimulus-factor level with each judgment rating. (For example, the passengers' actual frequencies of eye contact with the inspector were correlated with the subjects' ratings of their deceptiveness, across all fifty passengers.)

Four results of the data analysis are important in regard to the question of introspective access to the causes of behavior: (1) The accuracy of actors' estimates of the influence of the various stimulus factors on their person judgments was determined by correlating the actor estimates with the actual impacts. Averaged across all subjects, judgments, and stimulus factors, the *actor-actual correlation* was +0.42, which indicates a moderate degree of accuracy.[1] (2) The accuracy of observers' estimates of the influence of the stimulus factors was determined by randomly pairing each observer with an actor subject, and computing the correlation of the observers' estimate with the paired actors' actual impacts for each stimulus factor and each judgment. Overall, the average *observer-actual correlation* was +0.35. By the rationale suggested by Nisbett and Wilson (1977), the fact that the actor-actual correlation was higher than the observer-actual correlation (0.42 versus 0.35) suggests that the actors had some additional information, not known to the observers, that enabled the actors to be more accurate. This additional information might come from introspective access to the effect of stimulus factors during the actual judgment process. (3) To find out how much observer estimates were similar to actors' estimates, these two variables were correlated with each other. The *actor-observer correlation* of +0.48 indicates that actors' and observers' estimates were moderately similar, perhaps due to both of them using the same a priori theories about the relationship between stimulus factors (such as smiling, eye contact) on person judgments (such as friendliness). However, the correlation is far from perfect, which suggests that to a moderately large degree the actors and observers used different information in estimating the influence of the various stimulus factors on person judgments. (4) The most direct test of introspective access involved measuring the degree to which the actors' stimulus-factor impact estimates were accurate due to information not known to observers. This was done by computing a partial correlation between actors' estimates and actual impacts, while statistically controlling for observers' estimates. (In other words, the observers' accuracy was subtracted from the actors' accuracy, to determine how much residual actor accuracy remained.) The *actor-actual partial correlation* was +0.31, which indicates that actors' estimates of the impact of various stimu-

lus factors on their person judgments was based on a significant amount of information not known to the observers. Presumably, this additional information involved introspective access to their judgment process, though as we will see, other sorts of privileged information may also have been involved.

Similar results have been found in other studies that employed different judgment tasks but similar within-subject experimental designs. Wright and Rip (1981) had high-school juniors read descriptions of thirty-two colleges that varied on five stimulus factors (tuition cost, size, distance from home, living patterns, [percent of students living on campus], and median SAT scores). Then the subjects judged the colleges in terms of how much they wanted to apply to them for admission. Finally, they estimated the impact of each stimulus factor on their judgments. The correlation of actors' estimates with actual effects of the stimulus factors was higher (+0.38) than the observer-actual correlation (+0.32).

Gavanski and Hoffman (1987) had subjects read brief profiles of sixty-four male university students, which varied on six stimulus factors (such as whether they smoked, how hard they worked, whether they often swear when annoyed). Subjects judged how much they thought they would like each of the students (on a thirteen-point scale), then estimated the impacts of the different stimulus factors on their judgments. Actor-actual correlations were much higher (+0.81) than observer-actual correlations (+0.24), for observers who had not seen the student profiles or the actors' judgments.

On the other hand, a study by Wilson, Laser, and Stone (1982) did not find evidence for introspective access. Every day for five weeks, subjects filled out questionnaires in which they rated their mood (on a seven-point scale from "very bad" to "very good") and several predictor factors, such as their health, personal relationships, workload, amount of exercise, and amount of sleep the previous night. Finally, they estimated what impact each of the predictor factors had had on their mood over the five weeks. The correlation of actors' estimates with actual impacts was slightly lower (+0.42) than the observer-actual correlation (+0.45).

Overall, then, most of the published studies using within-subjects designs—where each subject judges numerous stimuli that vary on several stimulus factors—have found actors' estimates of the impacts of stimulus factors on their judgments to be somewhat more accurate than observers' estimates. To the extent that actors' and observers' judgments are similar, the results suggest that they are using the same a priori theories to explain their judgments, as Nisbett and Wilson (1977) claimed. However, to the extent that actors' estimates are more accurate than observers, the results suggest that actors are using privileged information—perhaps from introspection—that is not available to the observers to explain the actors' judgments.

