Accountability and Judgment Processes in a Personality Prediction Task

Philip E. Tetlock and Jae Il Kim
University of California, Berkeley

In this experiment, we investigated the impact of accountability—social pressures to justify one's views to others—on cognitive processing in a personality-prediction task. Subjects were presented with the responses of actual test-takers to 16 items from Jackson's Personality Research Form (PRF) and asked to predict how these individuals responded to an additional set of 16 items from the same test. Subjects were assigned to a no-accountability condition (they learned that all of their responses would be anonymous), a preexposure-accountability condition (they learned of the need to justify their responses before seeing the test-takers' PRF responses), and a postexposure-accountability condition (they learned of the need to justify their responses after seeing the test-takers' PRF responses). Preexposure-accountability subjects reported more integratively complex impressions of test-takers, made more accurate behavioral predictions, and reported more appropriate levels of confidence in their predictions than did either no-accountability or postexposure-accountability subjects. We conclude by considering possible psychological mediators of these effects as well as the broader theoretical implications of the findings for the development of contingency models of judgment and choice.

A burgeoning research literature has emerged on the cognitive structures and processing rules that people use to form impressions of others (cf. Fiske & Taylor, 1984; Markus & Zajonc, 1985; Nisbett & Ross, 1980). Relatively little attention has been directed, however, to the communicative functions of social cognition. People often expect to discuss and explain their impressions of others with important social audiences. Participation in this "communication game" (Higgins & McCann, 1984) can profoundly affect the information to which people attend, the complexity and types of rules they use to process this information, and the conclusions they ultimately draw from it (see Forgas, 1982; Higgins & McCann, 1984; Tetlock, 1985a).

In this experiment, we explore the impact of accountability—social pressures to justify one's views to others—on how people process information in a personality-prediction task. In this task, subjects are presented with the responses that actual people made to 16 items from the Achievement and Affiliation scales of the Personality Research Form (PRF; Jackson, 1967, 1971). Subjects were given an opportunity to form an impression of each test-taker and to report their impressions in a free-response format. Subjects were also asked to predict how each test-taker responded to an additional set of 16 items selected from the same PRF scales and to rate their confidence in each prediction.

How well is it reasonable to expect subjects to perform this task? Most research on social judgment gives cause for pessimism. This body of work suggests that subjects will fall prey to a variety of serious errors and biases: They will be too quick to draw conclusions about the personalities of test-takers (Jones, 1979), they will be too slow to revise their impressions in response to contradictory evidence (Asch, 1946; Nisbett & Ross, 1980), they will discount evaluatively inconsistent information (Asch, 1946; Schroder, Driver, & Streufert, 1967), they will base their predictions on only a small—and probably unrepresentative—sampling of the behavioral evidence (Abelson & Levi, 1985), and they will be too confident in the correctness of their predictions (Fischhoff, 1982; Koriatt, Lichtenstein, & Fischhoff, 1980). People, in short, are depicted as cognitive misers (Fiske & Taylor, 1984) whose preference for simple, easy-to-execute heuristics (the principle of least effort) renders them vulnerable to a long list of inferential shortcomings.

There are, however, good theoretical and empirical reasons for not fully accepting this portrait of the social-information processor. A number of experiments indicate that people can, under certain conditions, be motivated to process information in more complex, self-critical, and effort-demanding ways (Borgida & Howard-Pitney, 1983; Harkness, DeBono, & Borgida, 1985; McAllister, Mitchell, & Beach, 1979; Rozelle & Baxter, 1981; Showers & Cantor, 1985; Tetlock, 1983a, 1983b, 1985b; Zajonc, 1960). Of special interest here is research that suggests that holding subjects accountable for their judgments and decisions can sometimes have this effect. Evidence has now accumulated that social demands for accountability can markedly affect the cognitive strategies that people use in making judgments and decisions. Accountable decision makers are more likely than unaccountable ones to use cognitively complex rules in choosing among response options (McAllister et al., 1979; Tetlock, 1983a), to be more aware of the determinants of their judgments (Cvetkovich, 1978; Hagafors & Brehmner, 1983), to display greater consistency and stability of judgment (Hagafors & Brehmner, 1983), to process persuasive messages in detail rather than to rely on their general evaluation of the message's source (Chaiken, 1980), and to be more discriminating and responsive to evidence in evaluating others (Rozelle & Baxter, 1981; Tetlock, 1983b, 1985b).

This research was supported by National Institutes of Mental Health Grant RO1 39942 to Philip E. Tetlock.

Correspondence concerning this article should be addressed to Philip E. Tetlock, Institute of Personality Assessment and Research, 3657 Tolman Hall, University of California, Berkeley, California 94720.
The study reported here extends this line of "debiasing" research by assessing the impact of accountability on a set of theoretically interrelated dependent variables in a personality-prediction task: (a) the integrative complexity of the initial impressions that subjects form of the test-takers whose responses to the PRF subjects are asked to predict; (b) the accuracy of subjects' predictions; and (c) the appropriateness of the confidence ratings that subjects attach to their predictions. In the following paragraphs, we sketch our major hypotheses concerning how accountability should affect each of these dependent variables and the interrelations among them.

