ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES
Vol. 65, No. 1, January, pp. 1–17, 1996
ARTICLE NO. 0001

Effects of Procedural and Outcome Accountability on Judgment Quality

KAREN SIEGEL-JACOBS

University of Colorado, Boulder

AND

J. FRANK YATES

University of Michigan, Ann Arbor

Individuals are said to be accountable whenever their performance is monitored and there are consequences (either tangible or intangible) associated with that evaluation. We propose that there are at least two distinct types of accountability. One focuses on justification of the procedure used to arrive at an action (procedural accountability, or PA); the other focuses on the quality of the outcomes of that action (outcome accountability, or OA). In three experiments, subjects judged the likelihood that each of a set of individuals held a particular attitude on the basis of background information about those individuals. Results from Experiments 1 and 2 suggested that PA encourages people to take more of the available information into account. Whether or not this effect was beneficial, however, depended on how relevant that information was to the target judgment. OA had only detrimental effects, increasing the amount of noise (or "scatter") in subjects' judgments and thus leading to lower accuracy overall. Experiment 3 extended the work on PA, focusing on its relationship to the effects of outcome feedback. Results indicated that PA significantly reduced the tendency to be overly responsive to such feedback by reducing the variability in judgment unrelated to the target event. Practical and theoretical implications are discussed.

We thank Susan Gelman, Marlys Lipe, Colleen Seifert, and Edward Smith for their comments and suggestions on an earlier version of this paper. This work was supported in part by the Arthur F. Thurnau Endowment at the University of Michigan, by National Institute for Mental Health training Grant MH16892 at the University of Michigan, and by National Institute for Mental Health training Grant HHS5T32MH14617-17 at the University of Colorado. Correspondence and reprint requests should be addressed to Karen Siegel-Jacobs, US WEST Advanced Technologies, Inc. 4001 Discovery Drive, Boulder, CO 80303. E-mail: Ksiegel@advtech.uswest.com. or to J. Frank Yates, Judgment and Decision Laboratory, Department of Psychology, University of Michigan, 525 East University Avenue, Ann Arbor, MI 48109-1109. E-mail: jfyates@umich.edu.

Consider the following situations. In the first, a physician must decide between alternative treatments for a patient with cancer. She is aware that no matter what she decides, a suboptimal outcome may lead to a malpractice suit. In a second scenario, a stock broker is asked by a new client how likely it is that a given stock will increase in value over the next few months. The broker knows that he is most likely to have a loyal customer if his client makes money on the basis of his advice. A common feature of both scenarios is that a person's judgments or decisions are being monitored and evaluated for their quality. Furthermore, the evaluation implies some consequences for the judge or decision maker. When an individual takes this anticipated evaluation into consideration, we say that he or she feels "accountable" for his or her actions.[1]

The consequences associated with the evaluation of a judgment or decision need not actually be tangible in order for accountability to occur. For example, it may be enough that the accountable individual cares about the evaluator's level of regard. Essential to the present definition of accountability, however, is that the evaluation be done by someone other than the judge or decision maker. While it is certainly conceivable that one could feel accountable to him- or herself, the present discussion was not intended to generalize to such a situation.

Interestingly, although a great many real-world judgment and decision-making situations clearly include some level of accountability, most laboratory studies of judgment and decision making are deliberately designed to minimize this factor. For example, subjects are generally told that any responses they make during the course of the experiment will be kept strictly confidential and anonymous, and there are sel-

[1] Although we generally prefer the term "accountability," we do not intend to distinguish it conceptually from a need to "justify" one's actions.

0749-5978/96 $18.00

dom any real consequences for making "good" versus "bad" judgments or decisions. This disparity between typical laboratory conditions and the circumstances of so many real-life judgment and decision-making situations highlights an important reason for addressing the impact of accountability on behavior. Specifically, to the extent that behavior under conditions of accountability differs from behavior in the absence of accountability, our ability to generalize from laboratory studies to real-life situations is constrained.

A substantial body of research already exists which suggests that accountability does indeed affect judgment and decision behavior. In his insightful review of this literature, Tetlock (1985) makes the case that whether accountability is harmful or beneficial depends in large part on the degree of ambiguity in the task. In those situations where an accountable individual knows what response the evaluator will find acceptable (unambiguous), there is a tendency to conform to that standard (the "acceptability heuristic") in order to win approval.

When the evaluator's standard is unknown (ambiguous), however, simple conformity is obviously no longer an option. Assuming that the judge or decision maker still cares about winning approval, his next best option will be to find the most *defensible* course of action available. He may either try harder to discern what the particular evaluator might approve of under the given circumstances, or he might attempt to arrive at the most broadly defensible response possible. In either case, it is clear that he will have to process information more carefully (or "vigilantly," to use Tetlock's (1985) terminology) than would be required in a situation involving simple conformity. This increased vigilance is then hypothesized to produce effects such as higher levels of judgment accuracy, judgment consistency, complexity of thinking, and amount of information processed.

Tetlock's classification scheme does indeed have considerable explanatory power. If we stop here, however, we potentially ignore an important additional variable. Specifically, a closer examination of the literature reveals that there are in fact two distinct types of accountability which to date have not been distinguished. In the first type, which we call procedural accountability (PA), evaluation is based solely on the quality of the *procedure* that a judge or decision maker uses in arriving at a response, regardless of the quality of the outcome of that response. For example, in the medical domain, procedural accountability might take the form of requiring a physician to justify how a particular course of treatment was chosen, regardless of whether the patient ultimately got better or not. Evaluation in the second type of accountability (outcome accountability, or OA), on the other hand, is based exclusively on the quality of the *outcome* of a response (e.g., how the patient fared under the chosen treatment regime), without regard to the nature of the procedure used to arrive at that response.

Given that people generally want others to view them in a positive light (Baumeister, 1982), we might expect either kind of accountability simply to act as a general motivator, with the result being that people try harder to do a task well. There are at least two reasons, however, to expect the alternative kinds of accountability to have different effects. First, while outcome accountability may provide an additional incentive to produce a positively evaluated response, there is no guidance inherent in the manipulation as to how to achieve that goal (somewhat like simply shouting "get a hit" to the batter in a baseball game). Procedural accountability, on the other hand, actually suggests a method for enhancing performance—try changing the nature of the procedures being used (check your stance, think about your grip, etc.).

A second difference between the two kinds of accountability involves the perceived ease of improving performance. Since most real-world judgment and decision problems involve some level of irreducible uncertainty Hammond, 1995, even with the best possible procedure, there is almost always a chance of obtaining a suboptimal outcome. Therefore, if evaluation is based solely on outcome quality, an individual who has used the optimal procedure might still be considered to have failed. When an individual merely has to arrive at a justifiable *procedure* for approaching a problem, on the other hand, the uncertainty inherent in the situation does not pose the same handicap, as a suboptimal outcome does not necessarily cast doubt on the quality of the process.

According to Janis and Mann's (1977) conflict theory, when a decision has important consequences, but the decision maker feels that there is a good chance of failing to produce an adequate solution (as in outcome accountability), a high level of stress is induced. High levels of stress in turn have been shown to be detrimental to various aspects of judgment and decision processes (Keinan, Friedland, & Ben-Porath, 1987; Rothstein, 1986; Yates, 1990; Young & Yates, 1991). When a decision maker is confronted with a problem for which there is hope of finding a suitable solution, on the other hand, conflict theory predicts that the individual will feel a moderate amount of stress. Although we typically associate stress of any kind with negative effects, relatively low levels of stress have actually been shown to be beneficial, encouraging the decision maker to carefully and systematically consider relevant information (Janis & Mann, 1977; Yates, 1990; Young & Yates, 1991).

Based on the above distinctions, there is reason to

expect procedural accountability to have more beneficial effects on performance than outcome accountability. Procedural accountability gives the judge or decision maker a possible starting point for enhancing performance and, because it is unaffected by the uncertainty inherent in many situations, may seem like a more solvable (and therefore, potentially less stress inducing) problem than being held accountable for outcome quality.

In order to determine the extent to which the above prediction has held true in the accountability literature to date, it was necessary to categorize studies post hoc (as involving either procedural or outcome accountability). The following guidelines for classification were constructed based on our definition of the two kinds of accountability. Studies where the judge or decision maker was held accountable for the outcome, without regard to the procedure used, were classified as outcome accountability studies. Studies where the individual was asked to justify or explain how or why he or she arrived at a particular response (regardless of how satisfactory that response might be) were put into the procedural category.