Types of Information People Use to Explain Their Own Behavior

What type of information might people use to explain the causes of their own behavior—such as their judgments of people? Wilson and Stone (1985) distinguished between several types of information: (1) *Shared theories*

are those theories that are known by both actors and observers in the same culture. Shared theories include a priori theories—folk-psychology theories—that are learned by virtually all members of the culture, and also theories that can be derived by all members from their similar experiences. Actors and observers would be expected to have the same shared theories. (2) *Privileged information* is known only to actors—the people whose behavior is in question—and not by observers. There are three types of privileged information. (a) *Introspective data* on the "workings of one's own mind," such as remembered instances of particular stimuli and responses, and the sequence of decisions and affective reactions (feelings) that lead one to make particular responses to particular stimuli. (b) *Covariation data*, the accumulated knowledge of one's different responses to different stimuli, that can be used to infer a cause-and-effect relationship. (For example, you might notice that you rather consistently judge giggly people to be nervous, and so you infer that giggling causes you to judge people as nervous. Such an after-the-fact inference from covariation data is not the same thing as introspecting thoughts on specific occasions where someone's giggling caused you to conclude that they were nervous.) (c) *Idiosyncratic theories*, unique to the individual (such as "Coffee makes me sleepy."), that may be wrong due to faulty data analysis or faulty inference.

Wilson and Stone (1985) acknowledged that recent research shows that actors often use privileged information—information not available to observers—in trying to explain their own behavior. This conclusion departs significantly from the original Nisbett and Wilson (1977) argument that causal explanations are based only on shared a priori theories. However, Wilson and Stone concluded that *the use of privileged information does not necessarily make actors more accurate than observers* in explaining their (the actors') behavior. Sometimes actors analyze covariation data incorrectly, or use idiosyncratic theories that are wrong. Furthermore, in most cases in which actors' reports are more accurate than observers' reports, the differences are not large, and observers' and actors' reports are moderately highly correlated. The implication is that shared theories still account for a significant portion of actors' reports, though actors also use privileged information.

Covariation assessment. In cases in which actors use privileged information and their reports are more accurate than those of observers, it is difficult to determine whether the privileged information used by actors is based on actual introspection of ongoing decision processes or on after-the-fact inferences based on accumulated covariation data. The evidence suggests that covariation data can be used under some circumstances, though it does not account for all privileged information reports. Research on human inference processes has indicated that, in a variety of situations, people are usually poor at making accurate inferences based on covariation data (Nisbett & Ross 1980). The larger the number of stimulus factors that can vary, the harder it is to determine their relationship to people's responses. Also, the greater the time interval between different stimulus-response instances, the harder it is to accumulate useful covariation data, since some instances will be forgotten or remembered incorrectly.

Wilson and Stone (1985) ranked several self-attribution studies in

terms of how hard it would be for actors to assess stimulus-response covariation, and they concluded that the easier it was for actors to assess covariation, the greater the advantage of actors over observers in accurately explaining the actors' judgments (regarding persons, colleges, and so forth). This conclusion has been substantiated in a more recent study. Gavanski and Hoffman's (1987) procedure—in which subjects judged how much they would like students, based on brief profiles—made it relatively easy for subjects to detect covariation. They judged a large number of profiles in a relatively short time period, and the profiles varied on only six "yes-no" stimulus dimensions (such as "Does the student smoke?"). Subjects (actors) were unusually accurate in judging the impact of stimulus factors (actor-actual correlation +0.81), whereas observers with no covariation information were very inaccurate (observer-actual correlation +0.24). Furthermore, an observer group that was given covariation information—they read the student profiles and examined an actors' judgments of the profiles—were almost as accurate as the actors themselves in estimating the impact of the different stimulus factors on the actors' judgments (covariation observer-actual correlation +0.71). The superior accuracy of observers with covariation data over observers without covariation data shows that people can use covariation data in situations of this type, and suggests that the actors also used covariation data. However, other data analyses indicated that the accuracy of actors' reports could not be attributed entirely to covariation information. The actors used additional, privileged information, not known to the covariation observers, and in fact they used covariation information somewhat less than covariation observers did. Apparently, actors' reports were based partly on their introspections of individual judgments, in addition to their after-the-fact covariation analysis of the relationships between stimulus factors and their responses.

Wilson and Stone argued that although people might be able successfully to assess stimulus-response covariation in relatively simple, controlled laboratory situations, it might be difficult or impossible to do so in the real world. In real-life situations, the number of varying stimulus factors would often be so great, and/or the time intervals between relevant stimulus-response instances would be so great, that it would be hard to notice any consistent stimulus-response covariation. (For example, you might not notice that a particular food causes you to feel lethargic, because that particular food is consumed along with a variety of other foods that do not affect your mood.) Thus, in real-life situations, actors might have little or no advantage over observers that could be attributed to privileged access to covariation data by actors. On the other hand, as Gavanski and Hoffman's (1987) analysis suggests, the superior accuracy of actors' causal attributions is not due only to the use of covariation data. Actors have additional private information—presumably from introspection of stimuli, responses, feelings, and ongoing thought processes. Thus, we would expect that actors' explanations of their own behaviors should be more accurate than those of observers, even in real-life, real-world situations in which covariation assessment is difficult or impossible.