On the basis of past research, we expected subjects who felt accountable for their views to form more integratively complex impressions of the test-takers. What exactly does this mean? Integrative complexity has traditionally been defined in terms of two cognitive structural variables: the conceptual differentiation and integration of information (Schroeder et al., 1967; Streufert & Streufert, 1978; Suedfeld, 1983; Tetlock, 1979, 1981, 1985b). Differentiation is thought to be a function of the number of alternative interpretations or perspectives that a person draws upon in understanding a problem. For instance, in the personality-prediction task used here, subjects could form undifferentiated impressions of test-takers by focusing on only one of the two major facets of personality measured by the PRF scales (achievement or affiliation) and by offering unqualified assessments of the standing of test-takers on that trait dimension (e.g., this is an ambitious or this is a sociable person). Subjects could form highly differentiated impressions by clearly distinguishing between the two facets of personality measured by the PRF scales and by recognizing the need for situational qualifications on trait assessments (e.g., this person is introverted, but does enjoy certain types of social events). Integration is a function of the number of causal or conceptual connections that a person identifies between (among) differentiated perspectives. (Differentiation is thus a necessary but not sufficient condition for integration.) A low-integration response would merely list the perceived attributes of the test-takers. A high-integration response would attempt to specify how the differentiated attributes—in various combinations or interactions—could produce particular behavioral outcomes.

Experimental work suggests that accountability can, under certain conditions, motivate integratively complex thought. For instance, Tetlock (1983a) found that subjects who expected to justify policy positions to an audience with unknown views were much more likely to think about those policy issues in integratively complex ways than were subjects who felt accountable or subjects who knew the views of the audience to whom they were accountable (and, thus, had the "cognitively lazy" option of conformity). Tetlock (1983b, 1985b) found that accountable subjects were less likely to fall prey to primacy effects in an impression-formation task and to the fundamental attribution error in Jones's essay-attribution paradigm. Tetlock attempted to explain these findings by invoking the notion of preemptive self-criticism. He argued that accountability demands place subjects in a self-critical mental set in which they actively try to anticipate the objections or counterarguments that might be raised to their positions. As a result, subjects pay close attention to the evidence, are careful to refrain from judgment on the basis of incomplete information, and make persistent efforts to integrate contradictory or inconsistent information into their overall impression of the evidence.

If accountability promotes integratively complex processing of information about test-takers, then it may also be reasonable to expect accountability to improve the accuracy of subjects' predictions of test-takers' responses and the appropriateness of the confidence ratings that subjects express in those predictions. The test-takers chosen as stimulus persons in this study all received moderate-range scores on both the Affiliation and Achievement scales of the PRF, which means the test-takers agreed with approximately half of the high-achievement and high-affiliation items and disagreed with the other half. The test-takers, in short, are complex stimuli; simple trait generalizations are not likely to provide subjects with prediction formulas that exceed chance levels of accuracy. There is no guarantee, of course, that subjects who form integratively complex (multidimensional, qualified) impressions of test-takers will perform better than chance. Subjects may form inaccurate complex impressions or may fail to use all the information at their disposal to generate optimal predictions (cf. Abelson & Levi, 1985). Nonetheless, there are good logical grounds for supposing that some degree of integratively complex processing of test information is at least a necessary, if not sufficient, condition for consistently performing at above-chance levels on the personality-prediction task. Moreover, there are some empirical grounds for drawing a similar conclusion. Although previous studies are not entirely consistent on this issue, judges characterized by high cognitive complexity have been found to be better performers in inferential tasks that require role taking and empathic skills (cf. Hale & Delia, 1976; Press, Crockett, & Delia, 1975).

Accountability may also improve the appropriateness of the confidence ratings that subjects attach to their predictions via the intervening variable of integrative complexity. Past work suggests that subjects will tend to be excessively confident in the correctness of their personality-test predictions. This tendency toward overconfidence is one of the most well-documented biases in the judgment and choice literature—a bias that holds up across an impressive range of experimental tasks and subject populations (Lichtenstein & Fischhoff, 1977, 1980). There is also relatively little dispute over the normative characterization of overconfidence as a bias. In a typical study, for instance, subjects are asked to assess the probability that various (general knowledge) statements are true. The researchers measure appropriateness of confidence by comparing these probability estimates with the observed relative frequencies of being correct (hit rates). A subject is well calibrated if over the long term, for all answers assigned a given probability, the proportion correct equals the probability assigned. Evidence reveals that subjects usually assign confidence ratings that are substantially greater than the hit rates indicate to be justified. Observed hit rates of .60, for instance, tend to be associated with assessed probabilities of .70; hit rates of .75 are often assigned probabilities as high as .90 or even 1.0 (absolute certainty).

Koriat et al. (1980) and Ross and Lepper (1985) have identified a key information-processing mechanism underlying the overconfidence bias: the tendency of people to focus primarily on belief-supportive evidence in assessing the likely correctness of their own knowledge. These investigators have found that
overconfidence effects can be substantially reduced by short-circuiting this process, for example, through instructions to think of possible reasons why one’s preferred answers might be wrong or to think of causal scenarios that could lead to outcomes different from those expected.

Accountability may improve the calibration of confidence ratings in a similar way by encouraging subjects to engage in preemptive self-criticism in which they carefully consider potential flaws or limitations in the impressions that they have formed of test-takers. Subjects who become more self-critical should report more integratively complex impressions of test-takers and more appropriate degrees of confidence in their predictions. The increased realism of confidence ratings, moreover, should be partly or even wholly mediated by the increased integrative complexity. High-complexity subjects should be much less susceptible than low-complexity subjects to the biased-scanning effect hypothesized to underlie overconfidence in judgment.

Serious conceptual problems arise, however, in testing this seemingly straightforward hypothesis. Perhaps the most difficult problem is that accountability may eliminate overconfidence not via increased integrative complexity of thought, but via the activation of a conservative response bias. Accountable subjects may simply become unwilling to express much confidence in any of their predictions so as to avoid the embarrassing possibility of their high-confidence answers being proved wrong. Accountability, in this view, may have a general depressing effect on confidence ratings.

