Journal of Counseling Psychology
1990, Vol. 37, No. 3, 343-351

0022-0167/90/$00.75

Anchoring Errors in Clinical Judgments: Type I Error, Adjustment, or Mitigation?

Michael V. Ellis, Erica S. Robbins, Deborah Schult,
Nicholas Ladany, and Janice Banker
Department of Counseling Psychology
University at Albany, State University of New York

The nature of anchoring errors in clinical judgments was clarified. Study 1 tested if gender mediates the occurrence of anchoring errors. Judgments from 103 undergraduate psychology students evidenced neither anchoring errors nor gender differences. Given the inability to replicate Friedlander and Stockman's (1983) study, two rival hypotheses were advanced: the adjustment hypothesis (practitioners adjust appropriately their clinical judgments after receiving new client information) and the adjustment mitigation hypothesis (initial anchoring effects are mitigated by an adjustment effect). The judgments from 157 psychologists in Study 2 affirmed the adjustment and mitigation hypotheses over the anchoring hypothesis alone. The mitigation process appears adaptable in clinical judgments.

One topic of interest to both researchers and practitioners is the accuracy of clinical judgments about clients. The empirical evidence indicates that clinical prediction is less reliable than actuarial prediction (Meehl, 1954) and is subject to faulty inferential processes and biases (e.g., Arkes, 1981; Bieri, Orcutt, & Leaman, 1963; Klayman & Ha, 1987). Theoretical explanations have been advanced to account for the biases and faulty inferences in the judgment processes (e.g., Klayman & Ha, 1987; Nisbett & Ross, 1980; Snyder, 1981) and include Tversky and Kahneman's (1974) heuristic principles for judgments under uncertainty. Recently, Tversky and Kahneman's anchoring heuristic was applied to clinical judgments and investigated (Friedlander & Phillips, 1984; Friedlander & Stockman, 1983; Levin, 1984). These studies, however, yielded equivocal results (cf. Bieri et al., 1963). The general focus of the two studies reported in this article was on clarifying the nature of anchoring errors in clinical judgment.

Tversky and Kahneman (1974) proposed three heuristic (decision-making) principles that one uses to make judgments and predictions when presented with incomplete information or faced with uncertainty, a ubiquitous occurrence in counseling situations (Pepinsky & Pepinsky, 1954). Relying on these heuristics effectively reduces the complexity of the judgmental task; however, they also lead to predictable and sometimes severe errors. The first judgmental heuristic, *representativeness,* is typically used when one assigns the probability that an event or object belongs to a specific class (e.g., ascribing a diagnosis to a client), which results in biases such as relying on stereotypes and the gamblers' fallacy (assuming chance events self-correct). The second heuristic, *availability,* incurs bias when the familiarity, retrievability, and salience of instances or scenarios become the basis for assessing the frequency or size of a class (e.g., number of morbidly obese clients) or the probability of an occurrence (e.g., predicting a suicide attempt). The third heuristic, *anchoring,* is usually used when making a series of numerical estimates or predictions (e.g., rating a client's prognosis after each session). Anchoring errors occur when initial estimates of a phenomenon bias subsequent estimates towards the initial values (insufficient adjustment).

Recently, Friedlander and her associates have systematically investigated anchoring errors in clinical judgments. Using a sample of experienced mental health professionals, Friedlander and Stockman (1983) found significant anchoring effects for an anorexic case but not for a suicidal case. Friedlander and Phillips (1984) employed undergraduate students to rate the same anorexia case. No significant anchoring effects were found. However, students were found to be significantly less confident in their ratings than the 1983 practitioner sample. Levin's (1984) unpublished research incorporated a valence factor (i.e., positive academic achievement information or suicidal information) into the same paradigm with a sample of psychology graduate students. Manipulation checks indicated that the suicidality manipulation worked whereas the positive manipulation did so marginally. No significant effects emerged for valence or anchoring on either dependent variable.

