The Consistency and Accuracy of Holistic Judgment

Clinical Decision Making with a Minimally Complex Task

John Ruscio
The College of New Jersey

Antonee R. Stern
Elizabethtown College

Reaching holistic judgments requires an ability to combine multiple sources of information in an interactive—rather than additive—manner, a cognitively challenging process unsupported by research in the judge-modeling tradition. In three experiments, we more directly tested individuals' ability to make holistic judgments by explicitly showing them how to do so. Participants were provided with full specifications for a judgment task and given specific instructions on how to generate accurate predictions. Relative to a comparison condition in which two cues were additively related to a criterion, holistic judgments based on two interacting cues were less consistent and accurate. These results were replicated and extended across educational levels, academic disciplines, and clinical experience. The inability of participants to make holistic judgments in a task of minimal complexity has implications for the practical utility of supplementing or replacing holistic judgment with statistical prediction rules in clinical practice.

In everyday activities as well as in professional capacities, people must make an astonishing number of decisions. Many of the most consequential decisions can be made in either an actuarial manner (i.e., statistical or mechanical prediction) or on the basis of unaided human judgment (i.e., intuitive or clinical prediction).[1] The statistical approach involves an algorithmic, mechanical combination of information designed to maximize accuracy. Applying the clinical approach, a human judge evaluates available information and arrives at a decision. For more than half a century, research evidence has consistently supported the benefits of statistical over clinical prediction (for recent reviews and a meta-analysis, see Dawes, Faust, & Meehl, 1989; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Grove & Meehl, 1996). In addition to the advantages of mechanical methods of data *collection*, evidence has long shown that mechanical methods of data *combination*, such as regression equations or actuarial tables, yield more valid decisions than do less systematic approaches, such as relying on unaided human judgment (Sawyer, 1966). However, despite overwhelming empirical support, statistical prediction rules (SPRs) nonetheless remain scarce in applied mental health settings (Meehl, 1986; Swets, Dawes, & Monahan, 2000). Practitioners clearly still prefer to use their heads rather than formulas when making important decisions.

One of the reasons underlying professionals' adherence to clinical prediction methods may be that they feel that a SPR cannot consider all relevant material in the complex manner that they believe is often warranted in clinical situations. Rather, practitioners may think that their judgment processes far exceed the complexity of a SPR, that they "take into account" a wide range of relevant information and integrate it in sophisticated ways (Summers, Taliaferro, & Fletcher, 1970). In other words,

1. This method does not necessarily involve a professional clinician; rather, the term is broadly applied whenever a human judge forecasts outcomes. For the sake of consistency with the literature on prediction, we will use the term *clinical* to describe any decision making based on unaided human judgment regardless of whether it actually involves a professional clinician.

Financial support for summer work on this research and the payment of graduate students in Study 3 was provided through the Provost's Office at Elizabethtown College; Ron McAllister's support is gratefully acknowledged. We would also like to thank Jennifer Mills for her assistance in collecting Study 3 data, Stefanie Skoniecki for also collecting Study 3 data and providing detailed comments on a draft of this paper, and Ayelet Meron Ruscio for constructive criticism and editorial assistance.

Correspondence concerning this article should be addressed to John Ruscio, Department of Psychology, The College of New Jersey, P. O. Box 7718, Ewing, New Jersey, 08628-0718; e-mail: ruscio@tcnj.edu.

clinicians may subscribe to a more "holistic" approach to judgment. Despite its popularity with practitioners, this style of judgment has had a long and checkered history in the research literature (Ruscio, 2003). Early studies on the clinical-statistical prediction controversy led to speculation about the conditions under which clinical judgment might prevail over statistical methods. Although unaided human judgment was acknowledged to be inferior to SPRs for making simple decisions, it was hypothesized to be superior when the information under consideration was more complex. More specifically, task characteristics involving configural relationships between variables—such as nonlinear relationships or interactions among variables—were expected to favor the highly trained and experienced professional (Meehl, 1954, 1967).

It is commonly asserted that the seasoned practitioner does not rely upon individual factors; rather, he or she considers the client as a complex whole. Thus, it is argued that holistic judgment cannot be reduced to additive main effects, for such effects would fail to adequately contextualize the available information. Rather, to reason holistically one must consider each piece of information in light of *all available information*. In essence, holistic judgments reflect a process of thinking and reasoning that is based on interaction effects. If the meaning ascribed to each piece of information truly depends upon all available information—without consideration of individual factors—then judgment requires an interaction term of the highest order. For example, with a half-dozen pieces of information available, clinical judgment would need to include a six-way interaction term.

The question, then, is just how well human judgment can handle information that actually does involve such interactions. There are at least two ways to address this question. First, one can look to studies that use paramorphic models of human judgment to search for evidence of interactive cue utilization among experts. Such studies have failed to uncover the interactive processes that are allegedly taking place (Goldberg, 1991; Slovic & Lichtenstein, 1971; Stewart, 1988; Wiggins & Hoffman, 1968). Instead, additive linear models are generally found to reproduce judgments as well as more complex models. Although there is ample evidence that clinicians engage in nonlinear processing (e.g., Ganzach, 1995; Goldberg, 1968; Slovic & Lichtenstein, 1971; Wills & Moore, 1994)—particularly with cognitively straightforward strategies that involve conjunctive or disjunctive rules (Dawes, 1964; Einhorn, 1971)—clinicians do not appear to draw on interactions to any substantial degree.

There are, however, several limitations to the evidence provided by these judge-modeling studies. First, because models of clinicians' judgments do not necessarily capture actual cognitive processes (Hoffman, 1960), the absence of interactive terms in a model does not rule out the possibility that information was processed in a holistic manner. Second, the replicability of interactive terms across such models has seldom been evaluated. Though interactions rarely account for even a few percent of the variance in judgments, even this could be due to the capitalization on chance occurring when dozens of interaction terms are entered into judge-model regression equations. Third, the search for interactions has focused almost exclusively on the reproducibility (consistency) of clinicians' judgments, leaving their accuracy—the more important criterion—largely unexplored. Thus, the modeling approach to studying clinical judgment yields suggestive, but not conclusive, evidence of the failure of human judges to successfully utilize cue interactions.

An alternative way to address this question is to explicitly teach judges about a specific cue interaction and then test their ability to reach consistent and accurate holistic judgments on that basis. Using this approach, we hypothesized that whereas participants would be capable of understanding the nature of a particular interaction effect and become reasonably confident in their ability to perform the operations required to generate predictions from interactive cues, when put to the test their holistic judgments would be relatively inconsistent and inaccurate even under conditions of minimal task complexity. Because clinical judgment is only fair to poor at the relatively simple tasks that have previously been studied—for example, making decisions based on a small number of valid predictors that are related to the outcome in an additive, linear manner—we contend it is unlikely that judgment will function well with tasks that are even more complex—that is, making decisions based on a large number of variables of differing validity that are related to the outcome in nonlinear or interactive ways (Dawes, 1994, 2001; Faust, 1984).

To experimentally test the consistency and accuracy of holistic judgments, we constructed a judgment task in which two cues unambiguously interacted to predict a criterion. Because configural relationships are extremely challenging for people to detect (Hammond & Summers, 1965; Summers, Summers, & Karkau, 1969), we fully and carefully explained the interactive nature of the cue-criterion relationship in the experimental task. As a benchmark for comparison, participants also made judgments in a task that was identical in all regards save for the cue-criterion relationship, which was additive rather than interactive in nature. In three studies, individuals of varying educational levels, academic disci-

plines, and degrees of clinical training and experience made judgments under one or both of these experimental conditions. Our central hypothesis was that judgments would be more consistent and more accurate when made on the basis of two cues that were additively, rather than interactively, related to a criterion.