A brief review of the studies incorporating procedural accountability reveals that it does indeed appear to have largely positive effects on judgment and decision quality. In particular, procedural accountability has been shown to increase the consistency with which a judgment strategy is executed (Ashton, 1992; Hagafors & Brehmer, 1983), encourage more complex and analytic modes of processing (Chaiken, 1980; McAllister, Mitchell, & Beach, 1979; Tetlock, Skitka, & Boettger, 1989), motivate subjects to take more of the available information into account (Tetlock, 1983; Tetlock & Boettger, 1989), and improve various aspects of overall judgment and decision quality (Arkes, Christensen, Lai, & Blumer, 1987; Simonson & Nye, 1992; Tetlock & Kim, 1987).

In contrast, studies employing outcome accountability have often revealed detrimental effects. In particular, subjects have been found to attempt to represent (or misrepresent) themselves in a more flattering light (Fandt & Ferris, 1990), to be less able to reach a satisfactory agreement in a negotiating context (Klimoski, 1972), and to be more wasteful in the distribution of resources to needy individuals (Adelberg & Batson, 1978). Even when the nature of the accountability is positive as opposed to negative (i.e., rewards for good performance rather than penalties for bad), outcome accountability still appears to do more harm than good. For example, Arkes, Dawes, and Christensen (1986) report a reduction in within-subject consistency as a result of incentives.

In summary, it appears that whereas procedural accountability has generally proved to be beneficial in previous research, outcome accountability has had largely detrimental effects on judgment and decision quality. If in fact this difference is found to hold, one important implication is that, even within one of Tetlock's (1985) categories (e.g., ambiguous tasks), we cannot make accurate predictions regarding the effects of accountability without also attending to the *type* of accountability demanded. Therefore, our first goal was to determine empirically whether procedural and outcome accountability induced differential effects given the exact same ambiguous judgment situation.

Our second empirical goal was to explore some of the mechanisms by which procedural accountability operates. In particular, many investigators have claimed that procedural accountability alters (most often by increasing) the amount of information considered in making a judgment or decision. However, in each of these studies, information use has been inferred from other indices, such as a reduction in the primacy effect in impression formation (Tetlock, 1983), an increase in the dilution effect (Tetlock & Boettger, 1989), analysis of subject-generated "thoughts" protocols (Tetlock, Skitka, & Boettger, 1989), and analysis of the weights produced by constructing simple linear models of subjects' judgments (Weldon & Gargano, 1988). There are many other examples of studies involving indirect measures (e.g., Chaiken, 1980; Rozelle & Baxter, 1981), but none has directly monitored information use "on-line." Therefore, we sought to test directly the claim that procedural accountability encourages more complete processing of external information.

Finally, we addressed the fact that in everyday life, accountability demands often go hand-in-hand with some form of feedback about performance. For example, we are most often asked to justify our actions when there is some indication that desired outcomes have not been obtained. One of the most well documented results from the feedback literature is that (case-by-case) *outcome* feedback actually tends to reduce judgment accuracy, mainly because it causes a reduction in the consistency with which people apply their judgment policies (Arkes, Dawes, & Christensen, 1986; Brehmer & Kuylenstierna, 1978; Hammond, Summers, & Deane 1973; Schmitt, Coyle, & King, 1976).

Recall that procedural accountability, on the other hand, was found to *increase* judgment consistency in a number of studies (e.g., Ashton, 1992; Hagafors & Brehmer, 1983). While little has been offered by way of an explanation for this effect, two alternatives seem plausible. First, procedural accountability may encourage people to think harder about the best strategy from the outset, making them less willing to abandon that strategy. Alternatively, people might simply fear appearing "wishy-washy" by constantly shifting from one

strategy to another. In either case, the prediction is that procedural accountability should diminish the negative effects of outcome feedback on consistency.

The following experiments were designed to address each of the issues outlined above. In all cases, we chose to use probability judgments rather than decisions as our dependent variable, where a judgment is defined as an opinion about the status of some event in the real world. There are two reasons for this choice. First, a great many decisions actually rest on judgments. Therefore, in a sense, judgments are the more fundamental of the two classes of problems. Second, it is much easier to establish the quality of judgments than the quality of decisions. This is because the event about which one is making a judgment ultimately either occurs or does not occur, making the correctness of the judgment unambiguous (Yates, 1990). A good decision, on the other hand, has been defined as one that produces an outcome at least as satisfactory as would have been achieved had any other alternative been selected (Yates, 1990). Since in many cases it is not possible to *know* what outcome would have obtained given an alternative other than the one selected, however, it is often extremely difficult to determine decision quality in practice.

EXPERIMENT 1

Experiment 1 was designed to address the first two issues discussed above—type of accountability and its effect on information use. We presented all subjects with the same task, in which they judged the likelihood that each of a set of individuals held a particular attitude, based on some background information (cues) about that individual. No accountability, procedural accountability, and outcome accountability conditions were all included. We then monitored information use directly, in order to ascertain whether any differences that might be observed given different types of accountability could be traced to systematic variations in how externally provided information was being used. As noted previously, although the idea that accountability operates in part by increasing the amount of information to which judges attend is by no means an original one, researchers have thus far relied on indirect measures to test this claim (e.g., Chaiken, 1980; Rozelle & Baxter, 1981; Tetlock, 1983; Tetlock & Boettger, 1989; Tetlock, Skitka, & Boettger, 1989; Weldon & Gargano, 1988). While the results of these studies are persuasive, it is possible to produce each of the effects by means other than increasing the amount of information attended to. Therefore, in this experiment, the use of externally provided information (cues) was monitored.

In order to make specific predictions, it is helpful first to introduce the dependent measures we employed, and it is to this topic that we now turn.

Dependent Measures

At first glance, judgment accuracy appears to be a fairly simple concept—a judgment is either right or wrong. As numerous individuals have shown, however, the ability to make accurate probability judgments is in fact a relatively complex skill, entailing numerous component skills (Murphy, 1973; Yates, 1990). One of the most commonly used measures of *overall* probability judgment accuracy is the "Brier score" (Brier, 1950), also known as the "mean probability score" ($\overline{PS}$). A more detailed explanation of $\overline{PS}$ and its components is provided in the Appendix. Here, we merely attempt to convey the general ideas behind the measures, as all serve as dependent measures in the present studies.

Suppose we have some event about which judgments are being made, such as rain occurring on a given day. We will call this event the "target event." We can indicate the true state of the world with respect to this target event by using an "outcome index," designated "d." We assign the value 1.0 to d whenever the target event does in fact occur (e.g., it rains on the day in question) and the value 0 on all occasions when the target event does not occur. $\overline{PS}$ is simply the squared difference between the probability assigned to the target event and the value of the outcome index for that instance, averaged over all judgment occasions. Ideally, we would prefer a judge who assigned high probabilities to events that did occur and low probabilities otherwise, with the optimal value of $\overline{PS}$ being zero.

Murphy (1973; see also Yates, 1990, Chapter 3) showed that $\overline{PS}$ can be decomposed into three components of interest, according to the following formula:

$$\overline{PS} = \text{Var (d)} - \text{Discrimination Index (DI)} + \text{Calibration Index (CI)}. \quad (1)$$

The first component of the Murphy decomposition, Var (d), or the outcome index variance, reflects the variability inherent in the target event. As long as the target event is something whose outcome is determined in the real world (also known as an "external" target event), the outcome index variance will be constant for a given set of judgment occasions. For example, even if two judges assign very different chances of rain for the same ten days, their *outcome index variances* will still be identical.

Discrimination, the second contributor to $\overline{PS}$, refers to the extent to which probabilities assigned when the target event does occur differ from those assigned when it does not. The higher the discrimination index (DI), the better. So, for example, if on all occasions

when the target event occurs (e.g., it rains), a judge assigns a likelihood of 0.8, and whenever the target event fails to occur (no rain), the judge assigns the value 0.2, the judge is exhibiting perfect discrimination.

Psychologically, discrimination is in a sense a measure of the judge's clairvoyance, since perfect discrimination implies the ability to distinguish infallibly those occasions when a target event will occur from those when it will not. From a "consumer's" perspective, good discrimination is clearly a desirable trait in a judge. If the judge described above were our local weather forecaster, we could in theory avoid ever carrying an umbrella needlessly or getting soaked in an unanticipated downpour.

The last component of Murphy's decomposition of $\overline{PS}$ is known as the calibration index, or CI. Perfect calibration means that, over all items for which the judge says the probability of the target event occurring is .XX, the actual relative frequency of occurrence is XX%. For example, it should rain on 80% of the days for which a perfectly calibrated judge says the chance of rain is 0.80. Calibration amounts to being able to quantify properly the uncertainty in a given situation. It can be improved by keeping track (intuitively or otherwise) of the correspondence between observed relative frequencies and one's numerical assignments. Calibration can be degraded by, among other things, poor memory, reliance on weakly predictive information, and inconsistency—factors that affect other elements of judgment accuracy too.