In each article the authors advanced explanations for the equivocal results. Friedlander and Stockman (1983) concluded that the severity of the pathology for the suicidal client overshadowed the potential for biased judgments whereas anchoring biases did emerge for the less severe anorexic case. Friedlander and Phillips (1984) suggested that more experienced, confident judges may be more susceptible to anchoring errors. Finally, Levin (1984) speculated that the lack of sig-

This research was supported in part by a grant from the Campus Incentive Funds of the University at Albany, State University of New York.

We express our gratitude to Bruce E. Wampold for suggesting the mitigation hypothesis and corresponding analysis and to Douglas Strohmer for his comments on an earlier draft of this article.

Erica Robbins is currently a psychologist at Four Winds Hospital, Saratoga Springs, New York.

Correspondence concerning this article should be addressed to Michael V. Ellis, Department of Counseling Psychology, ED 220, University at Albany, State University of New York, 1400 Washington Avenue, Albany, New York 12222.

nificant results was attributable to the participants' lack of counseling experience and to the saliency of the suicidal manipulation. We therefore began Study 1 with this question: Why were anchoring errors found in only one of three counseling analogue studies?

Study 1

After a close examination of the previous studies, two things became apparent: The gender composition of the practitioner sample (35% women) differed markedly from the student samples (79% and 65% women, respectively), and minimal attention had been paid to Type I and Type II error rates. These two observations stimulated our research hypotheses and design. The methodological issues could be dealt with in a straightforward manner. Type I and Type II error rates were attended to by using multivariate procedures and by performing a priori power analyses to ensure adequate statistical power (minimizing Type II errors; Cohen, 1988; Haase & Ellis, 1987).

Given the marked difference in the men:women ratios in the samples and in the particular case materials, we reasoned that gender could have mediated the occurrence of anchoring errors. To wit, Friedlander and her associates appear to have assumed that the anchoring heuristic would be utilized when the participants rated the case materials, exclusive of the other two heuristic principles. The tasks of rating sequentially the client's level of functioning and prognosis, however, would probably necessitate making a series of intuitive judgments that involved other heuristic principles. Given that Tversky and Kahneman's (1974) availability, representativeness, and anchoring heuristics are combined, the more similar a client is to classes or stereotypes both familiar and salient to the person who is making judgments, the more probable that the case will be assigned to and rated on the basis of the class or stereotype; the class or stereotype serves to anchor subsequent ratings of the case.

Applying this logic to the aforementioned case materials and samples, gender becomes a defensible mediator of anchoring bias. Anorexia is widely recognized as a disorder of early to late adolescence that almost exclusively afflicts young women (American Psychiatric Association, 1987; Strober, 1986). It seems reasonable to argue that anorexia is more familiar and salient and, hence, a more readily available classification of clients for women (especially late adolescent and young adult women) than men. As such, women are more likely to be cognizant of the life-threatening potential of anorexia (Strober, 1986) and adjust their judgments accordingly. Men, on the other hand, tend to find anorexia less salient and more difficult to recall. They are thus more likely to assign an anorexia case to a more readily available and less severe client category or stereotype (e.g., dysthymia) and subsequently not adjust sufficiently their ratings. This could explain why anchoring effects were found for the anorexia case by Friedlander and Stockman (1983) in a predominately male sample and not in Friedlander and Phillips's (1984) and Levin's (1984) predominately female samples.

The purpose of Study 1 was to test further Tversky and Kahneman's (1974) theory through a replication and extension of Friedlander and Stockman's (1983) research. We hypothesized that anchoring errors would be found for men's ratings of an anorexia case and not for women's ratings. According to the representative principle, we also expected that women would be more confident in their judgments than would men; the more similar the case is to the client category or stereotype, the more confidence that will be expressed by those who rate the case. Combining our arguments with the assumption that a female case is more salient to female judges, we expected that if anchoring errors were to occur with a suicidal female client, they would be found only for male judges.

Method

Participants

With the effect size ($\hat{\rho}^2 = .102$) from Friedlander and Stockman (1983), the a priori power analysis indicated that 96 participants were required for power of .80 (Cohen, 1988). The final sample consisted of 103 volunteers from undergraduate abnormal psychology classes at a northeastern state university. The average age of the 40 male and 63 female students was 20.84 ($SD = 2.35$) years old. They had accumulated a mean grade point average of 3.06 ($SD = 0.41$) and 15.00 credit hours in psychology ($SD = 3.50$).