We also explored the potentially distracting role of irrelevant information in decision making. Previous research has suggested that human judges often find it difficult to ignore available but irrelevant information, thereby diluting the quality of judgments (Nisbett, Zukier, & Lemley, 1981; Ruscio, 2000). To extend these findings to the context of holistic reasoning, irrelevant information was added to the two relevant cues during half of one of the judgment tasks. Moreover, to determine whether the addition versus removal of irrelevant information midway through the task had a differential effect on judgments, order was not only counterbalanced but also included as a factor in this experimental design.

Finally, research consistently finds that individuals feel more confident in their judgments than their accuracy levels warrant (Dawes, 1994; Dawes et al., 1989; Faust & Ziskin, 1988; Oskamp, 1965; Ruscio, 2000). This is problematic because inflated confidence can be mistakenly perceived as a gauge of accuracy. If confidence outstrips the efficacy of a decision-making strategy, professionals may be less likely to seek and adopt more useful approaches. In the present studies, we did not explore the accuracy-confidence relationship within a calibration framework. Rather, we assessed participants' confidence levels using a subjective rating scale to determine whether they were sensitive to any differences in accuracy that emerged across experimental conditions.

STUDY 1

Method

Design. Participants were randomly assigned to either an additive main effect (subsequently referred to as "linear") or an interactive cue condition. All participants made two series of judgments, one using two relevant cues and one using six cues, including the same two relevant cues plus four irrelevant cues. The order of these series was counterbalanced across participants and treated as a factor in the design. Thus, a three-way mixed factorial design (cue condition × relevancy × order) was employed, with relevancy serving as a repeated measure.

Participants. Sixty-two undergraduate students at Elizabethtown College participated in the study for partial fulfillment of the experimental participation requirement in their introductory psychology course. Demographic data were not collected, but most students at this college are between 17 and 22 years old and approximately two-thirds of them are women; younger students and women are disproportionately likely to enroll in the introductory psychology course.

Materials. To test participants' performance rather than learning, the judgment task involved variables that were familiar, easily interpretable, and whose distributions and interrelationships were fully described. Participants were asked to predict the responsiveness to treatment of 60 hypothetical children diagnosed with Attention-Deficit/Hyperactivity Disorder (ADHD) given a fictitious new drug called Attevil. Scores on two relevant cues, attention deficit and hyperactivity, were provided for each child. The normal T score distribution of all variables—cues and criterion—was graphically displayed and described in the following way:

> For each child, his or her scores will range along a scale that has an average of 50, as shown in the graph. . . . Scores near 50 are most common, with fewer and fewer children scoring at levels that depart from this average. Scores above 70 or 80 are extremely rare, as are scores below 20 or 30. This is true of *all* variables, including *Attention Deficit*, *Hyperactivity*, and the one that you'll be predicting, *Responsiveness to Treatment*.

For one series of 30 children, participants were told that attention deficit and hyperactivity scores were the only relevant information on which to base their judgments, and that they would therefore receive information on only these variables. In the other series, however, participants were also provided with some irrelevant information:

> For the 30 children to follow, you will be given scores on six factors, all of which were measured using a reliable and valid clinical assessment tool that has been standardized for use with children. Although only Attention Deficit and Hyperactivity are relevant to your predictions, many practitioners like to have access to additional, contextual information to better understand the whole child. You will therefore be given both types of information.

Explicit definitions of all available cues and the criterion were provided, and participants were informed that cues were independent of one another (see appendix).

The relation between the two relevant cues and the criterion was manipulated across linear and interactive cue conditions. In the linear condition, criterion scores

were constructed based on an additive main effects model that was fully explained to participants:

> [Attevil] is more effective for children with higher scores on Attention Deficit, Hyperactivity, or both. That is, the higher a child scores on these variables, the more good Attevil is likely to do. The best predictions of responsiveness to Attevil would therefore be made by considering how high both factors are. The simple graph [showing two additive main effects; see Figure 1, top panel] summarizes this. . . . As you can see, predictions should be higher with increasing levels of Attention Deficit, and predictions should also be higher with increasing levels of Hyperactivity.

In the interactive condition, criterion scores were constructed based on a model with no main effects but a crossover interaction that was also carefully explained:

> [Attevil] is more effective for children with limited impairments, and not effective for those with more severe impairments. That is, for children with high scores on *either* Attention Deficit *or* Hyperactivity, *but not both*, Attevil is likely to be quite helpful. Those with only minor impairments on both factors, as well as those with severe impairments on both factors, will benefit little if at all from Attevil. The best predictions of responsiveness to Attevil would therefore be made by considering the combination of both factors. The simple graph [showing an interaction effect; see Figure 1, bottom panel] summarizes this. . . . As you can see, predictions should be higher when *either* Attention Deficit *or* Hyperactivity is high, moderate when one or both factors are intermediate, and low when both factors are high.

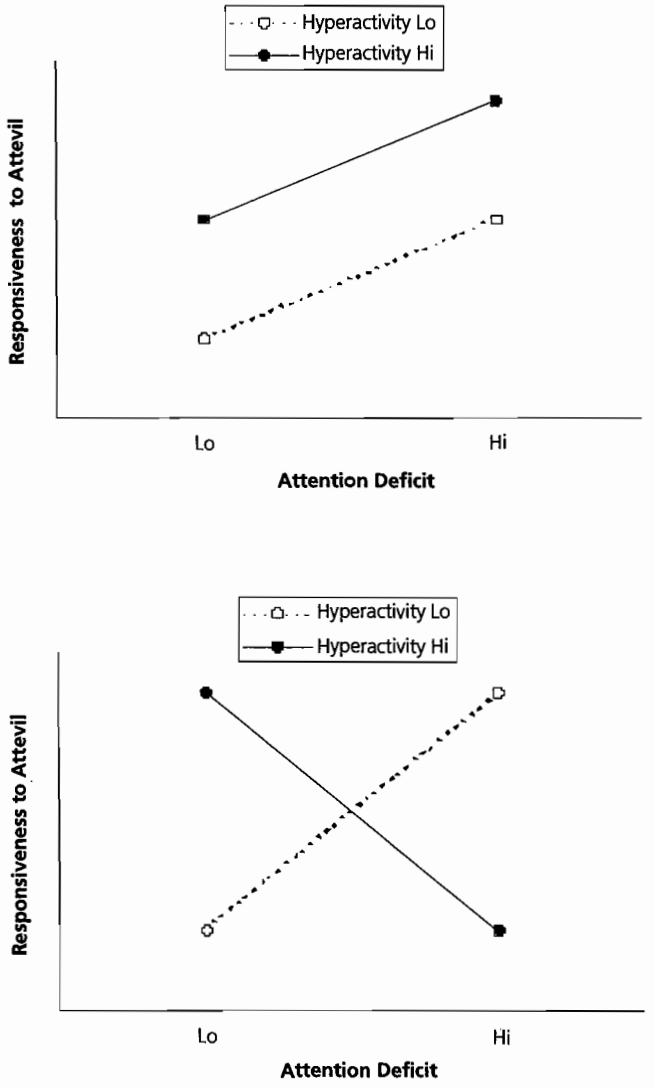

Figure 1. Graphs included in the instructions for Study 1. The top panel was used to illustrate the linear cue condition (two additive main effects), whereas the bottom panel was used to illustrate the interactive cue condition (a crossover interaction effect).