An alterative decomposition of the mean probability score, known as the *covariance decomposition* (Yates, 1982), was also employed (see Appendix for formulas). One of the main features of the covariance decomposition is that it allows separate examination of probabilities assigned on those occasions when the target event *does* occur versus judgments made when the target event does *not* occur. Clearly, we would want judgments to be *different* in the two cases and would generally prefer a judge who assigned high probabilities for those cases when the target occurred, and low probabilities when it did not. Subtracting the average target-negative judgment from the average target-positive judgment yields the first component of the covariance decomposition: *slope*. A judge who always assigns the value 1.0 when the target event occurs and assigns 0 when it does not occur will achieve the ideal value of 1.0 for slope, as well as the ideal value of 0 for $\overline{PS}$.

A second desirable trait in a judge would be that he or she say the same thing every time the target event was going to occur, and similarly, assign a single (albeit different) judgment on those occasions when the target event does not occur. In other words, any variance present in a set of judgments ideally should be linked to the occurrence or nonoccurrence of the target event. This particular feature of judgment accuracy is captured by the second component of the covariance decomposition—*scatter*. In essence, scatter reflects the extent to which judgments vary around their conditional means. Factors that would contribute to high (i.e., poor) scatter include inconsistent responding and incorporating irrelevant information when making a judgment. Again, a judge who always assigns the value of 1.0 when the target event occurs and assigns 0 when it does not occur will achieve the ideal value of 0 for scatter.

The third component of the covariance decomposition is *bias*. Bias simply reflects the difference between the average judgment rendered and the actual base rate or relative frequency with which a target event occurs over a given set of occasions. Someone who overestimates the likelihood of the target event occurring will show positive bias, while a judge who underestimates will have negative bias. Our ideal judge, who always assigns the value of 1.0 when the target event occurs and assigns 0 when it does not occur, will have no bias at all.

Finally, the last component of the covariance decomposition is the outcome index variance (Var (d)) first introduced as part of the Murphy decomposition (see Eq. (1) above). This component will not vary across subjects or conditions in the present experiment.

The above measures can be combined according to the following formula to obtain $\overline{PS}$:

$$\overline{PS} = \mathrm{Var(d)} + \mathrm{Bias}^2 + \mathrm{Var(d)(Slope)(Slope} - 2) + \mathrm{Scatter} \quad (2)$$

The most important thing to note is simply that $\overline{PS}$ will be lowest (best) when slope is high, when scatter is low, and when bias is nil.

The last set of dependent measures for Experiment 1 concerned information (cue) use. Specifically, the particular cues viewed, as well as the number of times any given cue was viewed, were recorded for each judgment.

We are now in position to make some specific predictions. Recall the general prediction that procedural accountability would be beneficial to judgment quality, due to increased attention to the process by which judgments are made. In terms of our dependent variables, the global prediction suggests that subjects may think more critically about the usefulness of the various items of available information when making their judgments, which in turn would be expected to lead to better discrimination (ability to distinguish target-positive from target-negative instances). Subjects in

this condition may also exhibit better calibration, since giving more thought to the usefulness of the available information should increase sensitivity to the level of uncertainty present for each judgment.

The predicted effects of outcome accountability are less positive. As suggested above, the combination of increased incentive to do well (i.e., produce the best outcome, or judgment), in the absence of guidelines for how to accomplish that goal, may be stressful enough to produce a performance decrement. Specifically, in accordance with previous research on the effects of stress (Keinan, Friedland, & Ben-Porath, 1987; Rothstein, 1986; Yates, 1990; Young & Yates, 1991), we expect more erratic (less consistent) responding on the part of subjects in this condition, which should be evidenced by increased scatter in their judgments.

Method

Subjects. A total of 122 undergraduates at the University of Michigan participated in order to fulfill a requirement of their introductory psychology course. Of the initial group of subjects, 55 failed to pass a post-experimental check of the accountability manipulation (discussed below). All results reported below were computed using only the remaining 67 subjects.

Procedures. Each subject was asked to play the role of a trial lawyer, and was given the following background information about a hypothetical case:

> Your client, Ms. Jones, has an incurable, degenerative disease, and wishes to stop receiving the treatments which are keeping her alive. However, the hospital is unwilling to take the necessary steps to help her. Consequently, your client has opted to take the matter to court in hopes that a jury will order the hospital to cease treatment at her request. Obviously, the selection of the jury for this case is of the utmost importance. You will want to choose the most sympathetic jury possible. Ideally, you would simply reject any potential jurors who admit to being opposed to suicide under the circumstances, and accept those who believe suicide is acceptable in such a case. Unfortunately, this is not possible, since you will not be given direct access to the potential jurors. Instead, all that you have is a file on each juror, containing his or her responses to seven preset questions which are asked of prospective jurors for *all* cases in the court. You will have to use the information in the file to the best of your ability to determine how likely it is that each juror believes suicide should be allowed under such circumstances.

All subjects made judgments for the same set of potential jurors. The first item of information in each prospective juror's file was that person's age. Item 2 concerned the extent of the juror's education, ranging from no formal schooling up to 8 or more years of college. Political party affiliation (from strong Democrat to strong Republican) was the third item. Item 4 concerned whether or not the prospective juror favored allowing abortion in cases where there was a strong chance of a serious defect in the baby. The fifth item was whether or not the prospective juror owned a gun. The sixth item concerned the prospective juror's feelings about premarital sex, ranging from "always wrong" to "not wrong at all." Finally, Item 7 was the potential juror's gender. Subjects were told explicitly that these questions were asked of all prospective jurors for all cases in the court, meaning that the responses were not necessarily relevant for the particular judgment they were being asked to make.

The subject was allowed to view as many (or as few) of the items, or "cues," as he or she wished for each prospective juror, and to view any given cue as many times as desired. However, only one cue value could be viewed at a time. On each trial, a screen appeared with the labels corresponding to the seven categories of available information (see Fig. 1). Subjects entered the number next to each label in order to view the actual cue value for that juror. Subjects were allowed to select any number of cues (including repetitions) in any order, and could view cues for as long as desired. When finished viewing cues, the subject pressed a key to move on to the judgment screen, where he or she entered the probability that this prospective juror believed a person should be allowed to end his or her own life given an incurable disease. Judgments were made for 48 jurors in all.

The set of jurors for whom judgments were made was drawn from a larger pool of people who answered all eight questions (seven cues and the criterion) in the 1989 General Social Survey (Davis & Smith, 1989). The sample was selected randomly with the constraint that the multiple correlation between the criterion and the seven predictors be within 0.02 of that for the full set of interviewees. Care was also taken to ensure that the correlations between cues and the criterion, as well

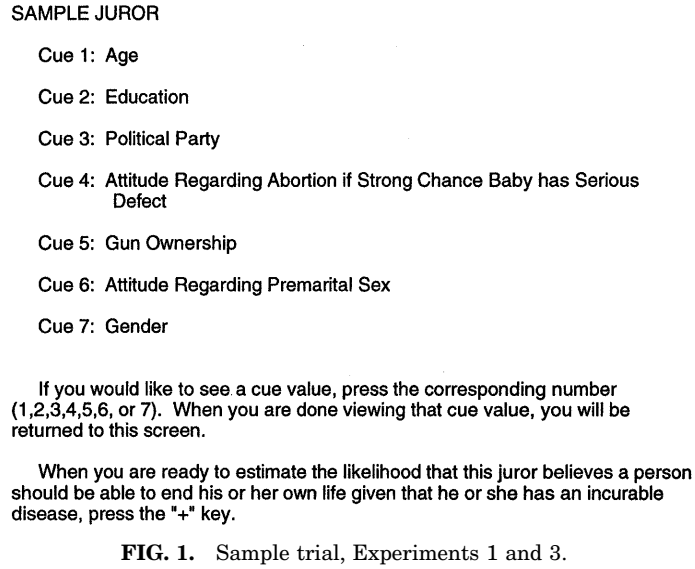

FIG. 1. Sample trial, Experiments 1 and 3.