In evaluating these and other conclusions based on comparisons of actors' and observers' explanations of actors' responses, recall that it was as-

sumed that the observers were of the same age, gender, and culture, and had educational backgrounds similar to the actors. Under these conditions, observers' explanations have been either less accurate, or at best, equally as accurate as actors' explanations. But observers may be more accurate than actors in explaining the actors' behavior, in cases where the observers are experienced psychologists armed with accumulated knowledge based on clinical experience, research results, and formal psychological theories that are more valid than the folk-psychology theories used by ordinary observers.

Finally, it should be noted that the process of assessing stimulus-response covariations for one's own behavior is, in a sense, an introspective process. Regardless of whether covariation analysis is done as a continuing process throughout the series of behavioral events, or done afterward, it is a private mental process involving the attempt to interpret one's own experience and behavior. In that sense it is interpretive introspection, which I linked earlier with reflective consciousness.

CONCLUSIONS

Wilson and Nisbett (1977) argued that people do not have direct introspective access to the causes of their behavior, and that their attempts to explain their own behavior are based on a priori theories about why people do what they do and feel what they feel. More recent research leads to a less extreme conclusion. The research indicates that in some cases actors' explanations of their own behavior are more accurate than observers' explanations, due to privileged information access by actors. This privileged information access includes some degree of introspective access to pertinent causal stimuli and thought processes by actors. Also, actors usually have better access to stimulus-response covariation data about their own behavior. Thus, people's attempts to explain their own responses are not based entirely on their a priori theories. A priori theories are, nonetheless, an important component of people's causal explanations, as is shown by the similarity of explanations by actors and observers. In some cases, privileged information access by actors does not improve the accuracy of their explanations, if they misinterpret the information and make incorrect inferences. Introspective access to the causes of behavior is limited.[2]

Two mental systems. To the extent that people are unable to introspectively report the causes of their behavior, the results of these experiments are consistent with theories that distinguish between a global-database conscious awareness system (CAS; Schacter 1989) or conscious interpreter system (Gazzaniga 1985) versus nonconscious modules and/or an executive system that actually controls behavior. In a similar vein, Wilson (1985) suggested that there are two mental systems. One, the *behavior production system*, is responsible for producing behavior, especially unregulated (nonvolitional) behavior. It operates largely at a nonconscious level, so its processes are not directly available to introspection. The other, the *verbal explanatory system*, attempts to assess and interpret our feelings and behavior, using whatever

information is available to it. This system operates at a conscious level, so its conclusions are available for introspection and verbal reporting.

Although the behavior production system's processes are not directly introspectible, some of its inputs (stimuli) and outputs (such as intermediate results in thought sequences, and consequent affective reactions) have conscious consequences and are introspectible. Thus, people would be expected to have some degree of introspective access to stimuli and thoughts related to the causes of their behavior.

Conditions that Promote Introspective Access

Under what conditions would we expect people to be more accurate in making introspective verbal reports on the causes of their own behavior and emotional reactions? Introspective (or retrospective) access to pertinent information should be better under conditions that promote attention to pertinent stimuli and thoughts as they occur, and that increase the likelihood that such information will be stored in memory. The distinction between controlled (effortful) processes and automatic processes is useful here. We would expect better introspective access for novel tasks that require flexible, controlled processing that has a large conscious component. Introspective access would be less likely for habitual tasks that are performed largely through automatic, nonconscious processes. Smith and Miller (1978) made a similar point:

> A dimension that can distinguish situations in which correct self-reports will be possible from those in which they will not . . . is the degree to which the subject is asked to report on tasks that are novel and of interest. Tasks that are novel and engaging for subjects, such as choosing a college or solving challenging problems, often seem to evoke accurate introspective awareness of process. Tasks that are, on the other hand, overlearned, routine, or of minimal interest may well be "run off" in such a way that subjects cannot report on intervening processes (pp. 360–61).

Given that the task is a nonautomatic one that engages one's attention, awareness of causal factors should be increased by conditions that make covariation between responses and pertinent stimuli or thoughts easier to detect. Covariation detection should be easier when there are relatively few pertinent stimulus dimensions and they are highly salient (conspicuous, novel, interesting), and when one has an opportunity to experience several variations on the stimuli, with consequent variations in responses, over a relatively short time interval.