One method of disentangling the impression-management and the information-processing effects of accountability is to vary the timing of the accountability manipulation. Subjects in the present experiment learned of the need to justify their impressions of the test-takers either prior to or only after exposure to the test responses. If the effects of accountability are mediated entirely via impression management (the increased reluctance of accountable subjects to express confidence in general), then these effects should occur with approximately equal strength among both pre- and postexposure-accountability subjects. If the effects of accountability are mediated entirely via the integrative complexity of the initial impressions that subjects form of the test-takers, then these effects should occur only among preexposure-accountability subjects. Finally, it is possible that both sets of processes occur. If this is the case, both pre- and postexposure-accountability will improve the calibration of confidence ratings, with greater improvement likely to occur in the preexposure-accountability condition.

The foregoing analysis still fails, however, to address an interesting theoretical possibility. It assumes that any improvement in confidence calibration among postexposure-accountability subjects should automatically be attributed to impression management of response thresholds. This assumption is not warranted. Although postexposure accountability may not be as effective as preexposure accountability in motivating self-critical, integratively complex thought (postexposure-accountability subjects may have ignored, forgotten, or assimilated important items of contradictory evidence to which they now lack access), postexposure accountability may still have some power to stimulate careful reflection on the judgment task. We need, therefore, an additional method to distinguish improvements in confidence that result from strategic response shifts from those that result from vigilant information processing. The solution is provided by the Brier method of partitioning the covariation between confidence and accuracy-of-judgment scores. (See Method section for more detail.) The Brier partition permits identification of three statistically distinct determinants of the confidence-accuracy relation: knowledge, calibration, and resolution. Of special interest here is the distinction between calibration (the ability of subjects to assign probabilistic confidence ratings to predictions that on average correspond to the likelihood of those predictions being correct) and resolution (the ability of subjects to assign predictions to confidence categories such that the proportion of correct predictions within different categories is maximally different). Subjects who are simply reluctant to report confidence in their predictions in general may improve their calibration scores, but only at the cost of degrading their resolution scores. By contrast, subjects who think about the test-takers in more integratively complex ways may be able to improve their calibration scores without a concomitant decline in their resolution scores (indeed, they may even be able to improve both their calibration and resolution scores). The key point here is that integratively complex processors should become selectively, not indiscriminately, less confident. Integratively complex judges will report lower confidence mainly on those test items for which reasonably compelling arguments exist in favor of both true and false predictions.

To summarize, this study examines the impact of accountability (pre- and postexposure) on how subjects think in a personality-prediction task. The three dependent variables of special interest are (a) the integrative complexity of subjects’ impressions of the test-takers, (b) the accuracy of subjects’ predictions, and (c) the appropriateness of subjects’ confidence in their predictions.

Method

Subjects

Undergraduate students (N = 60) at the University of California participated in the study in return for course credit or money ($3.50). Twenty subjects were randomly assigned to each of the three experimental conditions: no accountability, preexposure accountability, and postexposure accountability. Subjects were tested individually.

Procedure

The experimenter began each session by informing subjects that they would be participating in a study of the person-perception process, that is, how people form impressions of others from various types of information. Subjects were then told that they would be given the responses that three actual persons had made to 16 items drawn from a personality test—Jackson's (1967, 1971) Personality Research Form (PRF; see Table 1 for a list of these items). Eight of the items had been drawn from the Affiliation scale of the PRF and 8 from the Achievement scale. Subjects were asked to read the responses of each test-taker to these 16 items and then to write a brief personality sketch of that individual (at least three sentences). Subjects were given 12 min to perform this task for each test-taker. Order of presentation of test-taking protocols was randomized. After subjects had completed their written assessments of all three test-takers, they were asked to predict the responses of each test-taker to an additional set of 16 test items. Again, 8 items came from the
Table 1
True-False Test Items Presented to Subjects in the Impression-Formation Phase of the Study

<table>
<thead>
<tr>
<th>Item no.</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>I don’t really have fun at large parties.</td>
</tr>
<tr>
<td>2.</td>
<td>I seldom set standards which are difficult for me to reach.</td>
</tr>
<tr>
<td>3.</td>
<td>I do not let my work get in the way of what I want to do.</td>
</tr>
<tr>
<td>4.</td>
<td>I will not be satisfied until I am the best in my field of work.</td>
</tr>
<tr>
<td>5.</td>
<td>I am quite independent of the people I know.</td>
</tr>
<tr>
<td>6.</td>
<td>I would work just as hard whether or not I had to earn a living.</td>
</tr>
<tr>
<td>7.</td>
<td>I go out of my way to meet people.</td>
</tr>
<tr>
<td>8.</td>
<td>I have rarely done extra fighting in connection with my work.</td>
</tr>
<tr>
<td>9.</td>
<td>People consider me to be quite friendly.</td>
</tr>
<tr>
<td>10.</td>
<td>I would not be very good at a job which required me to meet people all day long.</td>
</tr>
<tr>
<td>11.</td>
<td>I choose hobbies that I can share with other people.</td>
</tr>
<tr>
<td>12.</td>
<td>I truly enjoy myself at social functions.</td>
</tr>
<tr>
<td>13.</td>
<td>I enjoy difficult work.</td>
</tr>
<tr>
<td>14.</td>
<td>I try to work just hard enough to get by.</td>
</tr>
<tr>
<td>15.</td>
<td>I seldom put out extra effort to make friends.</td>
</tr>
<tr>
<td>16.</td>
<td>People should be more involved with their work.</td>
</tr>
</tbody>
</table>

Affiliation and 8 from the Achievement scales of the PRF (see Table 2 for a list of these items). Subjects were also asked to rate their confidence in their predictions on subjective probability scales that ranged from 50% (completely unsure whether my prediction is correct) to 100% (completely confident that my prediction is correct). Subjects were instructed to make the confidence ratings of predictions immediately after each prediction. Subjects were allowed to refer to their written assessments of the test-takers, but not to the original 16 items on which those written assessments were based.