TABLE 1
Experiment 1: Cue/Criterion Correlations for Sample and Population

Cue	Source		
	Sample	Population	Significance (p)
Abortion, defect	0.39	0.41	<.01
Premarital sex	0.30	0.36	<.01
Education	0.17	0.17	<.01
Age	0.16	0.17	<.01
Gender	0.09	0.08	<.01
Own gun	0.04	0.05	.53
Party identification	0.02	0.02	.74

Note. Criterion = junior's feelings regarding the permissibility of suicide in cases where the individual has an incurable disease.

as among the cues themselves, mirrored those for the population as closely as possible. The correlations between cues and the criterion, for both the sample and the population are shown in Table 1. Because it is difficult to interpret these correlations without knowledge of how each question was coded in the original survey, we summarize the direction of the relationships here: People who favored allowing abortion in cases of fetal defects were more likely to favor allowing suicide, as were people who were more tolerant of premarital sex. The more education a respondent had, and the younger the respondent, the more likely he or she was to favor allowing suicide. Males were more likely to favor suicide than females, as were gun owners. Finally, the closer one's political identification with the Democratic Party was, the more likely that person was to favor allowing suicide in cases of terminal illness.

Also shown in Table 1 are the results of significance tests to determine whether each cue-criterion correlation differed reliably from zero for the population of participants in the survey. As can be seen, using an alpha level of 0.05, only the first five cues are actually valid predictors of the criterion. Both valid and nonvalid predictors were made available in order to make the task more consistent with the kinds of judgments people face every day, in which the determination of what information is truly useful is a key part of the process.

Subjects were given no explicit information regarding cue-criterion relationships. However, all were told at the outset of the experiment that the set of prospective jurors was in fact a random, and thus representative, sample of Americans. The background information they would see in each case was real, and the juror's true attitude toward suicide under the given circumstances was on record.

Each subject was randomly assigned to one of three conditions with the constraint that if two subjects participated at the same time, they were both assigned to the same condition. All subjects were in fact scheduled to participate in pairs, in hopes of augmenting any accountability effects in the procedural accountability condition. Poor subject attendance, however, resulted in 28 of the 122 subjects being run individually—seven in the no-accountability (NA) condition, 15 in the outcome accountability (OA) condition, and six in the procedural accountability (PA) condition.

PA subjects were informed that when the experiment was over, the items they chose to examine for a subset of the prospective jurors would be displayed on the screen, along with the judgment made in each case. An interview would then be conducted concerning why and how they used the information they did to make their judgments. They were asked to sign a consent form permitting the experimenter to tape-record the interview for use in further analysis, and taping equipment was placed in plain view.

OA subjects were informed that a score would be computed for them based on the accuracy of their judgments. Once the entire experiment had been completed, they would receive feedback about their performance relative to the other subjects in the study. Furthermore, the people with the top five scores (out of approximately 40 total participants in this condition) would receive a bonus prize of 10 dollars each. There was no mention of a postexperimental interview. Finally, NA subjects were simply reminded that all of their responses would be kept confidential and anonymous.

Manipulation check. The last thing each subject did was fill out a manipulation check questionnaire. The manipulation check given to subjects in the NA and PA conditions is shown in Fig. 2. Subjects were asked whether and to what extent they believed they would have to justify their responses post-experi-

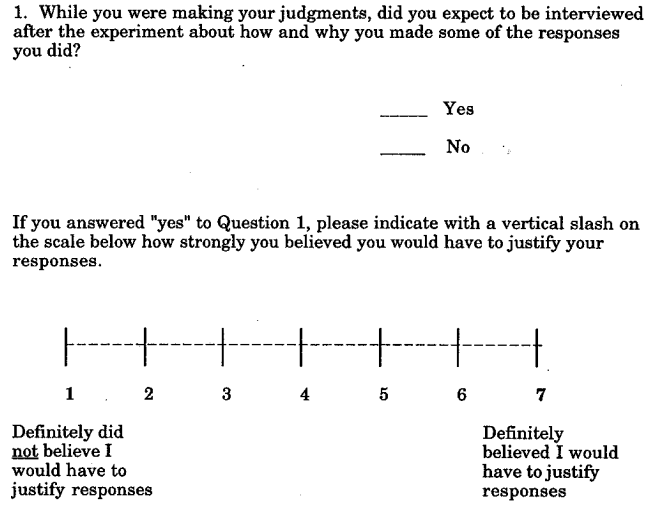

FIG. 2. Manipulation check, Experiment 1.

mentally. Four PA subjects who either answered "no" to Question 1, or who indicated that they thought the chance of being interviewed postexperimentally was less than "4" (midpoint of the scale), were excluded from the data analysis. The highest number of failures occurred in the NA condition, where a total of 26 subjects had to be eliminated because they indicated that they believed they *would* be interviewed postexperimentally. This anticipated accountability occurred even though we at no time made any mention of an interview and in fact attempted to assure these subjects of the opposite by stressing the anonymity and confidentiality of their responses.

The manipulation check for OA differed only in that subjects were also asked whether or not they believed they would be eligible for the bonus if their score was in the top five. Twenty-five subjects in the outcome accountability condition either responded that they did *not* believe in the bonus manipulation, or, as in the no accountability condition, indicated that they also expected to be interviewed at the end of the experiment concerning the procedure they used. Therefore, these subjects were not included in the remaining analyses.

All manipulations (bonuses and interviews) were in fact legitimate.

Results

Two subjects (one from the PA condition and one from the OA condition) had mean probability scores almost twice as high as would be expected due to chance responding. According to Grubb's test (see Dunn & Clark, 1974), both cases were considered outliers, and hence the data from these two subjects were not included in further analyses. Since cue usage was exceedingly variable, even within condition, we also used Grubb's test for the cue usage data. This resulted in nine subjects being excluded from the analysis of age, seven from the analysis of gender and premarital sex, three from the analysis of abortion, and two from the analysis of party identification.

There was no significant difference in the average number of cues used per judgment across conditions. Differences did emerge, however, for individual cues. Over the 48 possible opportunities to use each cue, PA subjects viewed the premarital sex cue significantly more often on average ($M = 47.0$) than did subjects in either the OA condition ($M = 43.5$), $t(34) = 2.19$, $p < .05$, or the NA condition ($M = 36.2$), $t(40) = 1.85$, $p < .05$. PA subjects also viewed the gender information ($M = 45.4$) and used the party identification information ($M = 43.2$) significantly more often than did subjects in the NA condition ($M = 37.5$ and 35.6, respectively, $t(41) = 2.35$, $p < .05$, and $t(44) = 1.85$, $p < .05$). Finally, OA subjects also viewed the gender cue ($M = 46.9$) significantly more often than did subjects in the NA condition ($M = 37.5$), $t(33) = 2.31$, $p < .05$. No other differences with respect to cue usage were significant.

Table 2 shows the results for the Murphy decomposition of $\overline{PS}$. There were no significant differences between conditions with respect to discrimination. However, PA subjects showed significantly better calibration than either the NA or the OA subjects. PA subjects also had significantly better mean probability scores than did OA subjects. The difference between PA and NA subjects with respect to $\overline{PS}$ was marginally significant.

Since the data violated assumptions of normality in many cases, only non-parametric tests are reported in Table 2. Parametric tests were conducted, however, with the only difference in results being an additional significant difference (rather than a marginally signifi-

TABLE 2

Experiment 1: Effects of Varying Types of Accountability on Accuracy, Mixed Cue Diagnosticity

Measure	Medians: No accountability (NA) condition	Medians: Procedural accountability (PA) condition	Medians: Outcome accountability (OA) condition	*U* (NA vs PA)	*U* (NA vs OA)	*U* (PA vs OA)
$\overline{PS}$ (Brier score)[a]	.264	.258	.286	2.00***	0.64	5.11*
Discrimination Index[b]	.046	.050	.057	0.15	1.36	0.75
Calibration Index[a]	.081	.057	.101	3.16*	1.74***	6.06**
Slope[b]	.114	.130	.121	1.10	0.41	0.25
Scatter[a]	.053	.046	.073	0.01	3.73*	4.04*
Bias[c]	.129	.131	.113	0.58	0.11	0.23

[a] Indicates small values are better.
[b] Indicates large values are better.
[c] Indicates zero value is best.
* $p < .05$.
** $p < .01$.
*** $p < .10$.

cant one) between no accountability and outcome accountability subjects with respect to calibration.

Table 2 also shows the results for the covariance decomposition. There were no significant differences between conditions in slope or bias. Both no-accountability subjects and procedural accountability subjects did show significantly less scatter than did outcome accountability subjects. Parametric tests revealed an identical pattern of results. The data were also analyzed using only subjects who had participated in pairs. Again, the same pattern of significant results was found.

Discussion

As predicted, PA subjects were better calibrated than NA subjects. This result is consistent with the results of a study by Tetlock and Kim (1987), who also used the Murphy decomposition to determine the effects of accountability on judgment quality.

Surprisingly, however, PA subjects did not have better discrimination than NA subjects. The data on cue usage suggest the most likely explanation for this result. Specifically, while procedural accountability promoted use of two of the cues that *are* valid predictors of feelings about suicide (premarital sex attitude and gender) relative to no accountability subjects, it also encouraged the use of one of the nonvalid cues (party identification). Recall that discrimination requires the ability to distinguish reliably those occasions when a target event will occur from occasions when it will not. Factoring irrelevant information into the equation can only be harmful in this regard.