We would expect people to be able to report the causes of their behavior more accurately if they introspect on their thought processes while the task is in progress, rather than waiting until the task is finished and relying entirely on retrospection. In the self-attribution studies that I described, subjects did not know that they would be asked to report on the influence of stimulus factors until after the judgment tasks were finished. Thus, their reports were based on retrospection. Some studies have examined the effects of concurrent introspection by comparing the accuracy of subsequent causal explanations for subjects who are given advance instructions to attend to

pertinent stimuli and thoughts while they make their judgments (such as person judgments) versus subjects without prior attention instructions.

Surprisingly, advance attention instructions do not necessarily increase the accuracy of causal explanations (such as stimulus-factor impact estimates). In some cases, advance attention instructions reduced the accuracy of causal explanations (Gavanski & Hoffman 1987). An interpretation of this outcome is that introspection during the judgment process can interfere with the judgment process, making the judgments less reliable, so that the relationship between causal stimuli and judgments is less consistent and harder for subjects to detect. Thus, introspection on our thought processes during a task can change the natural flow of thoughts and make introspective reports less accurate. This mutual interference between concurrent introspection and task-related thought processes would be expected to be greater the more complex the task at hand. Thus, reports of concurrent introspections of task-related thoughts should be more accurate for simpler tasks—provided that the task is novel and interesting enough to engage our attention.

SUMMARY

Nisbett and Wilson (1977) were concerned with the self-attribution problem, concerning the processes by which people attempt to explain the causes of their own behavior. As a criterion for *introspective access* to the causes of one's own behavior, they proposed that access would be shown if subjects ("actors") who respond in a particular experimental situation can subsequently make verbal reports on the causes of their responses that are more accurate than reports made by observers who have only general information about the situation. They reviewed numerous experiments that, they argued, supported three major conclusions: (1) People do not have introspective access to the causal relationships between stimuli and their responses. That is, they cannot accurately report, from introspection, *which* stimuli affected their responses, and/or they cannot report *how* the stimuli affected their responses. (2) Reports of effects of stimuli on responses are based not on introspection, but on *a priori theories* (prior beliefs) about the causal connections between the stimuli and responses. (3) When people's reports on stimulus-response relationships are correct, it is because their a priori theories happen to be correct, not because of correct introspection.

Critics of the Nisbett-Wilson position argued that their supporting research used between-subjects designs, in which each subject experienced only one stimulus condition. Thus it would be hard for subjects to judge causal effects of stimuli because they had no chance to observe that different stimuli were associated with different responses (stimulus-response covariation). Several more recent experiments have used within-subjects designs, in which each subject responds on several occasions to different stimuli (such as judgments of the person characteristics—friendliness, etc.—of different people). Most within-subjects experiments have found actors' reports on the causes of their behavior to be somewhat more accurate than observers' reports. More recent research suggests that—contrary to Nisbett and Wilson's

argument—people's reports on the causes of their behavior are not based entirely on a priori theories. Observations of stimulus-response covariation, and introspective access to pertinent events, are also important. Introspective access is expected to be more accurate for behavior in novel, interesting tasks, as contrasted with habitual, automatic task performances.

ENDNOTES

[1]For readers who are unfamiliar with correlation coefficients: A correlation coefficient (r) is a mathematical index of the degree to which two different variables (such as measures of behavior or experience) covary, or change together, such that high scores on one variable (variable A) tend to go with high scores on the second variable (variable B), and low scores on the first variable go with low scores on the second one. To compute the correlation it is necessary to have a set of scores (both A and B) for each member of a group of subjects. A perfect correlation, $r = +1.0$, means that the higher the value (or magnitude) of variable A, then the higher the value of variable B, so you could predict B if you knew A. (Or, a perfect -1.0 correlations means that high scores on A go with low scores on B.) At the other extreme, if $r = 0.0$, then the two variables are unrelated, so knowing the value of variable A would be useless for predicting B. Usually a correlation of about 0.7 to 1.0 is considered to be high, 0.5 to 0.69 is moderately high, 0.3 to 0.49 is moderate, 0.2 to 0.29 is low, and below 0.2 is very low to negligible.

[2]Wilson (1985) raised the question whether people have introspective access to *mental states*, such as attitudes, desires, and feelings that influence their behavior. He argued that inconsistencies between verbally reported attitudes and certain unregulated behaviors—behaviors not under voluntary control, such as spontaneous facial expressions and physiological responses—indicate that people sometimes do not have accurate introspective access to mental states. Furthermore, attempts to increase the accuracy of attitude reports through close introspective attention to the reasons for one's attitudes, can actually reduce the accuracy of attitude reports, under some conditions (Wilson, Dunn, et al. 1989). Space limitations prevent me from describing this research, but you should read the articles cited here if you want to learn more about the limitations of introspection.