Accountability Manipulation

Subjects in the no-accountability condition were assured prior to reading the test-takers’ PRF responses that their impressions of the test-takers would be completely confidential and not traceable to them. Even the experimenter, it was emphasized, would not know how they had responded. Subjects in the preexposure-accountability condition learned before reading any of the test-takers’ PRF responses that the researchers were interested in conducting detailed interviews with subjects to explore “the types of information that people use to form impressions of others.” These subjects expected to participate in such an interview after they completed the experimental tasks and were asked to sign a form granting permission to audiotape the interview “for future data analysis purposes.” Subjects in the postexposure-accountability condition received exactly the same instructions, but only after they had read all three test-takers’ PRF responses and had provided written personality sketches of the test-takers.

Dependent Variables

There were three major dependent variables. Integrative complexity. Two trained coders, unaware of the hypotheses to be tested, rated the integrative complexity of the free-response personality sketches that subjects provided for each of the three test-takers. The coding system used for this purpose has demonstrated reliability and construct validity. It has been successfully applied in a number of research contexts to test hypotheses concerning both personality and situational determinants of complexity of information processing (see Schroder et al., 1967; Streufert & Streufert, 1978; Suedfeld, 1983; Suedfeld & Tetlock, 1977; Tetlock, 1979, 1981, 1983a, 1984, 1986; Tetlock, Hannum & Micheletti, 1984).

As noted earlier, integrative complexity is defined in terms of both conceptual differentiation and integration. Differentiation refers to the number of alternative interpretations or perspectives that a person considers in analyzing an event or issue. Coders are trained to recognize that undifferentiated responses can be generated by simple evaluative rules. For instance, a person might take an undifferentiated view of a test-taker by focusing on only one major theme running through the test responses (e.g., a person who loves to take charge, to get his way, to win arguments). A more differentiated statement would recognize either contradictory evidence on the test-taker’s standing on a trait dimension (e.g., this person is dominant in some situations but quite accommodating in others) or the difficulty of capturing the complexity of the test-taker’s personality with a single trait label (e.g., this person needs the company of others but also needs to feel in control of most social situations). Integration refers to the development of complex connections among differentiated characteristics. The complexity of integrations depends on whether the person perceives the differentiated characteristics as existing in isolation (low integration), in simple interactions (moderate integration), or in multiple, contingent patterns (high integration). For example, high-integration statements might include explicit statements on how two or more trait dimensions interact to shape behavior (e.g., this person needs both to dominate and to be in the company of others. When the two needs come into conflict—because dominant behavior might provoke an unpleasant scene—he or she will probably back off unless the issue is very important to him or her).

Integrative complexity scores ranged from 1 to 7 (1 = low differentiation and integration; 3 = moderate/high differentiation, low integration; 5 = moderate/high differentiation, moderate integration; 7 = high differentiation and high integration). Scores of 2, 4, and 6 represent transition levels that are assigned when there is evidence of implicit

Table 2
True-False Test Items Presented in Personality-Prediction Phase of the Study

<table>
<thead>
<tr>
<th>Item no.</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>17.</td>
<td>As a child I worked a long time for some of the things I needed.</td>
</tr>
<tr>
<td>18.</td>
<td>I try to be in the company of friends as much as possible.</td>
</tr>
<tr>
<td>19.</td>
<td>Sometimes I have to make a real effort to be sociable.</td>
</tr>
<tr>
<td>20.</td>
<td>I don’t mind working while other people are having fun.</td>
</tr>
<tr>
<td>21.</td>
<td>Often I would rather be alone than with a group of friends.</td>
</tr>
<tr>
<td>22.</td>
<td>It doesn’t really matter to me whether or not I become one of the best in my field.</td>
</tr>
<tr>
<td>23.</td>
<td>I spend a lot of time visiting friends.</td>
</tr>
<tr>
<td>24.</td>
<td>My goal is to do at least a little bit more than anyone else has done before.</td>
</tr>
<tr>
<td>25.</td>
<td>I don’t spend much of my time talking with people I see every day.</td>
</tr>
<tr>
<td>26.</td>
<td>My friendships are many.</td>
</tr>
<tr>
<td>27.</td>
<td>I am not really very certain what I want to do or how to go about doing it.</td>
</tr>
<tr>
<td>28.</td>
<td>In my work I seldom do more than is necessary.</td>
</tr>
<tr>
<td>29.</td>
<td>I trust my friends completely.</td>
</tr>
<tr>
<td>30.</td>
<td>When I see someone I know from a distance, I don’t go out of my way to say hello.</td>
</tr>
<tr>
<td>31.</td>
<td>I often set goals that are very difficult to reach.</td>
</tr>
<tr>
<td>32.</td>
<td>People seldom think of me as a hard worker.</td>
</tr>
</tbody>
</table>
differentiation (e.g., use of qualifiers, recognition of uncertainty, information seeking) or implicit integration (e.g., allusions to interactions between different attributes of the test-taker's personality). High inter-rater agreement existed between coders ($r = .87$). Disagreements were resolved by relying on the judgment of a third coder.