Data were constructed such that the criterion was highly and equally predictable in both cue conditions. Two cues were constructed as vectors of random normal deviates, and their intercorrelation was checked to ensure that it was negligible (< .05). These two cues were combined—along with an additional random normal deviate that served as an error component—to form a criterion with the desired degree of predictability (see below). In the linear condition, an additive main effects model (cue 1 + cue 2 + error) was used to construct the criterion, whereas in the interactive condition, a multiplicative model (cue 1 × cue 2 + error) was used to construct the criterion. Across all 60 cases of information, regression equations derived from the two relevant cues predicted the criterion at $R = .73$, $R^2 = .53$ in the linear condition and $R = .74$, $R^2 = .54$ in the interactive condition.[2] Additional vectors of random normal deviates served as irrelevant cues, which were negligibly correlated with either criterion, $|r| < .08$, $p \geq .539$ for each. The 60 cases were randomly split into two series of 30 cases each.

Procedure. Participants arrived for the experiment alone or in small groups of up to 10 people, but completed the experiment individually. Each participant was assigned to either the linear or interactive cue condition through random distribution of the instruction sets. The nature of the judgment task was explained in detail both verbally and in writing, and the experimenter answered all participants' questions before asking them to begin the judgment task. The order of relevancy conditions was counterbalanced by

2. In the interactive condition, the cues were entered as main effects on the first step of a regression analysis, and neither predicted the criterion: $\beta = .01$ and $.04$, $p = .952$ and $.778$, respectively. It was the interaction term, entered on the second step of the regression analysis, that predicted the criterion, $\beta = 6.73$, $p < .001$.

randomly distributing packets of cue information for the first series of 30 children. Participants worked at their own pace, making 30 judgments and an overall rating of confidence reflecting the perceived accuracy of all 30 judgments (from 1, "low confidence," to 7, "high confidence"). When finished, they returned this packet of information and received the next, with the experimenter once again answering any questions before they continued working. After completing the second series of judgments and its associated confidence rating, participants were debriefed, thanked for their time, and excused.

Results

Analyses were conducted to examine differences in the consistency and accuracy of judgments across experimental conditions. For each participant, consistency was calculated as the *R* value of the judge-model regression equation that best predicted his or her judgments from the available cues (Cooksey, 1996). Consistency was calculated separately for each series of 30 judgments. To help normalize the distributions of consistency values, they were transformed using Fisher's *r'*. All analyses were performed on these transformed values, though means and standard deviations are reported in the more familiar correlational units.

Consistency was analyzed using a 2 (cue condition: linear vs. interactive) × 2 (relevancy condition: two relevant cues vs. two relevant plus four irrelevant cues) × 2 (task order: removal vs. addition of irrelevant cues) mixed model ANOVA with repeated measures on relevancy. As anticipated, there was a strong main effect for cue conditions, $F(1, 58) = 20.20$, $p < .001$, $\eta^2 = .26$, such that judgments were more consistent in the linear condition ($M = .79$, $SD = .20$) than in the interactive condition ($M = .65$, $SD = .15$). There was also a main effect for relevancy, $F(1, 58) = 15.95$, $p < .001$, $\eta^2 = .22$, such that judgments were more consistent when just two relevant cues were available ($M = .77$, $SD = .17$) than when two relevant plus four irrelevant cues were provided ($M = .66$, $SD = .25$). There was no main effect or interaction involving the order of the relevancy conditions, all F's < 1.

Each participant's judgments were correlated with the appropriate criterion measure—additive or interactive—to determine the accuracy of prediction (Cooksey, 1996). These accuracy scores were also transformed to Fisher's *r'* for analysis; *M*s and *SD*s are once again presented in correlational units. Judgmental consistency was an excellent predictor of accuracy, $r(60) = .76$, $p < .001$. Accuracy scores were subjected to the same ANOVA model described for consistency scores. There were no effects of relevancy or order on accuracy, Fs < 1. There was, however, a strong main effect for cue conditions, $F(1, 58) = 32.97$, $p < .001$, $\eta^2 = .36$, such that judgments were more accurate in the linear condition ($M = .50$, $SD = .27$) than in the interactive condition ($M = .18$, $SD = .18$). Another way of expressing this large difference in accuracy is to compare it to chance-level prediction of the criterion. Participants were scored as surpassing chance if their accuracy scores for both series of judgments exceeded the critical value for a correlation coefficient with $df = 28$ at $\alpha = .05$, which is .36. By this measure, 22 of 31 participants (71%) in the linear condition surpassed chance-level accuracy, whereas only 3 of 31 (10%) did so in the interactive condition, $\chi^2(1, N = 62) = 24.20$, $p < .001$, $\phi = .63$.

Next, participants' confidence ratings were analyzed to determine whether the large differences in consistency and accuracy revealed across cue conditions were accompanied by appropriate differences in subjective confidence.[3] Although confidence was correlated with consistency, $r(58) = .31$, $p = .016$, and marginally correlated with accuracy, $r(58) = .22$, $p = .097$, there was no difference between the confidence ratings made in the linear condition ($M = 3.78$, $SD = 1.05$) and those made in the interactive condition ($M = 3.68$, $SD = 1.21$), $t(58) = .34$, $p = .735$.

Finally, we checked whether any of the above results could be attributed to failure to understand the task, inattention, poor task motivation, or related factors. Judgmental consistency is arguably the best measure of whether a participant understood and was adequately engaged in the task. Five participants in the linear condition and seven participants in the interactive condition demonstrated poor consistency—operationalized as $R < .40$, which corresponded to a qualitative break in the distribution of consistency scores—for at least one series of judgments. Removal of these 12 individuals from the sample had no substantive effect on any of the results reported above.

Discussion

As expected, participants had considerable difficulty integrating information that was in fact interrelated interactively. Even under the relatively simple conditions of this experiment (e.g., written directions, a graphical summary,

3. One participant in each cue condition did not make one or both confidence ratings.

and verbal explanations specifically geared to the participants' immediate task), judgments made from two interactive cues were notably less consistent and less accurate than those made from two additive cues. Moreover, although confidence was weakly related to consistency and to accuracy, it was nonetheless equivalent across cue conditions. In other words, not only were participants comparatively poor at reaching sound holistic judgments, but their confidence ratings reflect no awareness of this.

The other experimental variable of interest, informational relevancy, produced mixed results. The availability of irrelevant information negatively influenced the consistency of judgments but not their accuracy. Given this finding, it was not surprising that the order of tasks—either adding or removing irrelevant information—had no discernible influence on consistency or accuracy. Perhaps instructions emphasizing the relevance of two cues and the irrelevance of the others enabled participants to selectively utilize the relevant information.

STUDY 2

In addition to replicating the results of Study 1, we wished to test whether individuals with widely varying educational experiences differ in their ability to make holistic judgments. Because we recruited participants from introductory psychology classes for Study 1, their educational levels did not differ much. Moreover, we had not recorded participants' academic disciplines. Thus, any performance differences attributable to educational experiences could not have been detected in Study 1. To determine whether certain educational experiences may facilitate holistic judgment, we recruited participants from a broader population for this second investigation. Consistent with our finding that undergraduate students performed relatively poorly at a holistic judgment task, we did not expect this performance to differ across individuals' educational levels or academic disciplines. In order to maximize the opportunity for reliable and valid judgments and thus provide a particularly "risky" test of our primary hypothesis (Popper, 1959), we simplified the procedure by eliminating irrelevant information and requiring just one series of judgments to make the task as straightforward as possible.