Why did procedural accountability increase the use of nondiagnostic as well as diagnostic cues? We suspect that the cue-use results may have had more to do with subjects' assumptions about the evaluator's values than with subjects' actual assessments of cue validity. As we have discussed, there is ample evidence that the desire to be liked or to win approval is a powerful motivating force in human behavior (see Baumeister, 1982, for a review). Note, however, that subjects in the procedural accountability condition were told nothing about the evaluator's standard. In the absence of any clear guidelines, these subjects may have assumed that the evaluator would adhere to the well-known aphorisms "when in doubt, check it out" and "leave no stone unturned." If so, then even if PA subjects were *aware* that some cues were essentially worthless, there would still be some incentive to try to incorporate them into their judgment policies, given that pending evaluation depended entirely on the assessed quality of the *procedure* employed.

A series of studies by Simonson and Nye (1992) supports this idea that the procedural accountability subject is driven, at least in part, by the evaluators' perceived standard. These investigators studied situations where an individual knows the rational procedure for making a decision, but where the norm is to decide nonrationally anyway. For example, although most of their subjects acknowledged that it was not rational to take into account sunk costs, nonaccountable subjects nevertheless clearly did so when making decisions. Accountable subjects, however, were significantly less likely to consider sunk costs in making their decisions, presumably because accountability encouraged them to act in accordance with what the *evaluator* would value (e.g., consider rational).

In contrast to procedural accountability, outcome accountability had only negative effects when compared with no accountability. Calibration was marginally worse, and scatter was significantly greater when the emphasis of the accountability was on outcomes. The latter finding is consistent with the claim that outcome accountability produces stress, which in turn has been shown to produce more variable or inconsistent response patterns (Keinan, Friedland, Ben-Porath, 1987; Rothstein, 1986). Although subjects in the outcome accountability condition did pay more attention to one of the valid predictors (gender), it was the least useful of the set, with a correlation of only .08 with the criterion.

Even more dramatic contrasts emerged when outcome accountability was compared directly to procedural accountability. Since procedural accountability improved calibration relative to no accountability, while outcome accountability was detrimental compared to no accountability, the difference between procedural and outcome accountability with respect to this measure was highly significant. Outcome accountability also produced significantly more scatter in comparison to procedural accountability. Finally, given the above two results, it is not surprising that outcome accountability led to significantly poorer scores with respect to overall accuracy than did procedural accountability.

The most obvious conclusion to be drawn from these results is that there is a need to attend to the nature of the accountability being manipulated before useful predictions about the effects of such accountability can be made. If the justification requirement focuses attention on the *procedure* used to make judgments, then accountability will be helpful in some regard (though this may depend in part on what the evaluator's standard is assumed to be). In contrast, outcome accountability had no beneficial effects whatsoever, and in fact was harmful in some respects. Although this result may seem counterintuitive (how can offering a reward for good performance hurt?), it is consistent with previous work on the effects of stress and incentives (Arkes, Dawes, & Christensen, 1986; Janis & Mann,

1977; McGraw, 1978). If an individual is motivated to do better, but does not know how to go about accomplishing this goal, consistency of behavior will be reduced and overall performance disrupted.

EXPERIMENT 2

In this experiment, we further investigated the hypothesis that the lack of improvement in discrimination for procedural accountability subjects in Experiment 1 could be explained by an increased effort to take into account nondiagnostic information. Specifically, in experiment 1 we presented subjects with a judgment task similar to that used in Experiment 1, but this time made available only relevant (valid) information. If our hypothesis is correct, then under these conditions, procedural accountability should increase discrimination relative to no accountability.

Method

Subjects. Subjects were 30 male and 51 female undergraduates at the University of Michigan. Fifty-four of the subjects took part in the experiment in order to fulfill a requirement of their introductory psychology course, and 27 received extra credit in a developmental psychology course for their participation.

Procedure. Subjects were led through the procedure individually. Their task was to imagine that they were helping a political candidate assess how a subset of his voting constituents felt about abortion. More specifically, for each of 48 voters, subjects had to judge how likely it was that the voter would favor permitting abortion given the condition that a family was too poor to afford additional children. Judgments were made on the basis of information about the highest educational degree each voter had obtained, the voter's personal income for 1988, and which of the following four attitudes regarding premarital sex best matched the voter's own feelings:

(1) It is always wrong for a man and a woman to have sexual relations before marriage.

(2) It is almost always wrong for a man and woman to have sexual relations before marriage.

(3) It is only sometimes wrong for a man and a woman to have sexual relations before marriage.

(4) It is not wrong at all for a man and a woman to have sexual relations before marriage.

The set of 48 voters for whom judgments were made was drawn from a larger sample of 272 people who answered the abortion question in the 1989 General Social Survey (Davis & Smith, 1989). The sample was selected randomly, with the constraint that the multiple correlation between the criterion and the three predictors be within 0.02 of that for the full set of interviewees. All information regarding stance on abortion, 1988 personal income, level of education, and attitude toward premarital sex was taken directly from the survey records.

The actual correlation in the data set between attitude toward premarital sex and the abortion question was 0.35, the correlation between educational level and abortion attitude was 0.20, and the correlation between income and abortion was 0.16. All correlations were significantly different from zero. The multiple R found by regressing all three predictors on the criterion was 0.38.

Subjects in the accountability condition were told that, following the experiment, some of the items would be selected at random, and they would be interviewed in detail concerning how they arrived at a judgment for each of these voters. They were also told that, with their permission, the interview would be tape recorded for use in future data analyses. Subjects who agreed to be audiotaped signed a consent form to this effect. Subjects in the no-accountability condition, in contrast, received no warning of the postexperimental interview. Rather, these subjects were reminded that their responses would be anonymous.

All subjects did in fact participate in an interview following the judgment task. The experiment was both preceded and followed by an unrelated task.

Results

Two-tailed Students' t tests were performed on the measures of overall accuracy, calibration, and discrimination, and on total time. The results are shown in Table 3.

Subjects in the accountability condition spent significantly longer on the task than did subjects in the no-accountability condition. There was no significant difference between conditions in terms of calibration or

TABLE 3

Experiment 2: Effects of Procedural Accountability on Accuracy Given Relevant Cues

Measure	Means: Accountability condition	Means: No-accountability condition	$t(79)$
Total time (min)	14.3	11.8	2.68**
$\overline{PS}$ (Brier score)[a]	.226	.272	−0.83
Discrimination Index[b]	.054	.043	2.41*
Calibration Index[a]	.070	.066	0.54

[a] Indicates small values are better.
[b] Indicates large values are better.
* $p < .05$.
** $p < .01$.

overall accuracy. Accountable subjects, however, did demonstrate significantly better discrimination. Since the characteristics of the sampling distributions for the accuracy measures are unknown, Mann–Whitney tests were also conducted. The pattern of results were exactly as found using Student's t tests.

Discussion

As predicted, when subjects were only given access to relevant information, introducing procedural accountability increased discrimination. Although cue use was not monitored directly in this experiment, the fact that accountable subjects also spent more time on the task suggests that this improvement in discrimination may have come about via increased attention to, and more complex processing of, the available information, as has already been suggested.

In contrast to the results of Experiment 1, however, in this experiment, procedural accountability did not improve calibration. As noted previously, theoretical analyses imply that one of the avenues by which calibration can be improved is through greater consistency. That might well be what happened here.

With respect to our measure of overall accuracy ($\overline{PS}$), there were no significant differences across conditions. Referring to Eq. (2), it is possible to see why this occurred. Notice that $\overline{PS}$ is a function of both discrimination and calibration. Although discrimination improved under conditions of accountability in the presence of relevant information, calibration was unaffected. In essence, calibration acted as an anchor, holding $\overline{PS}$ down in spite of the opposing pressure from discrimination.

In our final experiment, we turn to the effects of another variable with a history of affecting judgment consistency—the provision of outcome feedback.

EXPERIMENT 3

A great many real-world judgment situations that include a component of accountability also involve some kind of feedback, and in general, feedback is perceived as a useful tool for improving performance. As the judgment literature indicates, however, when feedback is given after each judgment (otherwise known as "outcome feedback"), it generally has a *negative* effect on accuracy (Hammond, Summers, & Deane, 1973; Schmitt, Coyle, & King, 1976). The main cause of this performance decrement appears to be an overresponsiveness to the feedback, sometimes called "error chasing." Subjects fail to appreciate that even the optimal strategy in a probabilistic judgment task will not always produce the correct response, and instead interpret any indication of error as a signal that they should alter their policies. Therefore, even if the optimal strategy is derived at some point in the judgment task, it is likely to be abandoned (as is any other strategy), which in turn would be expected to lower the overall consistency of judgments.