**Predictive accuracy.** Assessing the accuracy of person perception is a complex undertaking (cf. Cline, 1964; Cronbach, 1955). For instance, when judges predict the responses of a number of individuals to a number of test items, it is useful to decompose accuracy scores into three logically and statistically distinct sources of variation: (a) the ability of judges to predict the average responses (across items) of given individuals (in Cronbach's terms, differential elevation); (b) the ability of judges to predict the average responses (across individuals) to given items (in Cronbach's terms, stereotype accuracy); and (c) the ability of judges to predict the differences among test-takers on each item considered separately (in Cronbach's terms, differential accuracy). Each of these components was estimated by different terms in the $3 \times 3 \times 16$ (test-takers $\times$ items) mixed-design analysis of variance (ANOVA). (The dependent variable—incorrect or correct prediction—took on the values of 0 and 1, respectively.) The main effect for test-takers yields an estimate of differential elevation; the main effect for items yields an estimate of stereotype accuracy; the Test-Takers $\times$ Items interaction yields an estimate of differential accuracy.

**Appropriateness of confidence ratings.** Assessing the appropriateness of confidence ratings also involves a complex data analytic process (for an overview of methods for assessing appropriateness of confidence, see Lichtenstein & Fischhoff, 1977). The measures chosen for the current study are the calibration and resolution components of the Brier score proposed by Murphy (1973) and a simple index of over- and underconfidence.

Murphy (1973) has demonstrated that the Brier score can be partitioned into three additive parts. For two-alternative tasks (in which predictions can be scored as either right or wrong), the computational formula takes the following form:

$$\frac{1}{N} \sum_{i=1}^{N} (r_i - c_i)^2 = \bar{c}(1 - \bar{c}) + \frac{1}{N} \sum_{t=1}^{T} n_t(r_t - c_t)^2 - \frac{1}{N} \sum_{t=1}^{T} n_t(\bar{c}_t - \bar{c})^2. \quad (1)$$

The key terms in this formula are as follows: $N$, the total number of probability assessments made by a given judge (48); $r$, the probability assigned to a particular prediction ($i = 1 \ldots 48$); $c$, which is 1 if the prediction was correct, 0 otherwise; $\bar{c}$, which is the proportion of the total predictions that were correct; $T$, the number of probability categories that judges used to express their confidence in their predictions (six categories: $0.5, 0.6, 0.7, 0.8, 0.9, 1.0$); $n_t$, the number of probability assessments that fell into the $t$th category; $r_t$, the numerical value of the probabilities in the $t$th category ($0.5, 0.6, 0.7, 0.8, 0.9, 1.0$); and $\bar{c}_t$, the proportion of the probability assessments in the $t$th category that were attached to the correct alternative.

The term on the left of Equation 1 is the total Brier score, which is a general measure of the goodness of fit between the probability assessments and accuracy scores. This term can be decomposed into the three terms on the right of the equation. The first term on the right is a measure of knowledge. If a judge was always correct in his or her predictions, this component would have a value of zero. The second term on the right measures calibration. It is the weighted average of the mean square differences between the proportion of predictions correct in each probability category and the probability value of that category. A judge would receive the ideal calibration score of zero if he or she consistently got 50% of the predictions assigned to the .5 category correct, 60% of the predictions assigned to the .6 category correct, and so forth. The third term is resolution. It is an index of the judge's ability to sort the predictions into probability categories such that the proportions of correct answers are maximally different. It is the variance of the proportion of correct predictions across the $T$ categories; the larger, the better.

As Lichtenstein & Fischhoff (1977) have noted, the Brier procedure is a proper scoring rule that rewards people for responding with their own beliefs. People who try to respond strategically in order to improve their scores encounter a trade-off between calibration and resolution. For instance, our judges might be able to improve their calibration scores through simple strategies (such as reporting only confidence levels of .5 and .6) but only at the cost of degrading their resolution scores.

Because the calibration score, a squared index, is insensitive to whether a judge is over- or underconfident, we also computed a measure of over/underconfidence: the mean confidence rating of a judge minus the proportion of the judge's predictions that were correct. A positive score reflects overconfidence and a negative score, underconfidence.

**Results**

We subjected each of the major dependent variables to a repeated measures ANOVA.\(^1\) Figure 1 presents the mean integrative complexity scores as a function of accountability and the stimulus person being judged. An ANOVA of these scores revealed a main effect for accountability, $F(2, 57) = 4.20, p < .02$. Planned contrasts indicated that as predicted, preexposure-accountability subjects reported more integratively complex impressions of test-takers ($M = 2.33$) than did both no-accountability subjects ($M = 1.90$) and postexposure-accountability subjects ($M = 1.81$). The respective $F$ ratios for these comparisons were 6.62 and 5.46 (both $p$s $< .05$). The no-accountability

\(^1\) The $F$ tests reported here may be biased, in either a positive or negative direction, because the repeated-measures independent variables of test-taker and test items are more appropriately thought of as random, not fixed, effects (cf. Clark, 1973). We conducted quasi-$F$ ratios for all effects for all dependent variables to check on whether this potential bias could have been responsible for any of the effects identified by conventional ANOVA procedures. The quasi-$F$ test results led to conclusions identical to those of the standard $F$ tests.
Figure 2. Mean accuracy-of-prediction scores as a function of accountability and stimulus person.

and postexposure-accountability conditions did not differ from each other ($F < 1$). No other effects were significant.