Method

Design. Participants were randomly assigned to either a linear or an interactive cue condition, which formed the independent variable. Educational level (freshman/sophomore, junior/senior, faculty) and academic discipline (professional studies, social science, physical science, arts and humanities) were also recorded as subject variables.

Participants. One hundred twenty-one individuals at Elizabethtown College participated in this experiment. Eighty-eight of these (72%) were undergraduates, many of whom participated in exchange for experimental credit in an introductory psychology course. There were 47 freshmen and sophomores and 41 juniors and seniors. The remaining 33 participants (28%) were faculty members. Participants represented a broad range of academic disciplines: 51 (42%) were from professional studies (e.g., business, education, occupational therapy), 33 (27%) were from social sciences (e.g., psychology, sociology, anthropology, social work), 23 (19%) were from physical sciences (e.g., biology, chemistry, physics, mathematics), and 10 (8%) were from arts and humanities (e.g., English, philosophy, history, music); 4 (3%) were undergraduates whose field of study was as yet undecided. Faculty members were fairly evenly spread across these four disciplinary domains (*n*s ranged from 7 to 10).

Materials and Procedure. The materials and procedure differed from those of Study 1 in only three ways. First, no irrelevant information was provided: All participants were given only the two relevant cues of attention deficit and hyperactivity. Second, the task instructions were strengthened and further clarified. In the linear cue condition, the original instructions were followed by this additional note: "The best predictions of responsiveness to Attevil would therefore be made by considering these two factors independently of one another; considering the combination of these two factors is unnecessary for successful prediction." In the interactive cue condition, the added note read: "The best predictions of responsiveness to Attevil would therefore be made by considering the combination of both factors; considering either factor alone is insufficient for successful prediction." Because some informal pilot testing of these new instructions suggested that they were even more clear to participants than the graphs used in Study 1 to depict additive main effects or a crossover interaction, the graphs were dropped. Third, only one series of judgments was made, along with one confidence rating, for the first series of 30 cases of information from Study 1.

Results

Participants' consistency and accuracy scores were calculated in the same manner as in Study 1 and again transformed using Fisher's r' for analysis, with *M*s and *SD*s presented in correlational units. Analyses tested the extent to which consistency and accuracy varied as a function of cue conditions, educational level, and academic discipline. Because some cell sizes were prohibitively small when crossing the three educational levels with the four academic disciplines, a series of four two-way ANOVAs was conducted. Cue conditions were crossed separately with educational level and with academic discipline, once using consistency as the dependent variable and once using accuracy.

In the analysis of cue conditions × education for consistency, there was a main effect for cue conditions, $F(1, 115) = 28.54$, $p < .001$, $\eta^2 = .20$, such that judgments were more consistent in the linear condition ($M = .80$, $SD = .23$) than in the interactive condition ($M = .67$, $SD = .17$). Although there was no main effect for education, $F(2, 115) < 1$, there was a marginal interaction, $F(2, 115) = 2.66$, $p = .074$, $\eta^2 = .04$, that showed faculty to be particularly consistent in the linear condition (see Figure 2, top graph).

In the analysis of cue conditions × education for accuracy, there was a main effect for cue conditions, $F(1, 115) = 27.83$, $p < .001$, $\eta^2 = .20$, such that judgments were more accurate in the linear condition ($M = .48$, $SD = .29$) than in the interactive condition ($M = .27$, $SD = .22$). There was also a marginal main effect for education, $F(2, 115) = 2.96$, $p = .056$, $\eta^2 = .05$. Post hoc comparisons of means using Tukey's HSD with $\alpha = .05$ revealed that judgments made by faculty ($M = .47$, $SD = .24$) were more accurate than those made by freshmen and sophomores ($M = .31$, $SD = .31$); judgments made by juniors and seniors ($M = .38$, $SD = .27$) were of intermediate accuracy and did not differ from the other two groups. There was no interaction of cue conditions and education, $F(2, 115) = 1.56$, $p = .214$. However, because difference in the ability to make holistic judgments was the focal point of this research, additional tests of simple effects for education were conducted within cue conditions. In the linear condition, there was an effect for education, $F(2, 115) = 3.79$, $p = .025$, whereas in the interactive condition, there was no effect for education, $F(2, 115) = 1.01$, $p = .367$. As shown in Figure 2 (middle graph), the difference in accuracy across educational levels is primarily attributable to the proficiency of faculty members in the linear condition, paralleling the results obtained for consistency.

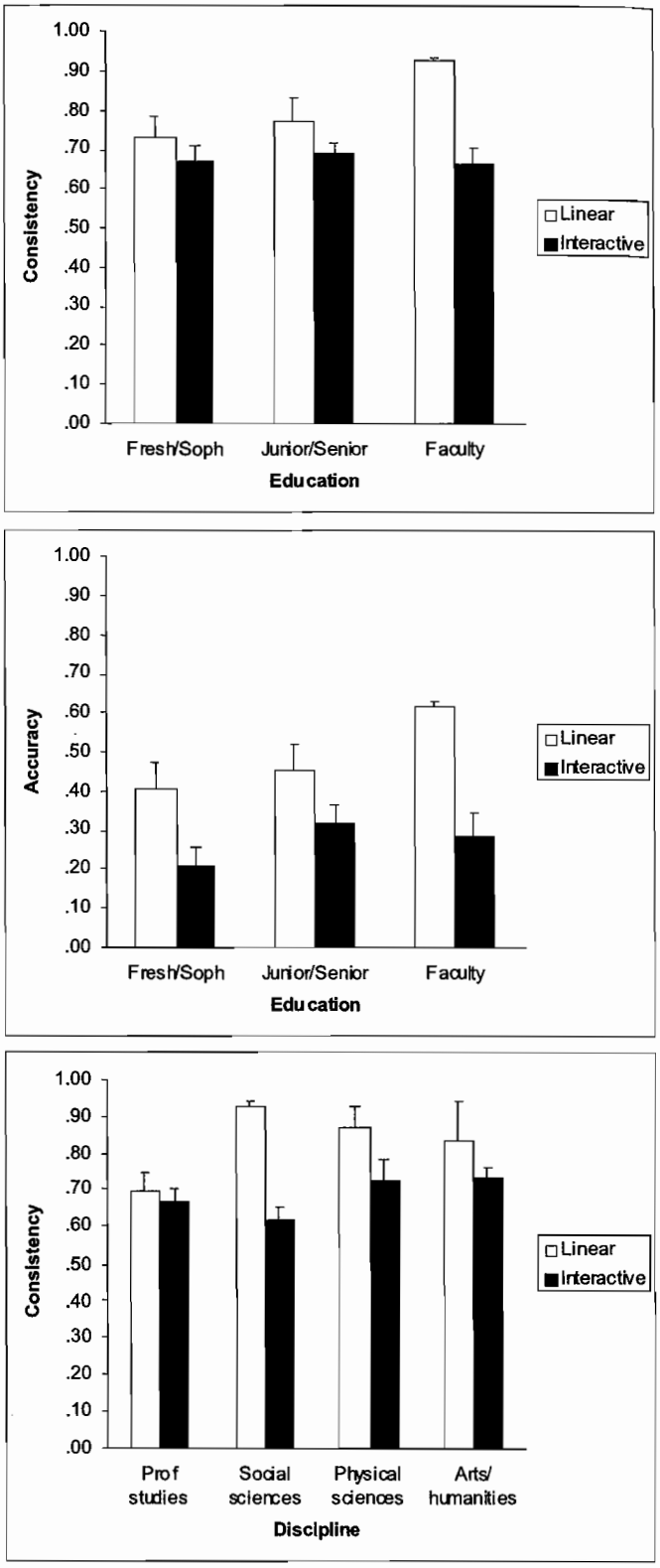

Figure 2. Top: Consistency of judgments across cue conditions and educational levels. Middle: Accuracy of judgments across cue conditions and educational levels. Bottom: Consistency of judgments across cue conditions and academic disciplines. Each error bar represents one standard error of the mean.