Procedure

The experiment entailed a 2 (sex) × 2 (time) × 2 (case) factorial design in which case was a repeated measures factor. The time factor referred to when the salient information appeared in the case materials, either early (Interview 1) or late (Interview 4). Thus, there were four groups: male early (ME; $n = 20$), male late (ML; $n = 20$), female early (FE; $n = 32$), and female late (FL; $n = 31$). The two cases appeared in counterbalanced order (by case and by time) with male and female participants randomly assigned separately to one of the four orders.

Adhering to Friedlander and Stockman's (1983) procedures, we informed the participants that the study assessed "factors influencing clinical judgment" and we then presented the two cases (Joanne and Gina). Each case comprised detailed written synopses from five consecutive psychotherapy sessions. The materials differed in the order of case presentation and at which point the salient pathognomonic information was introduced in the sequence of interviews. After the participants read each session summary, they rated the client's level of functioning and prognosis and indicated their degree of confidence in making these ratings. Participants were encouraged to refer to earlier sessions as they proceeded through the materials in order to avoid recency effects (Arkes, 1981). In addition, a private practice setting was described to eliminate the potential contamination of contextual cues (Bieri et al., 1966).

Case Materials

The case materials and manipulations were identical to those of Friedlander and Stockman (1983, p. 640), who used a panel of expert practitioners to validate both cases. Although the two women described in the materials differed on marital and occupational status, age, and level of disturbance, both clients were presented as motivated and capable of behavior change and insight. Any cues of client progress in therapy were left ambiguous. Only noninferential data, with many direct client quotes, were presented in the cases. Infor-

mation about previous treatment or about the therapist's orientation, characteristics, or interventions was not provided.

Joanne, a 44-year-old, single free-lance sportswriter, was the more disturbed, characterological case. The pathognomonic information for Joanne consisted of: (a) a history of suicidal ideation; (b) prior suicide attempts; and (c) the client's statements that she thought about dying and had made recent suicidal gestures. Gina, the less disturbed case, was a 23-year-old librarian who sought help because she was "depressed and nervous about getting pregnant." The salient information manipulated for Gina consisted of: (a) a history of hospitalization and forced feeding; (b) a prior anorexia nervosa diagnosis; and (c) the client's statements that she was dieting and had already lost 22 pounds.

Dependent Measures

After reading each session summary, the participants rated the client's level of functioning on the Global Assessment Scale (GAS; Endicott, Spitzer, Fleiss, & Cohen, 1976). Behavioral anchors relevant to mood and social and cognitive functioning were provided in 10-point intervals along a hypothetical continuum (1–100) of mental illness to health. Higher ratings indicated fewer symptomatic behaviors and less need for help. As for reliability, Endicott et al. (1976) reported high interclass correlations (*r*s = .61–.91) by using vignettes, case records, and adult patients. Adequate concurrent validity was achieved with a variety of symptom and severity ratings, wherein the GAS demonstrated the greatest sensitivity to change (Endicott et al., 1976).

The Prognostic Scale (PS; Friedlander & Stockman, 1983) was anchored at each level (1 = *superior* to 7 = *grossly impaired*). The participants were directed to "rate the highest level of adaptive functioning which could be expected for this client (i.e., prognosis)" (p. 641). The Confidence Scale (CS), which ranged from 1 (*totally confident*) to 7 (*not at all confident*), was adapted from Friedlander and Stockman. Participants were instructed to "rate your confidence in the two judgments you just made" for each session summary. Although no reliability data exists for the PS and CS, they were used in the three previous studies.

After completing the study, the participants provided demographic information and then defined anorexia as a check of their recognition of the disorder presented in Gina's case. We judged all participants to have a sufficient understanding for the purposes of the study.