Figure 2 presents the mean accuracy-of-predictions scores. The ANOVA of these scores was unusually complex; it took the form of a 3 (accountability) x 3 (stimulus person) x 16 (items) design. The key results can, however, be stated fairly simply. Preexposure-accountability subjects made more accurate predictions ($M = 69.8$%) than did both no-accountability subjects ($M = 59.1$%), $F(1, 57) = 9.63, p < .01$, and postexposure-accountability subjects ($M = 62.3$%), $F(1, 57) = 8.42, p < .01$. Although this improvement was a general one (it was not limited to a particular test-taker, a particular set of test items, or a particular set of Test-Taker x Item combinations), it is necessary to append some qualifications to the main effect of preexposure accountability on judgmental performance. As Figure 2 reveals, certain test-takers were easier to predict than others, $F(2, 114) = 11.29, p < .01$ (what Cronbach called a differential elevation effect). Similarly, certain items were easier to predict than other items, $F(15, 855) = 2.85, p < .01$ (what Cronbach called a stereotype accuracy effect), and certain Test-Taker x Item combinations were easier to predict than other combinations, $F(30, 1710) = 3.16, p < .01$ (what Cronbach called a differential accuracy effect). The accountability manipulation interacted with these two latter effects. Preexposure accountability improved accuracy to a greater degree on certain items than on others, $F(30, 855) = 1.78, p < .01$. (Specifically, Items 17, 25, 26, 27, and 30 in Table 2 showed the greatest improvement.) Preexposure accountability also improved accuracy to a greater degree on certain Test-Taker x Item combinations than on others. The Test-Taker x Item interaction was significantly greater in the preexposure-accountability condition than in the no-accountability and postexposure-accountability conditions, $F(60, 1710) = 1.68, p < .01$. To summarize in Cronbach's (1955) terms, preexposure accountability was especially likely to improve stereotype accuracy (accuracy in predicting responses to particular items) and differential accuracy (accuracy in predicting particular Test-Taker x Item combinations).

We used analysis of covariance (ANCOVA) to test the possibility that the superior predictive performance of preexposure-accountability subjects was at least partly attributable to the tendency of these subjects to form more integratively complex impressions of test-takers. The ANCOVA revealed a substantial positive relation between integrative complexity (the covariate) and accuracy of performance, $F(1, 56) = 6.62, p < .05$. The earlier observed differences in predictive accuracy among experimental conditions continued, however, to be significant even after controlling for the integrative complexity variable. Preexposure-accountability subjects (adjusted $M = 66.7$%) were still more accurate than both no-accountability subjects (adjusted $M = 61.2$%) and postexposure-accountability subjects (adjusted $M = 62.4$%). The respective $F$ ratios for these comparisons were 5.78 and 5.12 (both $p < .05$).

Table 3 presents the mean appropriateness-of-confidence ratings that subjects offered for their predictions of test-takers' responses. These results indicated that the well-replicated overconfidence phenomenon did indeed emerge in the no-accountability and postexposure-accountability conditions of the experiment ($M$s = 13.6 and 9.6, respectively). But the overconfidence effect was substantially weaker in the preexposure-accountability condition ($M = 4.0$). The comparisons between the no-accountability and preexposure-accountability conditions, $F(1, 56) = 13.48, p < .01$, and between the preexposure- and postexposure-accountability conditions, $F(1, 56) = 9.02, p < .01$, were both statistically significant. An ANCOVA suggested that the more realistic confidence ratings of the preexposure-accountability subjects were partly attributable to the tendency of those subjects to form more integratively complex impressions of test-takers, $F(1, 56) = 9.89, p < .01$. Controlling for integrative complexity did not, however, eliminate the significant difference between the preexposure-accountability and no-accountability conditions, $F(1, 56) = 7.71, p < .01$, or the significant difference between the preexposure- and postexposure-accountability conditions, $F(1, 56) = 6.02, p < .05$.

Figure 3 presents the mean calibration and resolution scores derived from the Brier partitioning of the coarvarion between subjects' confidence ratings and accuracy scores. An ANOVA of calibration scores revealed a pattern strikingly consistent with

Table 3
Confidence Dependent Variables as a Function of Accountability

<table>
<thead>
<tr>
<th>Condition</th>
<th>Overconfidence scores$^a$</th>
<th>Calibration scores$^b$</th>
<th>Resolution scores$^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>No accountability</td>
<td>13.6</td>
<td>69.7</td>
<td>35.7</td>
</tr>
<tr>
<td>Preexposure accountability</td>
<td>4.0</td>
<td>28.4</td>
<td>41.9</td>
</tr>
<tr>
<td>Postexposure accountability</td>
<td>9.6</td>
<td>52.0</td>
<td>26.9</td>
</tr>
</tbody>
</table>

$^a$ Higher scores indicate that the subjects assigned lower subjective probability ratings to their answers that exceeded the objective probability of their answers being correct.

$^b$ Lower scores indicate superior calibration (ability to assign subjective probability ratings to answers that correspond to the objective probabilities of the answers being correct).

$^c$ Higher scores indicate superior resolution (ability to sort predictions into probability categories such that the proportions of correct answers are maximalistically different).
those for the dependent variables of integrative complexity and predictive accuracy. Preexposure-accountability subjects reported better calibrated confidence ratings than did no-accountability subjects, $F(1, 56) = 7.98, p < .01$, and postexposure-accountability subjects, $F(1, 56) = 6.43, p < .01$ (see also Figure 4). No other significant effects existed. We used an ANCOVA to explore whether the improved calibration scores among preexposure-accountability subjects were at least partly due to the more integratively complex styles of thinking of these subjects. This analysis indicated that subjects who formed integratively complex impressions of test-takers did indeed report better calibrated confidence ratings, $F(1, 56) = 6.98, p < .05$. Controlling for the integrative complexity variable, however, did not completely eliminate the tendency for preexposure-accountability subjects (adjusted $M = 32.9$) to be better calibrated than no-accountability subjects (adjusted $M = 66.5$) and postexposure-accountability subjects (adjusted $M = 52.0$). The respective F ratios for these comparisons were 6.42 and 4.12 (both $p < .05$).