In the analysis of cue conditions × discipline for consistency, the main effect for cue conditions again emerged, $F(1, 109) = 27.70$, $p < .001$, $\eta^2 = .20$. There was also a main effect for discipline, $F(3, 109) = 4.80$, $p = .004$, $\eta^2 = .12$. Post hoc comparisons of means using Tukey's HSD with $\alpha = .05$ revealed that judgments of those in the physical sciences ($M = .80$, $SD = .19$) were more consistent than those in professional studies ($M = .68$, $SD = .21$). The consistency of social scientists ($M = .77$, $SD = .20$) and those in the arts and humanities ($M = .78$, $SD = .18$) was intermediate and did not differ from that of the others. There was also an interaction between cue conditions and academic discipline, $F(3, 109) = 4.20$, $p = .007$, $\eta^2 = .10$. Whereas those in professional studies achieved similar consistency in both cue conditions, members of other disciplines (particularly the social sciences) made more consistent judgments using linear rather than interactive cues (see Figure 2, bottom graph).

In the analysis of cue conditions × discipline for accuracy, the main effect for cue conditions again emerged, $F(1, 109) = 20.86$, $p < .001$, $\eta^2 = .16$. There was no main effect for discipline, $F(3, 109) = 1.34$, $p = .264$, nor did discipline interact with cue conditions, $F(3, 109) = 1.75$, $p = .160$.

Consistency was a good predictor of accuracy, $r(119) = .58$, $p < .001$. Furthermore, in a comparison of participants' accuracy with chance-level guessing,[4] 47 of 60 participants (78%) in the linear condition surpassed chance ($r = .36$, as in Study 1), whereas only 21 of 61 (34%) did so in the interactive condition, $\chi^2(1, N = 121) = 23.69$, $p < .001$, $\phi = .44$.

Confidence ratings were analyzed as in Study 1.[5] Confidence was marginally related to consistency, $r(115) = .17$, $p = .070$, and unrelated to accuracy, $r(115) = .08$, $p = .374$, and once again it did not differ across the linear ($M = 3.95$, $SD = 1.64$) and interactive ($M = 3.63$, $SD = 1.38$) conditions, $t(115) = 1.15$, $p = .253$.

Finally, as in Study 1, removal of individuals with poor consistency did not substantively alter any of the results.

Discussion

The primary findings of Study 1 were replicated in Study 2. Judgments were more consistent and more accurate in the linear condition relative to the interactive condition. Moreover, judges in the interactive condition did not recognize their comparatively poor performance: Confidence was only weakly related to consistency, even less so to accuracy, and did not differ across cue conditions.

Several additional results were uncovered in the present study. Educational levels predicted both consistency and accuracy of judgments, but in a limited way. More specifically, faculty members made more consistent and more accurate judgments than did students, but only in the linear condition. Furthermore, differences in the quality of judgments across academic disciplines were slight. Physical scientists reached more consistent judgments than did those in professional studies, with those in social sciences and arts and humanities performing at intermediate levels. In addition, there was more marked divergence in consistency across cue conditions in some disciplines (particularly social sciences) than in others. However, these differences in consistency were fairly small and did not translate into a difference in accuracy, which is ultimately the more important outcome measure. Thus, under holistic conditions, judgments were of relatively poor quality for individuals with a broad range of educational experiences.

STUDY 3

In our final study, we wished to determine whether clinical training and experience was related to the consistency and accuracy of holistic judgments. Thus, in addition to recruiting undergraduate students, we solicited participants from graduate programs in clinical and counseling psychology. We also incorporated more stringent checks to ensure that participants understood the instructions—which were once again as straightforward as possible—in each experimental condition.

Method

Design. Participants made judgments in both linear and interactive cue conditions, which formed the within-subjects independent variable. Clinical training and experience (undergraduate student vs. graduate student in clinical or counseling psychology) was also recorded as a subject variable.

Participants. One hundred thirty-two individuals participated in this experiment. One hundred fourteen of these (86%) were Elizabethtown College undergraduates par-

4. In Study 2, participants made one series of 30 judgments, whereas in Study 1 they made two. It was therefore easier to surpass chance-level accuracy once in Study 2 rather than twice in Study 1.

5. Two participants in each cue condition did not make a confidence rating.

ticipating in exchange for experimental credit in an introductory psychology course and 18 of these (14%) were graduate students in clinical or counseling psychology programs at the Pennsylvania State University participating in exchange for a payment of $20. Among the undergraduates, 77% were women and ages ranged from 17 to 24 (M = 18.65, SD = 1.07); among the graduate students, 67% were women and ages ranged from 22 to 36 (M = 27.33, SD = 3.94). Graduate students were in their first through fifth years of doctoral training (M = 2.83, SD = 1.30). Nearly all of the undergraduates and some of the graduate students participated in group sessions, but each participant completed the experiment individually at his or her own pace.

Materials. The materials were similar to those of Studies 1 and 2. A series of judgments was made for 40 hypothetical patients who had been diagnosed with schizophrenia, admitted for inpatient treatment, and given a 6-week drug treatment. The nature of the drug was manipulated across experimental conditions, and instructions were even more extensive than those in Study 2 to ensure that participants understood the nature of the additive versus interactive cue-criterion relationships. Drug X was described as more effective in the treatment of individuals with more severe hallucinations and/or delusions (the two cues; see appendix for definitions as excerpted from the task instructions), whereas Drug Y was described as more effective in the treatment of individuals with severe hallucinations or delusions, but not both. Following a thorough explanation of these relationships in the instructions were a bullet-pointed summary of the cue-criterion relationships, an illustrative case, and a procedural recap.

Unlike the *T* score units of Studies 1 and 2, both of the present cues were normally distributed across integer values ranging from 0 to 10 and the criterion (psychological functioning after a 6-week drug treatment) was normally distributed across integer values ranging from 0 through 100; frequency distributions for cues and criterion were provided in the instructions. In the Drug X (linear) condition, the criterion was computed as an additive, linear function of the two cues plus error variance; in the Drug Y (interactive) condition, the criterion was computed as a strictly interactive function of the two cues plus error. Across the first 30 cases of information, both criterion variables were equally predictable from the two cues ($R^2 = .72$). The final 10 cases of information were randomly selected from the first 25 cases (with the scores for the two cues reversed to disguise this repetition) to afford an additional test of the consistency of judgments. The same set of 40 cases was used for each experimental condition, with the values for hallucinations and delusions reversed and the order of the cases reversed within each successive set of 10 cases to disguise this repetition.