Results

Manipulation Checks

A 2 (order) × 2 (case) repeated measures multivariate analysis of variance (MANOVA) with Interview 1 data was performed to determine if the counterbalancing procedure worked and if the two cases were initially judged differently. The analysis yielded nonsignificant effects for order of case presentation and nonsignificant interaction effects, Pillai's $V < .056$, Fs(3, 104) < 2.04, ps > .112, $\hat{\rho}_m^2$s < .029 ($\hat{\rho}_m^2$ is the shrunken multivariate effect size; see Cohen, 1988), and thus affirmed the counterbalancing procedure. A significant case effect was found, $V = .259$, $F(3, 104) = 12.12$, $p < .0001$, $\hat{\rho}_m^2 = .238$. In order to understand exactly how the two cases differed, the follow-up procedures consisted of both univariate t tests and standardized discriminant function coefficients (Haase & Ellis, 1987). Follow-up tests revealed that raters judged the client in the suicidality case as functioning lower and as having a worse prognosis and that raters were less confident of their judgments than in the anorexia case (space limitations preclude presenting follow-up results; they are available from Michael V. Ellis). Because there were no order effects, the data for this factor were collapsed in subsequent analyses.

In order to test the effectiveness of the manipulations, two multivariate t tests were conducted on ratings from the first interview, one per case. A Bonferroni adjusted alpha was used to protect against inflated Type I error rates ($\alpha_{pc} = .05 \div 2 = .025$). At Interview 1, there were significantly different ratings between the group exposed to the suicidality manipulation and the group with no initial manipulation, $V = .254$, $F(3, 104) = 11.83$, $p < .0001$, $\hat{\rho}_m^2 = .235$. In follow-up procedures we found that the multivariate effect largely comprised significantly healthier GAS scores and, to a lesser extent, better prognosis ratings for the early condition. Although in the expected direction, the early and late anorexia groups did not have significantly different initial ratings, $V = .042$, $F(3, 107) = 1.55$, $p = .207$, $\hat{\rho}_m^2 = .015$. Thus, the manipulations worked for the suicidality case but not for the anorexia case. Nonetheless, data from both cases were analyzed because the anorexia case has yielded anchoring errors in previous studies that have not incorporated manipulation checks.

Major Analyses

The judgments of interest were the GAS (level of functioning), PS, and CS ratings after Interview 5. When anchoring occurs, raters do not make sufficient adjustments from their initial judgments to their later ones. Theoretically, raters exposed to the pathognomonic information earlier rather than later will have significantly worse ratings at the last judgment point. Even though all participants were exposed to the identical information by the fourth session, the anchoring hypothesis was subjected to greater risk by using Interview 5 judgments (Friedlander & Stockman, 1983). The principle hypothesis under investigation was that male participants would show anchoring errors (e.g., ME < ML) whereas female participants would not have anchoring errors (e.g., FE = FL), especially for the anorexic case of Gina. To test this hypothesis a 2 (time) × 2 (gender) × 2 (case) repeated measures MANOVA was conducted. Of the seven effects potentially tested by the MANOVA, the four effects of the within-model factors were relevant specifically to the hypotheses. Neither the three-way interaction nor the two two-way interactions attained statistical significance, Vs < .031, Fs(3, 99) < 1.02, ps > .387, $\hat{\rho}_m^2$s < .016. The main effect for case was significant, $V = .103$, $F(3, 99) = 3.72$, $p = .014$, $\hat{\rho}_m^2 = .076$. Results of the follow-up procedures indicated that in comparison to Gina, Joanne was judged as functioning at a significantly lower level and that participants were less confident of their ratings of Joanne. In short, there were no reliable anchoring effects; hence none of the hypotheses were supported.

Discussion

The purpose of this study was to determine if gender served as a mediating variable of the anchoring heuristic when an-

choring errors occur. Even though we conducted this experiment with more than an 80% probability of detecting an anchoring effect of .10 or larger, anchoring errors were not found for either case. In fact, the largest observed effect size for anchoring errors was $\hat{\rho}_m^2 = .002$, a trivial effect (Cohen, 1988). Like Friedlander and Phillips's (1984) and Levin's (1984) findings, these results were in opposition to those of Friedlander and Stockman (1983), who found anchoring errors for the anorexia case. No significant differences were found for gender. The only significant result confirmed that the judges were less confident in their more pathological ratings of the suicidality case than the anorexia case, which is consistent with previous studies. Thus, our hypotheses were not supported, and our original question remained (i.e., why were anchoring errors in clinical judgments found in only one counseling analogue study?).