In contrast, we have already seen some evidence that procedural accountability can have the opposite effect on consistency. Specifically, as indicated by the results of Experiment 1, given a relatively rich domain with a variety of cues, procedural accountability resulted in significantly lower scatter than did outcome accountability. While the link between scatter and consistency is not one-to-one, the two are clearly closely related. For example, an individual who gives different responses on two occasions with identical cue values is likely to raise both her scatter and her level of inconsistency.

Since outcome feedback reduces consistency, whereas procedural accountability may enhance it, it is not a trivial task to predict what will happen when both factors are present. At least two outcomes are possible. First, accountability might temper the extent to which people chase errors in response to feedback. Subjects who seek to present themselves in a positive light are unlikely to want to appear indecisive and may attempt to maintain a more consistent policy as a consequence. Furthermore, if a procedural accountability demand causes subjects to give more thought to the optimal procedure from the outset, they may be less inclined to change it mid-stream. In either case, we are led to the same prediction—higher consistency and/or less scatter given the addition of procedural accountability to an outcome-feedback situation than would be seen in the absence of such accountability.

It is also conceivable, however, that accountability could increase the already excessive responsiveness to feedback. That is, because subjects expect to have to justify their procedures, they might be especially averse to any indication that the policy is flawed and therefore continue shifting that policy in an attempt to eliminate all errors. As already indicated, one expected consequence of such frequent (and potentially unwarranted) changes in policy would be an increase in the amount of scatter or noise in an accountable subject's judgments.

A study of Hagafors and Brehmer (1983) suggests at least partial support for this idea that accountability-plus-feedback might lead to increased error chasing. Subjects were asked to predict the level of a fictitious disease from the levels of two fictitious symptoms. Although accountability was found to increase the consistency of judgments, this was true only in the absence of feedback. Subjects who were given trial-by-trial feedback and were held accountable were no more consistent than were feedback subjects in the absence of accountability.

In interpreting this result, however, it is important to note that the way consistency was defined differs somewhat from its natural language meaning. Consistency is generally thought of as behaving in the same way given similar circumstances—behavior which exhibits high test–retest reliability. In Hagafors and Brehmer's study, on the other hand, a linear regression analysis was performed on each individual's judgments, and the correlation between actual judgments and the judgments generated by the linear model was taken as a measure of consistency. The distinction is important for the following reason (cf. Lee & Yates, 1992). Suppose that the combination of accountability and feedback actually encourages subjects to develop more complex strategies—ones that emphasize nonlinear relationships as well as linear ones. If our sole measure of consistency is how well the subject's responses are predicted by a linear model, an accountable subject given feedback will appear to exhibit lowered consistency, when in reality the strategy has simply become more complex (i.e., nonlinear). Although linear models generally have been shown to account for the majority of the variance in judgment tasks, at least *some* subjects in all of these studies have demonstrated strong nonlinearities in their judgment policies (see Brehmer & Brehmer, 1988, for a review).

To address this issue in the present experiment, each subject made judgments under one of three possible conditions—procedural accountability alone, outcome feedback alone, and procedural accountability in the presence of outcome feedback. Unbeknownst to the subjects, a subset of the items were repeated at a later point in the task. The correlation between the initial judgments for these items and the judgments made for the repetitions thus provides a direct measure of consistency.

Based on the results of Hagafors and Brehmer's (1973) study, we would expect any beneficial effects of accountability on consistency to be lost when outcome feedback is introduced as well. If the apparent reduction in consistency in Hagafors and Brehmer's study actually reflects an attempt by subjects to develop more complex (e.g., non-linear) judgment strategies, however, then when consistency is measured by test–retest reliability and scatter rather than as degree of linearity in subjects' judgment policies, accountability may actually prove beneficial rather than harmful.

Method

Subjects. Fifty-eight subjects from the University of Michigan's Psychology Paid Subject Pool were paid $6.00 each in return for their participation in the experiment. Of the initial group, one did not speak English well enough to understand the task and one was under psychiatric care and felt that his medical condition affected his performance on the task. All results discussed below concern only the remaining 56 subjects.

Procedure. Subjects participated individually. The task was essentially the same as that used in Experiment 1, with two exceptions. First, subjects made judgments for 53 jurors rather than 48. The extra five judgments were creating by duplicating Jurors 19–23 in the slots for Jurors 41–45. Subjects were not made aware of this repetition.

Second, instead of comparing no accountability, procedural accountability, and outcome accountability, the three conditions in the present experiment were feedback, procedural accountability, and procedural accountability-plus-feedback. All subjects were interrupted eight times during the experiment, following their judgments for Jurors 6, 15, 16, 26, 29, 36, 48, and 53. These particular jurors were chosen at random, with the constraint that there be one interruption within the first six trials.

At each interruption, feedback (FBK) subjects saw a screen indicating what cue values they had selected for the previous judgment, the judgment they had made, and the juror's true attitude toward suicide. They were allowed to view this screen for as long as they liked before continuing the experiment.

Accountability (ACT) subjects were interrupted after each of the same jurors, and in each case, were also shown their judgment and the cue values they had selected. Unlike the FBK, subjects, however, ACT subjects were not told the juror's true attitude with respect to suicide. Instead, the experimenter merely asked the subject to explain why he or she had chosen the specific cues used, and how those cues were combined to arrive at the judgment rendered.

Finally, accountability-plus-feedback (ACTFBK) subjects were also interrupted after the eight critical jurors and were shown the same information as were FBK subjects. Like ACT subjects, ACTFBK subjects were asked to justify their judgments for each of these cases.

Subjects in the two conditions involving accountability were warned in the instructions that periodically they would be asked to justify judgments, and were asked to sign a consent form allowing the experimenter to tape-record the interviews. For one subject who did not wish to be recorded, the experimenter took notes on the subject's justifications.

FBK subjects were told in the instructions that periodically they would be given feedback. To emphasize the *lack* of accountability in this condition, the experimenter left the room while the subject did the task.

All subjects were told that they were free to adopt

whatever procedure seemed most appropriate to them for making their judgments. The purpose of this statement was to ensure that *all* subjects were equally aware of the need to have some kind of formal procedure for doing the task.

Manipulation check. All subjects responded to a postexperimental manipulation check, in which they were asked whether they thought they would be asked to justify their judgments at various points in the experiment. Two subjects in the accountability conditions indicated that they did not believe they would really be interviewed throughout the task about their procedure. These subjects were therefore excluded from subsequent analyses. Three other subjects in the feedback (no accountability) condition who nevertheless expected to be interviewed about their procedures were also dropped.

Results

Three subjects had mean probability scores almost twice as high (bad) as would be expected due to chance responding. Since these values are considered outliers according to Grubb's test (see Dunn & Clark, 1974), data from these subjects were not included in further analyses.

Median values for the mean probability score and its various components can be seen in Table 4. Since the data violated assumptions of normality, ANOVAs were conducted using the ranks as opposed to the original values, resulting in nonparametric tests (see Conover & Iman, 1981). The only significant differences to emerge concerned scatter, $F(2,46) = 5.32$, $p < .01$. Pairwise t tests using the ranks revealed that subjects in the feedback (FBK) condition exhibited significantly higher scatter than did subjects in either the accountability (ACT) condition, $t(31) = 3.14$, $p < .01$, or subjects in the accountability-plus-feedback (ACTFBK) condition, $t(30) = 2.35$, $p < .05$. Parametric tests yielded the same pattern of results.

FBK subjects averaged more cues viewed per trial ($M = 6.11$) than did subjects in either the ACT condition ($M = 5.29$) or the ACTFBK condition ($M = 5.83$). The difference between FBK and ACT subjects was significant, $t(30) = 2.01$, $p < .05$, while the difference between FBK and ACTFBK was not.

Test–retest reliability (consistency) was measured by computing the correlation between judgments made for Jurors 19–23 and judgments for the second presentation of those jurors (Jurors 41–45). The test–retest value for one subject in the ACT condition was found to be an outlier, and this subject's data were excluded from the test–retest analysis. A one-tailed Student's t test showed that test–retest reliability was significantly higher for subjects in the ACTFBK condition ($M = .75$) than for subjects in the FBK condition ($M = .53$), $t(31) = 1.73$, $p < .05$. Test–retest reliability also appears to be much higher in the ACT condition ($M = .70$) than in the FBK condition ($M = .53$), but this difference was only marginally significant according to a one-tailed test ($t(29) = 1.64$, $p < .06$). Closer examination of the test–retest data, however, reveals that the low reliability in the feedback condition is due almost solely to the exceedingly low (in fact negative) scores of two subjects. Consequently, when nonparametric tests are used, there are no longer any significant differences between conditions. While it is important to note that, according to Grubb's test (Dunn & Clark, 1974), these two scores do not meet the technical definition of outliers, this particular distribution of correlations suggests that a great deal of caution should be used in interpreting the test–retest results.