Procedure. Undergraduate students were recruited through sign-up sheets outside their introductory psychology classroom and tested in a psychology laboratory. Graduate students were recruited through direct e-mail contacts and follow-up phone solicitation, and an experimenter traveled to their university to conduct the experiment in a quiet location (e.g., office, library, or laboratory space). The experimenter explained the nature of the study, obtained informed consent, and then distributed instructions for the first task; the order of the linear and interactive conditions was counterbalanced. After participants recorded their judgments for all 40 cases in one experimental condition, they returned the instructions to the experimenter and proceeded to make two ratings on 7-point Likert scales: First they rated their confidence in the accuracy of these judgments, and then they rated how well they understood the instructions. Finally, participants were then asked to turn over the sheet on which they had recorded their judgments and write a paragraph-length summary of the instructions for that experimental condition. When finished, this procedure was repeated for the second experimental condition.

Results

Participants' judge-model consistency scores were calculated in the same manner as in Studies 1 and 2, using the *R* value from a regression equation that best predicted his or her judgments from the two cues for the 30 unique cases. (For this and all subsequent correlational measures, scores were once again transformed using Fisher's *r'* for analysis, with *M*s and *SD*s presented in correlational units.) A test-retest consistency score was calculated by correlating each participant's judgments across the 10 repeated cases (i.e., judgments on the final 10 cases were correlated with the prior judgments on the same 10 cases as they had appeared among the first 25). Accuracy was calculated across the 30 unique cases as the correlation between each participant's judgments and the criterion variable appropriate for each experimental condition.

Prior to conducting analyses, a research assistant carefully read each of the written summaries of instructions to identify participants who clearly misunderstood the directions in one or both experimental conditions.

This resulted in the exclusion of 8 undergraduate students' data, yielding a final $N = 124$. Analyses tested the extent to which consistency and accuracy varied as a function of cue conditions and clinical training and experience (henceforth referred to as "experience").

In the analysis of judge-model consistency scores across cue conditions and experience, there was a main effect for experience, $F(1, 122) = 4.40$, $p = .038$, $\eta^2 = .04$, such that graduate students ($M = .83$, $SD = .09$) made more consistent judgments than did undergraduates ($M = .74$, $SD = .18$). There was no main effect for cue conditions, $F(1, 122) = 2.45$, $p = .120$, $\eta^2 = .02$, nor an interaction between cue conditions and experience, $F(1, 122) = 2.72$, $p = .102$, $\eta^2 = .02$.

In the analysis of test-retest consistency scores across cue conditions and experience, there was also a main effect for experience, $F(1, 122) = 5.45$, $p = .021$, $\eta^2 = .04$, such that graduate students ($M = .83$, $SD = .11$) made more consistent judgments than did undergraduates ($M = .71$, $SD = .23$). There was no main effect for cue conditions, $F(1, 122) = .95$, $p = .332$, $\eta^2 = .01$, nor was there an interaction between cue conditions and experience, $F(1, 122) = .12$, $p = .730$, $\eta^2 = .00$.

In the analysis of accuracy across cue conditions and experience, there was no main effect for experience, $F(1, 122) = .35$, $p = .557$, $\eta^2 = .00$. There was a main effect for cue conditions, $F(1, 122) = 3.93$, $p = .050$, $\eta^2 = .03$, such that judgments were more accurate in the linear condition ($M = .55$, $SD = .37$) than in the interactive condition ($M = .49$, $SD = .18$). There was no interaction between cue conditions and experience, $F(1, 122) = .63$, $p = .430$, $\eta^2 = .01$.

Ratings of confidence in the accuracy of predictions were weakly correlated with judge-model consistency scores ($r[123] = .20$, $p = .029$, and $r[123] = .14$, $p = .132$ in the linear and interactive cue conditions, respectively), moderately correlated with test-retest consistency scores ($rs = .25$ and $.27$, $ps = .005$ and $.003$), and uncorrelated with accuracy ($rs = -.02$ and $.09$, $ps = .861$ and $.318$). Neither ratings of confidence nor ratings of the extent to which participants understood the instructions differed across cue conditions, both $ts < 1.40$. It is noteworthy that mean ratings on the latter were high in both the linear ($M = 5.23$, $SD = 1.23$) and interactive ($M = 5.34$, $SD = 1.25$) conditions, suggesting that the difference in accuracy across these conditions is not attributable to a differential understanding of instructions.

Finally, we examined whether the consistency and accuracy of judgments was related to the number of years of participants' graduate training. Because the correlational results did not differ across cue conditions, performance was averaged across the linear and interactive tasks. Participants' judge-model consistency scores were unrelated to the extent of their training, $r(16) = -.22$, $p = .379$, as were their test-retest consistency scores, $r(16) = .07$, $p = .775$. Surprisingly, accuracy decreased substantially with more years of clinical training, $r(16 = -.47, p = .047$.

As in Studies 1 and 2, results were reexamined after the removal of individuals with poor consistency, which was once again operationalized as $R < .40$ for either experimental condition, a value that represented a notable gap in the distributions of judge-model consistency scores for each condition and resulted in the additional exclusion of data for 25 undergraduates and 1 graduate student. Although some of the effects became a bit stronger (e.g., the main effect for cue conditions on accuracy scores) and some a bit weaker (e.g., the main effect for experience on test-retest consistency scores), there was no qualitative difference in any of the results reported above.

Discussion

Although this study failed to replicate the difference in the consistency of judgments that was observed in Studies 1 and 2, it did replicate the more important difference in accuracy. Once again, despite this difference in accuracy there was no corresponding difference in confidence ratings, suggesting that participants did not recognize their comparatively poor performance when making judgments in the interactive condition, and participants' confidence did not predict their accuracy within experimental conditions. These results were obtained in a study that included not only extensive, clear instructions but also more stringent checks to ensure that participants understood these instructions.

GENERAL DISCUSSION

The central issue addressed by these three experiments concerns the capacity of human judges to reach sound holistic judgments. In the first two studies, judgments made on the basis of interacting cues were of poor consistency and accuracy, and the latter finding held in the third study as well. The relatively high quality of judgments made from additively related variables in the comparison condition suggested that the poor showing of those in the interactive condition could not be attributed to other aspects of the task. Moreover, the judgment task in this research was the simplest holistic task that

we could conceive,[6] and the failure of participants to perform well under these conditions casts serious doubt on the efficacy of holistic judgment more broadly. We originally intended this experimental task to be the first in a series of increasingly complex judgment tasks to which higher-order interactions, nonlinear cue-criterion relationships, and other features would be added incrementally to test the limits of human judgmental ability. As suggested by the judge-modeling literature on configural cue utilization—not to mention the experience of statistics instructors—it appears that even the most straightforward of interaction effects presents a formidable cognitive challenge.

While the results are, on the whole, highly problematic for proponents of holistic judgment, we must be careful not to ignore potential individual differences. There are considerable variations in the extent to which people enjoy engaging in complex thought (Cacioppo, Petty, Feinstein, & Jarvis, 1996), and differences such as these can influence the quality of judgments (Ruscio, 2000). Thus, certain individuals may be more proficient at holistic reasoning than our aggregate results suggest. However, the fact that only 10% of the participants in Study 1 were able to replicably surpass chance-level guessing in the interactive condition supports the notion that even this minimally complex task is quite demanding. Moreover, given its weak association with consistency and accuracy in all three studies, subjective confidence in one's judgments seems unlikely to be a useful indicator of individual differences.