The results from the preliminary analysis were helpful in attempting to answer the question post hoc. The manipulation checks revealed significant group differences for the suicidality manipulation but not for the anorexia manipulation. The methodology that was used by Tversky and Kahneman (1974) and in the series of studies with Friedlander and Stockman's (1983) paradigm requires significantly different starting points for the two groups. That is, if the two groups do not have significantly different starting judgment values, their final judgments by definition cannot be significantly different. Hence, the only condition in which anchoring effects could have emerged in this study involved the suicidality case. This leaves us with a paradox. Friedlander and Stockman speculated that the more serious suicidality information obviated an anchoring bias whereas judgments of a moderately disturbed client did not. Therefore, it appears that to attain significantly different starting points, the salient information (manipulation) has to be sufficiently strong so that it consequently prevents anchoring effects from occurring. If this paradox is correct, then how can anchoring effects be assessed? One solution may be to design a study whereby the pathognomonic information is contrasted to healthy information rather than to a neutral condition in which no salient information is presented. This methodology ought to yield significantly different starting values even for a moderately disturbed client.

The preliminary analyses yielded another noteworthy finding. Presenting either the suicidality case first or the anorexia case first had no reliable effect on judgments about the other case. The suicidality case, however, was judged significantly more pathological than the anorexia case at both Interview 1 and Interview 5. Applying the anchoring heuristic, one may expect to find that when one judges two divergent cases, the first case ought to anchor judgments of the second case. However, no anchoring effects of this nature were manifested. Perhaps Bieri et al. (1963) were correct. When new case materials are presented, the raters have less of a need for remaining consistent with previous judgments, but they are free to adjust their judgments.

The speculative explanations posited to account for the contradictory results have focused on the severity of client disturbance (manipulations) and the differences between the professional and student samples (Friedlander & Phillips, 1984; Friedlander & Stockman, 1983; Levin, 1984). These explanations are applicable to Study 1 as well. Indeed, the results may not be generalizable beyond the nonclinician student sample (see Garb, 1989). Another explanation is that the anchoring errors observed by Friedlander and Stockman may be due to Type I error. That is, given a series of partial replications of the 1983 study, the single instance of anchoring effects may be attributable to chance occurrences and sample vagaries. Taken collectively, then, the results from Study 1 plus the three prior investigations conducted by Friedlander and her associates posed a more basic question: Do anchoring errors occur in clinical judgments?

Study 2

The purpose of Study 2 was to test rigorously the veracity of the anchoring hypothesis in a practitioner population. When developing the rationale for Study 2, we attempted to test multiple theories (see Klayman & Ha, 1987) and to test (rule out or control) rival explanations associated with the anchoring heuristic (e.g., methodological flaws or anomalies; Serlin & Lapsley, 1985). A multiple hypothesis experiment was designed to obviate making inferences from nonsignificant results (e.g., failing to reject the null hypothesis and concluding that there are no anchoring effects; see Cook & Campbell, 1979, pp. 44–50). The first theory, Tversky and Kahneman's (1974) anchoring hypothesis, has been discussed previously. A second competing theory is the adjustment hypothesis. The adjustment hypothesis was adduced from early research and theory on anchoring (e.g., Bieri et al., 1966; Bieri et al., 1963; Campbell, Hunt, & Lewis, 1957; Sherif & Hovland, 1961), which has suggested that in a succession of judgments, raters assimilate their judgments (i.e., raters respond and adjust their judgments to be consistent with the more recent manipulation). The perspective of counselors sufficiently adjusting their judgments of clients as they receive additional information coincides with Pepinsky and Pepinsky's (1954) model. In fact, Pepinsky and Pepinsky explicate and emphasize the process of testing and reformulating hypotheses (i.e., judgments) about clients. Extending the assimilation model to judgments of clinical cases, we arrived at a definition for the adjustment hypothesis: Practitioners would adjust appropriately their clinical judgments after being exposed to new salient information (i.e., anchoring errors do not generate meaningful effects). Hence, two groups of practitioners provided the same information initially and later exposed to different salient information would have significantly divergent final judgments. Conversely, two groups with significantly disparate initial judgments but exposed eventually to identical information would have equivalent final judgments.