TABLE 4

Experiment 3: Effects of Procedural Accountability and Feedback on Judgment Accuracy

Measure	Medians		
	Feedback (FBK) condition	Accountability (ACT) condition	Accountability plus feedback (ACTFBK) condition
$\overline{PS}$ (Brier score)[a]	.260	.247	.255
Discrimination index[b]	.045	.050	.047
Calibration index[a]	.063	.045	.058
Slope[b]	.134	.134	.105
Scatter[c]	.069	.032	.042
Bias[c]	.128	.081	.088

[a] Indicates small values are better.
[b] Indicates large values are better.
[c] Indicates zero value is best.

Discussion

As predicted, feedback alone led to significantly higher scatter than did accountability alone. Even more interesting from a practical standpoint, however, is the finding that the addition of procedural accountability was sufficient to significantly reduce these negative effects of feedback. This is shown by the fact that accountability-plus-feedback led to significantly lower scatter than was found given feedback alone.

Based on the results of parametric tests, a similar pattern appears to hold for test–retest reliability. Adding accountability to feedback resulted in significantly higher test–retest reliability than was observed with feedback alone. As was noted above, however, while these findings are congruent with the results concerning scatter, the test–retest results must be interpreted

with a great deal of caution, as they were driven largely by the data from two of the 16 subjects in the feedback condition, and do not hold when nonparametric tests are used. One recommended modification for future work would be to base the test–retest scores on more than five pairs of data per subject in order to obtain a clearer picture of the effect of accountability on this measure.

At this point, we cannot distinguish between two possible explanations for why procedural accountability reduces inconsistency (as indicated by lowered scatter) when added to feedback. The first alternative is that the anticipation of accountability leads people to invest considerable time and energy from the outset in pursuit of the optimal strategy. In essence, there is then a "sunk cost" effect (Thaler, 1980). That is, having "paid" for this strategy in terms of time and effort, there is a reluctance to abandon it, even in the presence of feedback indicating errors.

A second possible explanation for the increased consistency found when feedback subjects are made accountable involves a self-presentational motive. In particular, there is evidence that we value decisiveness—firmness of resolve. For example, Cialdini, Braver, and Lewis (1974) found that subjects in a persuasion task thought the person they were trying to persuade was more intelligent when he or she did not easily yield to influence. If it is true that the appearance of certainty is generally valued, we might expect accountable subjects to be especially averse to constantly changing policies throughout a task. Better to be slightly in error but appear decisive than to seem to flounder while in search of the optimal judgment policy.

It is also perhaps worth speculating about the connection between the present findings and those concerning the distinction between case-by-case outcome feedback and various forms of cognitive feedback and feedforward. With case-by-case outcome feedback, individuals are informed of the correct response after each response they personally generate. In contrast, with cognitive feedback and feedforward, subjects try to improve their judgments on the basis of information about the judgment situation or about their previous performance aggregated over multiple responses. The literature has shown that the latter are generally more effective learning tools than the former (cf. Balzer, Doherty, & O'Connor, 1989). It is conceivable that the greater beneficial effects of cognitive feedback and feedforward rest on the same mechanisms as do procedural accountability effects. For instance, both approaches seem to encourage the subject to examine critically and to seek consciously to improve his or her judgment policy on the basis of considerations directly bearing on that policy. In the case of cognitive feedback and feedforward, however, the pertinent information is provided by an external source, whereas in the case of procedural accountability, it is almost purely introspective.

GENERAL DISCUSSION

Taken together, the results of the experiments reported herein suggest that accountability affects at least two stages of the judgment process. First, there is evidence for accountability effects on the selection and evaluation of available information. In Experiment 1, we saw that procedural accountability led to superior calibration. This type of accountability also produced better discrimination in Experiment 2 (better ability to distinguish occasions when the target event would occur from those when it would not), again suggesting more complex information processing. Finally, procedurally accountable subjects in Experiment 1 were more likely than nonaccountable subjects to select both one of the *most* diagnostic cues and some of the *least* diagnostic cues.

This last result seems puzzling in light of the finding that procedurally accountable subjects also had the best calibration. A plausible explanation, however, is that the cue-use results have more to do with subjects' assumptions about the evaluators' values than the subjects' actual assessments of cue validity.

In addition to influencing information use, our results provide evidence for accountability effects on the consistency with which the subject's judgment policy is executed. First, in Experiment 1, it was shown that outcome-accountable subjects exhibited significantly more scatter than either subjects who were required to account for their procedures or subjects who were not held accountable at all. Recall that scatter reflects the amount of variability in a set of judgments that is unrelated to the occurrence (or nonoccurrence) of the target event. One of the primary mechanisms through which such scatter can be increased is by applying a judgment policy inconsistently.

Further evidence for accountability effects on policy execution was provided by Experiment 3, where the introduction of procedural accountability counteracted the tendency to become less consistent (exhibit more scatter) in response to outcome feedback. Although there was some suggestion that the addition of procedural accountability also ameliorated the detrimental effects of feedback on test–retest reliability, more data must be collected before a conclusive statement regarding this relationship can be made.

Notice that, while both procedural and outcome accountability appear to have had an effect on the consistency with which subjects executed their judgment policies, the effects are in opposing directions. Although procedural accountability in the context of out-

come feedback acted to enhance consistency (reduce scatter), outcome accountability alone increased the scatter. We feel that this last result is especially important in light of the fact that we introduced outcome accountability simply by offering a reward for improved performance. Such incentive schemes are no doubt commonly implemented in real-world situations, under the belief that the more motivated the individual is, the better performance will be. The studies reported here do not allow us to generalize beyond judgment situations in which the individual does not have access to a clear procedure for improving performance quality. Under such circumstances, however, our results suggest that outcome accountability should in fact be avoided if at all possible, even if the alternative is simply no accountability at all.

Given the rather salient failure of procedural accountability to affect *overall* accuracy, it might be argued that this manipulation should be avoided too. Based on our results, however, we disagree. Recall that in Experiment 2, procedurally accountable subjects showed significantly better discrimination than nonaccountable subjects. As we have indicated, good discrimination is one of the most desirable characteristics a judge can have. For example, suppose an individual employs a stock broker who says there is a 99% chance that a stock will go up in all cases where it actually does increase in value and assigns 98% on all occasions when the stock does not go up. Clearly, this broker's performance will be suboptimal in many respects. For instance, her calibration will be extremely poor, which may result in a high (bad) mean probability score as well. As long as her client knows her "code," however, he can in theory earn a substantial sum of money. This is because the broker's discrimination is perfect. As long as the employer invests every time his broker says 99% and holds back every time she says 98%, he should do quite well.

In Experiment 1, procedural accountability was also found to enhance calibration. In some respects, this is a much less striking result, at least from a pragmatic point of view. The reason is that, in many circumstances, calibration (unlike discrimination) is relatively easy to improve. For example, unless every stock under consideration does in fact increase in value, our broker described above inevitably will overestimate the likelihood of the target event. We can repair this problem, however, with an extremely simple instruction, such as telling the broker that, every time she thinks there is a 98% chance that a stock will go up, she should say "zero" instead. (This is because of all occasions when she assigns 98%, the stocks fail to increase in value). This should improve calibration dramatically, without any cost to the other components of accuracy.

Even though it is relatively easy to improve calibration "artificially," improved calibration in the present experiments did in fact seem to reflect a genuine difference in processing (specifically, a more thorough search of memory for relevant quantified judgements and corresponding relative frequencies) under conditions of accountability. This is because we at no time gave subjects the kind of instruction described above, and yet, procedurally accountable subjects still exhibited better calibration. To the extent that enhanced calibration is brought on by a genuine increase in the amount of effort given to learning from past experience, it might be argued that such an improvement is valuable as well.

APPENDIX

Formulas for the Mean Probability Score ($\overline{PS}$) and the components of the Murphy Decomposition

The formula used to compute $\overline{PS}$ is as follows:

$$\overline{PS} = (1/N) \sum_i (f_i - d_i)^2, \tag{3}$$

where f_i is the probability assigned on the i-th judgment, and d_i is the outcome index for that case.