An ANOVA of resolution scores revealed no significant effects, $F(2, 57) = 2.43$. This result strongly suggests that preexposure accountability did not improve calibration at the expense of resolution. If anything, there was a trend in the opposite direction. Preexposure-accountability subjects had better resolution scores ($M = 41.92$) than did no-accountability subjects ($M = 35.74$) and postexposure-accountability subjects ($M = 26.91$).

**Discussion**

The data add to the growing evidence that social pressures for accountability can, under certain conditions, motivate people to become more vigilant, thorough, and self-critical information processors. Preexposure accountability led subjects (a) to form more integratively complex impressions of stimulus persons; (b) to make more accurate predictions concerning the responses of these stimulus persons to personality-test items; and (c) to make better calibrated confidence ratings of their predictions. There is also considerable circumstantial evidence from the ANCOVAs that the effects of preexposure accountability on predictive accuracy and confidence calibration were in part mediated by the tendency of subjects in the preexposure-accountability condition to form more integratively complex initial impressions of the test-takers.2

The theoretical implications of these data are potentially far-reaching. There appears to be a growing recognition within cognitive social psychology that although the cognitive-miser metaphor highlights important facts (all other things being equal, people rely heavily on simple, low-effort heuristics), the metaphor may also obscure other equally important facts. People often use different processing rules when the stakes are high as opposed to low (cf. Borgida & Howard-Pitney, 1983; Chaiken, 1980; Erber & Fiske, 1984; Harkness et al., 1985; Petty, Cacioppo, & Goldman, 1981; Taylor, 1975; Tetlock, 1983a, 1983b). As Showers and Cantor (1985, p. 277) noted in their recent review of the literature, involvement can motivate people to show flexibility in adjusting their interpretations of events in response to changing information, to take control of their cognitive processes, to recognize multiple ways of encoding the same event.

2 The key phrase here is “in part.” The effects of preexposure accountability on perceptual accuracy and calibration were not completely eliminated when the role of integrative complexity was taken into account. This result suggests the operation of other mediating processes. One possibility worthy of investigation is that preexposure accountability promoted more elaborate or in-depth processing of the initial test-taker response information, thus making that information more available when subjects were asked to make predictions on later items to assess their confidence in those predictions. This hypothesis could be tested by incorporating measures of recall in future research designs.
or outcome, and even to rethink basic beliefs and values. To paraphrase Jenkins’s (1981) contextualist critique of cognitive psychology, the appropriate question may not be “What kind of machine is the human information processor?” but, rather, “What kinds of machines do people become when confronted with particular types of tasks in particular types of decision environments?” (see also Payne, 1982). People may be cognitive misers much of the time, but not all of the time, and it is theoretically and practically important to understand the situational (and individual-difference) qualifications that need to be attached to the cognitive-miser portrait of the social perceiver.

Viewed from this broad theoretical perspective, our results point to the need for a more complex—and less pessimistic—appraisal of the perceivers as an intuitive scientist. The time may have come for contingency theories of judgment and choice that specify how situational, task, and role variables interact to shape social-information processing (cf. McAllister et al., 1979; Payne, 1982). Contingency-theory approaches, however, are still relatively undeveloped. It is not sufficient, for example, simply to hypothesize that the higher the stakes, the more likely people are to think in complex, effort-demanding ways and the more resistant people will be to well-known errors and biases. That hypothesis would, in key respects, be wrong. Sometimes increasing the stakes in decision problems makes matters worse and encourages belief perseverance, overconfidence, and rigid, dichotomous thinking (Janis & Mann, 1977). Sometimes increasing the stakes has no effect at all (Fischhoff, 1982). And, sometimes, increasing the stakes leads to improved judgmental performance. Contingency theories need to specify when it is reasonable to expect these very different empirical outcomes (cf. Kruglanski & Freund, 1983).

The current data pose a number of puzzles for contingency theories of judgment and choice. For instance, postexposure-accountability subjects were under the same social demands as preexposure-accountability subjects, and presumably, they were every bit as motivated to make correct predictions and express appropriate levels of confidence. Yet they failed to do so. Only preexposure accountability improved judgmental performance. This result, moreover, should not be viewed as an isolated phenomenon. Tetlock (1983b, 1985b) has found that the timing of accountability manipulations is a crucial factor in determining the magnitude of primacy effects in an impression-formation task and of overattribution effects in a Jones’s essay-attrition paradigm. In those studies, as in the present one, preexposure accountability improved judgmental performance; postexposure accountability had no effect. Taken together, these findings point to two conclusions:

1. Accountability motivates subjects to process social information in more analytic and complex ways that can substantially reduce judgmental biases such as belief perseverance, the fundamental attribution error, and overconfidence.

2. There appear to be cognitive constraints on the power of accountability as a debiasing manipulation. Accountability is much more effective in preventing than in reversing judgmental biases. Once subjects have assimilated or integrated information into their impressions of a person or event, they have a hard time reinterpreting that information. Accountability has a marked impact on the initial impression-formation process (it places subjects in a vigilant mental set that confers some protection from certain inferential biases), but it has little impact after the initial processing has occurred (accountability cannot undo biased processing at an earlier analytic stage).