Might training or experience improve performance? In Study 2, neither educational level nor academic discipline was related to the accuracy of holistic judgments. There were disciplinary differences in consistency, but these did not translate into differences in accuracy. Faculty did achieve greater consistency and accuracy levels than students, but not when making holistic judgments from interacting cues. These results do not rule out the possibility that individuals with more specialized training and experience might fare better than those in the present studies.

A more direct examination of this hypothesis was provided by Study 3, the results of which failed to provide evidence that individuals with clinical training and experience were immune to the difficulties in making holistic judgments. Because this was a relatively small sample of graduate students, rather than a larger sample of seasoned practitioners, strong conclusions cannot be drawn from these data. However, judgment research suggests that training and experience lead to improvements in accuracy only in the presence of immediate, concrete, and unambiguous feedback (Dawes, 1994; Faust, 1986). Very little, if any, of the training received by mental health practitioners provides this type of feedback, and it is seldom actively solicited or otherwise encountered on the job (Faust, 1986; Ruscio, 1998b; see Smith & Dumont, 1997, for ways to improve judgment through training and practice).

In fact, our experience with several of the more practice-oriented faculty members in Study 2 and correlational evidence from Study 3 suggests not only that clinical training does not always lead to improved judgment, but that it also may have the opposite effect some of the time. Several of the faculty participants in Study 2 from mental health–related fields insisted that they needed more information to properly contextualize the limited data with which we provided them. Having taken the atypical step of providing *only* relevant information—which greatly simplifies the task—and explaining this fact to participants, we wondered where the preference for additional data would lead practitioners of a similar mindset.[7] Moreover, when told about the irrelevant information that had been made available to judges in the previous study, one participant expressed a strong desire to have access to it, even when the experimenter stressed that this information was useless for making these predictions. Much to our surprise, in Study 3 the number of years of graduate training was unrelated to consistency but predicted substantial *decreases* in accuracy. Again, in light of the restricted sample of individuals with clinical training and experience in this study, this correlational finding should be interpreted with due caution. In any event, it suggests an interesting avenue for future research.

Some clinicians may feel strongly about needing all available information precisely because of their training, which often emphasizes the need to consider all aspects of the client and his or her situation—family, educational, and health history; current signs, symptoms, stressors,

6. Our judgment tasks involved a fairly high level of predictability: A statistical integration of the two relevant cues achieved $R^2 = .54$ when predicting the criteria used in Studies 1 and 2 and $R^2 = .72$ when predicting the criteria used in Study 3. Although it is likely that accuracy levels would be higher if the criteria were even more predictable, it seems unlikely that the accuracy achieved in an interactive cue condition would increase substantially more than that achieved in a linear cue condition.

7. Although the manipulation of irrelevant information in Study 1 produced no impact on accuracy levels, this may be due to the heavy emphasis on its irrelevance in the task instructions. Previous research has shown that irrelevant information can dilute the quality of judgments (e.g., Ruscio, 2000), and it remains for future research to establish the conditions under which irrelevant information is particularly problematic.

and strengths; psychological, neuropsychological, and other test results; and so on—when making decisions or planning interventions (e.g., Lezak, 1995). This is sound advice so long as the information is relevant to the judgment at hand. Considering empirically irrelevant data can lead to a variety of judgmental biases (e.g., over- or underpathologizing, diagnostic overshadowing, or biases involving age, sex, race, or social class; Garb, 1999). Moreover, one only needs to contextualize information by way of holistic judgment when the variables do indeed interact. When they do not, consideration of relevant factors in an additive manner is both empirically sufficient and considerably less cognitively demanding.

Based on research regarding the small effects of training and experience on judgment (for reviews, see Garb, 1989, 1999), our interactions with participants, and the results of Study 3, we doubt that the poor performance we observed is in any way unique to the populations that we tested. We suspect that few professionals would succeed at this task, and that success at considerably more complex tasks would likely drop off precipitously. Nevertheless, it would be interesting to know whether practicing clinicians or other professional decision makers would make better or worse holistic judgments relative to our undergraduate, faculty, and graduate student participants.

Consistent with the decades-old literature on the modeling of clinical judgment, we are left wondering whether individuals do, in practice, actually perform the sophisticated feats of information integration that they profess. Might they in fact be attending to individual features rather than a complex whole and combining this information as best they can, perhaps through a crude approximation of an additive, linear model or a cognitively simpler heuristic of some kind? Research on social cognition (e.g., Nisbett & Wilson, 1977a, 1977b) and on the psychology of judgment and decision making (e.g., Faust, 1984; Slovic & Lichtenstein, 1971) suggests that we often lack insight into our cognitive processes. In each of the present studies, subjective confidence in the accuracy of predictions did not differ across groups of participants who achieved different—sometimes dramatically different—accuracy levels. Although the reasons for this poor match between accuracy and confidence are unclear, this finding nonetheless demonstrates that participants badly misread their own performance. Thus, our data are consistent with the accumulated evidence in the failed search for interactive judgment through the examination of paramorphic judge-models.

The examination of cognitive limitations in these studies bears directly upon the broader literature on clinical and statistical decision making. As in hundreds of other studies that have compared these two approaches to making predictions, participants in both of the present cue conditions—linear and interactive—achieved lower accuracy levels than did a simple SPR. The statistical integration of information in both experimental conditions accounted for more than half of the variance in the respective criterion measures. Participants made a respectable showing in the linear condition, but their judgments still explained only about one quarter of the variance in the criterion. Moreover, in the interactive condition, participants' judgments explained much less of the criterion variance than that explained by a simple SPR. This suggests not only that cue interactions can reduce the consistency and accuracy of human judgment, but also that a more mechanical method of data integration is *particularly* important under the very circumstances that are alleged by some clinicians to favor holistic judgment policies. If real-world information does indeed interact in clinical cases, SPRs can accommodate this knowledge and put it to good use—provided that practitioners can in fact describe the nature of the interaction effects. Fortunately, the development and implementation of SPRs is not an intrinsically difficult undertaking (Ruscio, 1998a; Swets et al., 2000).

Not only does holistic reasoning constitute a difficult judgmental feat, it may often be unnecessary. Dawes (1979) persuasively argues that, in the real world, variables tend to be related *monotonically*. That is, the *direction* of a variable's effect does not typically change as it interacts with other variables. Nonlinear relationships tend to be monotonic, too. The significance of monotonicity is that even linear SPRs are able to capture interactive and nonlinear effects well. A simple additive sum of linear main effects can be astonishingly potent even when variables are not additively or linearly related to the criterion. To demonstrate this point, we performed an exercise described by Yntema and Torgerson (1961): We constructed a data set including three uncorrelated factors (x, y, and z) with rectangular distributions including the integers from 1 to 10, resulting in 1,000 cases ($10 \times 10 \times 10$). We used this data set to address two questions. First, how well does a linear model based solely on additive main effects predict a criterion composed entirely of interactive relationships, $(x \times y) + (x \times z) + (y \times z)$? The multiple correlation coefficient is .97, which accounts for 94% of the criterion variance. Second, how well does this same linear model of additive main effects predict a criterion composed entirely of nonlinear effects, $x^2 + y^2 + z^2$? The multiple correlation coefficient is .98, which accounts for 95% of the criterion variance. Thus, the predictive power

of linear models with monotonic relationships leaves virtually no benefit to using the far more cognitively demanding strategy of making holistic judgments.