Murphy (1973; see also Yates, 1990, Chapter 3) showed that $\overline{PS}$ can be decomposed into three components of interest, according to the following formula:

$$\overline{PS} = \bar{d}(1 - \bar{d}) - (1/N) \sum_j N_j(\bar{d}_j - \bar{d})^2 + (1/N) \sum_j N_j(f_j - \bar{d}_j)^2 \tag{4}$$

$$= \text{Var}(d) \quad - \quad \text{DI} \quad + \quad \text{CI} \tag{5}$$

where $\bar{d}$ is the base rate or overall proportion of target event occurrences; f_j stands for the alternative judgment categories ($f_1 = 0\%$, $f_2 = 10\%$, etc.); N_j refers to the number of times a particular judgment category is used; and $\bar{d}_j$ corresponds to the proportion of times the target event occurs for a given judgment category (i.e., when a particular judgment category is used).

Formulas for Components of the Covariance Decomposition of the Mean Probability Score

The formula for computing slope is shown below:

$$\text{Slope} = \bar{f}_1 - \bar{f}_0, \tag{6}$$

where $\bar{f}_1$ is simply the average judgment assigned when the target event occurs, and $\bar{f}_0$ is the average judgment assigned when it does not.

The formula for scatter is:

$$\text{Scatter} = \frac{N_1\text{Var}(f_1) + N_0\text{Var}(f_0)}{N_1 + N_0}, \quad (7)$$

where N_1 is the number of judgment occasions on which the target event occurs, N_0 is the number on which it does not occur, $\text{Var}(f_1)$ is the variance of the judgments for the occasions when the target event occurs, and $\text{Var}(f_0)$ is the variance of the judgments when the target event does not occur.

The formula for bias is:

$$\text{Bias} = \bar{f} - \bar{d}, \quad (8)$$

where $\bar{f}$ is the average probability judgment over the sample, and $\bar{d}$ is the sample base rate.

The last components of the covariance decomposition is the outcome index variance (Var(d)) first introduced as part of the Murphy decomposition (see Eq. 4 and 5). This component did not vary across subjects or conditions in the present experiments.

REFERENCES

Adelberg, S., & Batson, C. D. (1978). Accountability and helping: When needs exceed resources. *Journal of Personality and Social Psychology,* **36,** 343–350.

Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior and Human Decision Processes,* **39,** 133–144.

Arkes, H. R., Dawes, R. M., & Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes,* **37,** 93–110.

Ashton, R. H. (1990). Pressure and performance in accounting decision settings: Paradoxical effects of incentives, feedback, and justification. *Studies on Judgment Issues in Accounting and Auditing. Journal of Accounting Research,* **28**(Suppl.), 148–180.

Ashton, R. H. (1992). Effects of justification and a mechanical aid on judgment performance. *Organizational Behavior and Human Decision Processes,* **52,** 292–306.

Balzer, W. K., Doherty, M. E., & O'Connor, R. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin,* **106,** 410–433.

Baumeister, R. F. (1982). A self-presentational view of social phenomena. *Psychological Bulletin,* **91,** 3–26.

Brehmer, A., & Brehmer, B. (1988). What have we learned about human judgment from thirty years of policy capturing? In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgment: The SJT view* (pp. 74–114). Amsterdam: Elsevier.

Brehmer, B., & Kuylenstierna, J. (1978). Task information and performance in probabilistic inference tasks. *Organizational Behavior and Human Performance,* **22,** 445–464.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review,* **78,** 1–3.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology,* **39,** 752–766.

Cialdini, R. B., Braver, S. L., & Lewis, S. K. (1974). Attributional bias and the easily persuaded other. *Journal of Personality and Social Psychology,* **30,** 631–637.

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician,* **35,** 124–129.

Davis, J. A., & Smith, T. W. (1989). *General social surveys.* [Machine-readable data file]. Chicago: National Opinion Research Center (Producer); University of Connecticut, Storrs, The Roper Center for Public Opinion Research (Distributor).

Dunn, O. J., & Clark, V. A. (1974). *Applied statistics: Analysis of variance and regression.* New York: Wiley.

Fandt, P. M., & Ferris, G. R. (1990). The management of information and impressions: When employees behave opportunistically. *Organizational Behavior and Human Decision Processes,* **45,** 140–158.

Ford, J. K., & Weldon, E. (1981). Forewarning and accountability: Effects on memory-based interpersonal judgments. *Personality and Social Psychology Bulletin,* **7,** 264–268.

Hagafors, R., & Brehmer, B. (1983). Does having to justify one's judgments change the nature of the judgment process? *Organizational Behavior and Human Performance,* **31,** 223–232.

Hammond, K. R. (in press). Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice. New York, Oxford University Press.

Hammond, K. R., Summers, D. A., & Deane, D. H. (1973). Negative effects of outcome-feedback in multiple-cue probability learning. *Organizational Behavior and Human Performance,* **9,** 30–34.

Janis, I. L., & Mann, L. (1977). *Decision making.* New York: Free Press.

Keinan, G., Friedland, N., & Ben-Porath, Y. (1987). Decision making under stress: Scanning of alternatives under physical threat. *Acta Psychologica,* **64,** 219–228.

Klimoski, R. J. (1972). The effects of intragroup forces on intergroup conflict resolution. *Organizational Behavior and Human Performance,* **8,** 363–383.

Klimoski, R., & Inks, L. (1990). Accountability forces in performance appraisal. *Organizational Behavior and Human Decision Processes,* **45,** 194–208.

Lee, J.-W., & Yates, J. F. (1992). How quantity judgment changes as the number of cues increases: An analytical framework and review. *Psychological Bulletin,* **112,** 363–377.

Libby, R., & Lipe, M. G. (1992, Autumn). Incentives, effort, and the cognitive processes involved in accounting-related judgments. *Journal of Accounting Research,* 249–273.

McAllister, D. W., Mitchell, T. R., & Beach, L. R. (1979). The contingency model for the selection of decision strategies: An empirical test of the effects of significance, accountability, and reversibility. *Organizational Behavior and Human Performance,* **24,** 228–244.

McGraw, K. O. (1978). The detrimental effects of reward on performance: A literature review and a prediction model. In M. R. Lepper & D. Greene (Eds.), *The hidden costs of reward* (pp. 33–60). Hillsdale, NJ: Erlbaum.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology,* **12,** 595–600.

Nisbett, R. E., Zukier, H., & Lemley, R. (1981). The dilution effect: Nondiagnostic information. *Cognitive Psychology,* **13,** 248–277.

Rothstein, H. G. (1986). The effects of time pressure on judgment in multiple cue probability learning. *Organizational Behavior and Human Decision Processes,* **37,** 83–92.

Rozelle, R. M., & Baxter, J. C. (1981). Influence of role pressures on the perceiver: Judgments of videotaped interviews varying judge

accountability and responsibility. *Journal of Applied Psychology,* **66,** 437–441.

Schmitt, N., Coyle, B. W., & King, L. (1976). Feedback and task predictability as determinants of performance in multiple cue probability learning tasks. *Organizational Behavior and Human Performance,* **16,** 388–402.

Simonson, I., & Nye, P. (1992). The effect of accountability on susceptibility to decision errors. *Organizational Behavior and Human Decision Processes,* **51,** 416–446.

Tetlock, P. E. (1983). Accountability and the perseverance of first impressions. *Social Psychology Quarterly,* **46,** 285–292.

Tetlock, P. E. (1985). Accountability: The neglected social context of judgment and choice. In B. Staw & L. Cummings (Eds.), *Research in organizational behavior.* Greenwich, CT: JAI Press. Vol. 1, pp. 297–332.

Tetlock, P. E., & Boettger, R. (1989). Accountability: A social magnifier of the dilution effect. *Journal of Personality and Social Psychology,* **57,** 388–398.

Received: June 3, 1993

Tetlock, P. E., & Kim, J. I. (1987). Accountability and judgment processes in a personality prediction task. *Journal of Personality and Social Psychology,* **52,** 700–709.

Tetlock, P. E., Skitka, L., & Boettger, R. (1989). Social and cognitive strategies for coping with accountability: Conformity, complexity, and bolstering. *Journal of Personality and Social Psychology,* **57,** 632–640.

Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior and Organization,* **1,** 39–60.

Weldon, E., & Gargano, G. M. (1988). Cognitive loafing: The effects of accountability and shared responsibility on cognitive effort. *Personality and Social Psychology Bulletin,* **14,** 159–171.

Yates, J. F. (1990). *Judgment and decision making.* Englewood Cliffs, NJ: Prentice–Hall.

Young, C. A., & Yates, J. F. (1991). *The effects of stress on judgment and decision making.* Unpublished manuscript, Department of Psychology, University of Michigan, Ann Arbor.