One may argue, however, that the anchoring and adjustment theories are not exclusive (this explanation was developed post hoc). Rather, both anchoring and adjustment heuristics can be used when making a series of clinical judgments. More specifically, in most instances the adjustment hypothesis is operative; any initial anchoring effects dissipate in a succession of judgments (Bieri et al., 1963, p. 623). However, when

a group of practitioners are presented two sets of contrasting salient information (e.g., anorexia and psychological health), an anchoring effect may be evoked initially but is mitigated by an assimilation effect. Thus, the initial salient information may serve to anchor clinical judgments at first, however new contrasting salient information may induce judgments partially adjusted towards the new anchor (cf. Bieri et al., 1966). We named this the *adjustment mitigation hypothesis* (*mitigation* for short) because the anchoring heuristic is incorporated into the adjustment hypothesis. The mitigation hypothesis is based on one assumption (i.e., neutral information affects negligibly any subsequent judgments) and one testable proposition: When two sets of salient information are presented at different times in a succession of judgments, the salient information presented later will be half as influential as the salient information presented earlier. Although there is no specific theory or empirical evidence to justify a 50% reduction, this seems a reasonable reduction in the impact of new salient information if both heuristics are in operation (see Bieri et al., 1966; Klayman & Ha, 1987).

Study 2 was designed to test the three hypotheses, in addition to redressing the rival explanations and previous methodological problems. For example, in the four previous studies, there were no attempts to sample randomly, nor were the actual response rates reported; hence, their generalizability was suspect. Therefore, a random sample was drawn from a population of practicing psychologists in an attempt to replicate Friedlander and Stockman's (1983) discovery of anchoring bias. A pilot study was undertaken to estimate the magnitude of a nontrivial anchoring or adjustment effect (see Hayes & Haas, 1988). With this effect size an a priori power analysis was computed to ensure adequate statistical power. Finally, a salient, psychologically healthy manipulation was developed and validated for the anorexia case. This manipulation was designed to engender significantly different starting judgments and thus to circumvent the paradox encountered in previous investigations. Moreover, the inclusion of the healthy manipulation made it possible to test the multiple hypotheses.

Method

Participants

A total of 792 professionals were randomly selected from among the 16,500 listed by the Council for the National Register of Health Service Providers in Psychology (1985); 57 (7.2%) were deceased or had no forwarding address. Of the 735 remaining professionals, 157 returned completed materials, which resulted in a 21.4% response rate. The mean age of the volunteer participants (61.8% men and 38.2% women) was 46.83 (SD = 9.86) years old. The majority of the respondents were licensed psychologists (98.7%) who had a doctorate (96.2%) in either clinical (82.5%) or counseling psychology (13.4%). They saw an average of 17.46 clients (SD = 11.39) per week. Their theoretical orientations were psychoanalytic or psychodynamic (25.0%), cognitive behavioral (15.2%), or eclectic (45.9%).

In order to assess if respondents differed from those who chose not to participate in the study (i.e., sample representativeness), the two groups were compared on data published in the *National Register* (i.e., sex, geographic region, theoretical orientation, and diplomate status). Respondents did not differ from nonrespondents in terms of sex or diplomate status, ts(765) < 1.36, ps > .17, or in terms of geographic region or theoretical orientation, χ^2s(11) < 13.49, ps > .27. Thus, there was no evidence to suggest that the respondents were not representative of psychologists listed in the *National Register*.

Dependent Measures

Client pathology was construed as a multidimensional construct (Haase & Ellis, 1987) that consisted of client level of functioning (GAS; Endicott et al., 1976) and client prognosis after treatment (PS). Both of these measures were described fully in Study 1.