In many areas of professional practice—including clinical assessment, case formulation, and treatment planning—individuals are urged to contextualize all available information about a client in order to make the most appropriate decisions about his or her case. Given that this process requires holistic judgment, our results suggest that this may place an unrealistic cognitive demand on the clinician. Research on the ability to detect configural relationships suggests that comprehensive assessments will be practically useful to the extent that the relevant contextual variables can be identified as explicitly as possible, and research on statistical prediction suggests that this information should be combined mechanically. The present research joins a large literature supporting the benefits of assessing only those characteristics relevant to the particular judgment or decision at hand and combining them using a cross-validated SPR. In contrast, a penchant for holistic judgment may divert practitioners' attention beyond what is relevant and pose unnecessary cognitive obstacles to the valid integration of clinical data.

REFERENCES

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin, 119*, 197–253.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications*. San Diego, CA: Academic Press.

Dawes, R. M. (1964). Social selection based on multidimensional criteria. *Journal of Abnormal and Social Psychology, 68*, 104–109.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571–582.

Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy built on myth*. New York: Free Press.

Dawes, R. M. (2001). *Everday irrationality: How pseudo-scientists, lunatics, and the rest of us systematically fail to think rationally*. Boulder, CO: Westview Press.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668–1674.

Einhorn, H. J. (1971). Use of nonlinear, noncompensatory models as a function of task and amount of information. *Organizational Behavior and Human Performance, 6*, 1–27.

Faust, D. (1984). *The limits of scientific reasoning*. Minneapolis: University of Minnesota Press.

Faust, D. (1986). Research on human judgment and its application to clinical practice. *Professional Psychology: Research and Practice, 17*, 420–430.

Faust, D., & Ziskin, J. (1988). The expert witness in psychology and psychiatry. *Science, 241*, 31–35.

Ganzach, Y. (1995). Nonlinear models of clinical judgment: Meehl's data revisited. *Psychological Bulletin, 118*, 422–429.

Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin, 105*, 387–396.

Garb, H. N. (1999). *Studying the clinician: Judgment research and psychological assessment*. Washington, DC: American Psychological Association.

Goldberg, L. R. (1968). Simple models or simple processes? Some research on clinical judgments. *American Psychologist, 23*, 483–496.

Goldberg, L. R. (1991). Human mind versus regression equation: Five contrasts. In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology* (Vol. 1, pp. 173–184). Minneapolis: University of Minnesota Press.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19–30.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2*, 293–323.

Hammond, K. R., & Summers, D. A. (1965). Cognitive dependence on linear and nonlinear cues. *Psychological Review, 72*, 215–224.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin, 57*, 116–131.

Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford University Press.

Meehl, P. E. (1954). *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.

Meehl, P. E. (1967). What can the clinician do well? In D. N. Jackson & S. Messick (Eds.), *Problems in human assessment* (pp. 594–599). New York: McGraw-Hill.

Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment, 50*, 370–375.

Nisbett, R. E., & Wilson, T. D. (1977a). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology, 35*, 250–256.

Nisbett, R. E., & Wilson, T. D. (1977b). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231–259.

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

Oskamp, S. (1965). Overconfidence in case-study judgments. *Journal of Consulting Psychology, 29*, 261–265.

Popper, K. A. (1959). *The logic of scientific discovery*. New York: Basic Books.

Ruscio, J. (1998a). Information integration in child welfare cases: An introduction to statistical decision making. *Child Maltreatment, 3*, 143–156.

Ruscio, J. (1998b). The perils of post-hockery. *Skeptical Inquirer, 22*, 44–48.

Ruscio, J. (2000). The role of complex thought in clinical prediction: Social accountability and the need for cognition. *Journal of Consulting and Clinical Psychology, 68*, 145–154.

Ruscio, J. (2003). Holistic judgment in clinical practice: Utility or futility? *The Scientific Review of Mental Health Practice, 2*, 38–48.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 66*, 178–200.

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance, 6*, 649–744.

Smith, D., & Dumont, F. (1997). Eliminating overconfidence in psychodiagnosis: Strategies for training and practice. *Clinical Psychology: Research and Practice, 4*, 335–345.

Stewart, T. R. (1988). Judgment analysis: Procedures. In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgment: The SJT view* (pp. 41–74). Amsterdam: North-Holland Elsevier.

Summers, D. A., Summers, R. C., & Karkau, V. T. (1969). Judgments based on different functional relationships between interacting cues and a criterion. *American Journal of Psychology, 82*, 203–211.

Summers, D. A., Taliaferro, J. D., & Fletcher, D. J. (1970). Subjective vs. objective description of judgment policy. *Psychonomic Science, 18*, 249–250.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*, 1–26.

Wiggins, N., & Hoffman, P. J. (1968). Three models of clinical judgment. *Journal of Abnormal Psychology, 73*, 70–77.

Wills, C. E., & Moore, C. F. (1994). Judgment processes for medication acceptance: Self-reports and configural information use. *Medical Decision Making, 14*, 137–145.

Yntema, D. B., & Torgerson, W. S. (1961). Man-computer cooperation in decisions requiring common sense. *IRE Transactions of the Professional Group on Human Factors in Electronics, HFE-2*, 20–26.

APPENDIX

Definitions of Cues

Studies 1 and 2

Attention Deficit: High scorers on this scale have difficulty maintaining their concentration or focusing their attention. Their mind wanders, they are easily distractible, and they have a hard time suppressing other thoughts that pass through their minds.

Hyperactivity: High scorers on this scale have poor impulse and motor control. They are constantly on the go, acting out—often in inappropriate ways—and reacting almost immediately to real or imagined events in the environment around them.

NOTE: In the information that you will be given, Attention Deficit and Hyperactivity are independent of one another. That means that there is no systematic relationship between these two scores, so that any given child may score at any level of Attention Deficit regardless of his or her level of Hyperactivity. Be sure to think about them as entirely separate factors.

Responsiveness to Treatment: This is what you will be predicting. High scorers on this scale show marked improvement in ADHD symptoms after they begin taking Attevil. To the mutual satisfaction of themselves, their parents, their teachers, and others, they experience significant reductions in problematic thoughts and behaviors.

[For the series of cases that included four irrelevant cues, the following definitions were also included.]

Family Problems: High scorers on this scale report considerable family discord. Their families are described as lacking love, quarrelsome, and unpleasant. They may even report hating members of their families. Their childhood may be portrayed as abusive and/or lacking in affection.

Depression: High scorers on this scale exhibit symptoms of depression. They suffer from feelings of discouragement, pessimism, and hopelessness that characterize the clinical status of depressed children as well as the basic personality features of hyper-responsibility and high personal standards.

Health Concerns: High scorers on this scale worry about their health and feel sicker than the average child. They complain of gastro-intestinal symptoms (such as constipation), neurological problems (such as dizziness), sensory problems (such as poor hearing), pain (such as headaches), and respiratory trouble (such as asthma).

Fears: High scorers on this scale indicate many specific fears. They typically fear things like the sight of blood; high places; animals such as snakes, mice, or spiders; leaving home; fire; storms and natural disasters; water; the dark; being indoors; and dirt.

Study 3

Delusions: The extent to which the patient strongly believes things that are known to be untrue (e.g., that thoughts are being inserted into her mind, or that he is Napolean).

Hallucinations: The extent to which the patient hears, sees, or feels things that aren't really there.