Why should the effects of accountability (or perhaps debiasing manipulations in general) be limited to the initial impression-formation stages? The key to the answer may lie in the cognitive processes that underlie the judgmental biases under investigation. It has been hypothesized that such biases are often the product of highly overlearned or automatic-processing routines of which people are largely or even entirely unaware (cf. Bargh, 1984). To use Nisbett and Wilson’s (1977) distinction, people may have access only to the products of these very rapid processes, not to the processes themselves. Accountability may be effective in debiasing social judgment only when people learn before automatic-processing routines are activated that they are accountable. Preexposure-accountability subjects have an opportunity to adopt a more self-conscious, controlled approach to how they will analyze the social information available to them. This self-consciousness disrupts automatic processing. By contrast, postexposure-accountability subjects, who have access only to the products of automatic processing, are in no position to correct the inferences they have already drawn. They also have little motivation to do so. The products of automatic processing provide plausible answers to the questions that are asked of them.

The foregoing argument suggests that contingency theories need to specify the limits on the flexibility of judgment and choice processes. Equally critical, contingency theorists need to clarify distinctions among independent variables that are often loosely classified as motivators of cognitive effort or work (e.g., personal involvement, decision importance, transmission—receiver set, monetary incentives, reversibility of outcomes, accountability). Sometimes these independent variables have quite similar effects (e.g., McAllister et al., 1979) and sometimes, quite dissimilar ones. For instance, the effects of accountability on overconfidence reported here are very different from the effects of decision importance on overconfidence reported by Sieber (1974). She found that increasing the importance of the decision-making task exacerbated the overconfidence effect—a result that she interpreted in terms of Hull—Spence theory (the tendency for arousal to increase the likelihood of dominant responses). Obviously, comparing two studies that differ in as many ways as do the present study and Sieber’s (1974) is a tricky undertaking. Perhaps Sieber’s manipulation (course grades were thought to be at stake) was more powerful than our manipulation and produced levels of arousal that interfered with, rather than facilitated, self-critical, integratively complex thought. Aroused is related in a curvilinear fashion to integrative complexity, with moderate levels most conducive to complex functioning (Schröder et al., 1967; Streufert & Streufert, 1978). Our accountability manipulation may have created an optimal level of arousal and Sieber’s, a suboptimal level. It is also possible, however, that the two manipulations differ in significant qualitative respects. More important than the general arousing properties of accountability may be the specific cognitive coping responses activated by the need to justify one’s views. Accountability may serve as a signal to subjects to take the role of the other toward their own mental processes and to give serious weight to the possibility that their preferred answers
are wrong. Accountability, in this view, does not simply motivate thought; it functions as a social brake on judgmental biases that occur in our less reflective moments.\(^3\)

This latter theoretical interpretation is bolstered by the results of the only two other studies in the research literature that have reported success in reducing overconfidence effects. Koriat et al. (1980) and Ross and Lepper (1985) found that encouraging subjects to engage in self-critical thought (think of reasons why you might be wrong) or differentiated thought (think of causal scenarios that could lead to different outcomes) led to more realistic levels of confidence. Our results are highly compatible with the work of these investigators and with what is already known of the effects of accountability on complexity of reasoning in other contexts (e.g., Tetlock, 1983a). The key mediating variable may not be the amount but, rather, the type of thought that the debiasing manipulation evokes.

Finally, a few words of caution are in order concerning the normative implications of our findings. Contingency theories of judgment and choice often posit that people should adjust the complexity of their reasoning processes as a function of the importance of the problem. It is assumed, moreover, that people should make these adjustments in a quasi-rational manner (cf. Abelson & Levi, 1985; Payne, 1982, for reviews). People should weigh the costs and benefits of different cognitive strategies and choose the strategy that strikes the right balance between the desire to make a correct decision and the desire to minimize mental effort. Although superficially reasonable, normative analyses of this sort leave major questions unanswered. It is not clear, for instance, (a) how one should go about quantifying mental effort; (b) whether complex thinkers are generally better at predicting the behavior of others (there is a need to determine the replicability of the complexity–accuracy relation across a much broader range of stimulus persons and judgment tasks); (c) whether most people perceive complex cognitive strategies as a means of increasing the probability of arriving at correct assessments of others; and (d) whether complex thinkers are able to capitalize on their more differentiated understandings of others to design behavioral strategies that are more likely to lead to desired outcomes. Unfortunately, it is not possible to provide even tentative answers to these questions. We still know remarkably little about the conditions under which people are well advised to rely on simple versus complex judgment strategies. Moreover, what knowledge we have is not altogether consistent: Sometimes simple strategies do quite well, and other times, quite poorly (Abelson & Levi, 1985; Kleinmuntz & Kleinmuntz, 1981). Normative contingency theories of judgment and choice— theories that tell us how we should think in different situations—appear to be a long way off in the future.

References


\(^3\) A comparison of work on cognitive tuning (Zajonc, 1960) and work on accountability provides another illustration of the need to distinguish carefully between independent variables that, on intuitive grounds, one might suppose have similar effects. For instance, placing subjects in a transmission set (leading them to expect to communicate their opinions to others) might not seem very different from making subjects accountable (leading them to expect to justify their opinions to others). The two manipulations can, however, have very different effects. Subjects in transmission sets tend to form more extreme dispositional attributions than subjects in a no-set control condition (cf. Harvey, Har-


Received April 14, 1986
Revision received August 4, 1986