Case Materials

The anorexic case (Gina) was used because this case was the only one to yield anchoring effects (Friedlander & Stockman, 1983). The case materials were identical to those used in Study 1 except for two minor changes. The client's name was changed to Cathy to minimize recognition of a twice published case (Friedlander & Phillips, 1984; Friedlander & Stockman, 1983). Cathy was identified as an architect rather than a librarian in order to be not only more consistent with current career opportunities available to women but also conducive to the healthy manipulation.

A paragraph to present salient, psychologically healthy information was generated consistent with the case materials. The initial paragraph was modified on the basis of evaluations by three counseling psychologists. In short, the healthy paragraph presented the client as self-reliant, optimistic, and conscientious and as a survivor; it stated that she had had previous successful counseling, had a good support network with two long-term and very close friendships, was quite successful vocationally, maintained a regular exercise program, and rarely drank alcohol (the full manipulation is available from Michael V. Ellis). A panel of nine expert mental health professionals with 5 or more years of experience who worked in a psychiatric inpatient facility validated the salient healthy information. After the professionals first read all five interview summaries (sans the manipulations) and rated the client's level of functioning (GAS) and prognosis (PS), the healthy paragraph alone was read and subsequently rated. The results indicated that the experts made adjustments in the predicted directions for both the GAS (Mdn = 13.60) and PS (Mdn = 0.56). The panel of experts rated the case as realistic (1 = *completely realistic* to 9 = *completely unrealistic*; M = 3.00, SD = 2.35) and the manipulation as indicative of modest psychological health (1 = *not psychologically healthy* to 9 = *totally psychologically healthy*; M = 5.66, SD = 1.73).

Pilot Study

In order to facilitate an a priori power analysis, a separate panel of nine expert mental health practitioners from the community indicated the "minimal change in the ratings that you would consider to be clinically significant" (i.e., a nontrivial effect). The panel (5 women and 4 men) had an average of 8.21 years of experience (SD = 5.53). When this panel assumed a private practice setting, the smallest clinically meaningful change in a client's level of functioning (GAS) averaged 14.56 (SD = 7.47), and in a client's highest expected prognosis (PS), it averaged 1.56 (SD = 0.53). The corresponding estimates of the population effect sizes were $\hat{\rho}^2$ = .236 and .360, respectively, which were quite large by counseling psychology standards (Haase, Ellis, & Ladany, 1989). Hence, a more conservative effect size on the basis of the 9-point GAS and 0.5 PS mean differences (i.e., $\hat{\rho}^2$ = .10 for both scales) was used in the power analysis to yield

a minimal sample size of 150 for power of .80 (Cohen, 1988). As we anticipated a 20% response rate, 750 psychologists were randomly sampled from the *National Register*. A replacement was randomly sampled for those who were deceased or had no forwarding address.

Procedure

Participants were randomly assigned to one of six conditions in a 2 (time) × 2 (salient information) factorial design with two extra conditions (see Himmelfarb, 1975). The participants assigned to groups in the 2 × 2 design were presented one of the two types of salient information (anorexic or healthy) in either Interview 1 (early) or Interview 4 (late). The participants in the two extra groups were exposed to both types of salient information; either the healthy information early and anorexic information late, or in reversed order (anorexic early and healthy late). Thus, the six groups were: healthy early (HE; $n = 30$), healthy late (HL; $n = 22$), anorexic early (AE; $n = 24$), anorexic late (AL; $n = 28$), healthy early–anorexic late (HEAL; $n = 24$), and anorexic early–healthy late (AEHL; $n = 29$).

The anonymous questionnaire materials were mailed individually to the randomly drawn sample. Within 3–4 weeks, a follow-up reminder was sent. No other follow-up attempts were made. Participants followed the procedures described in Study 1 with one notable exception, which was made in order that the study be more consistent not only with Tversky and Kahneman's (1974) formulations but also with clinical practice. After reading each interview summary, the participants first classified the client's problem with an Axis I and Axis II diagnosis (American Psychiatric Association, 1987). Participants then completed the following: five filler questions to indicate the types of judgments about clients that clinicians typically make (e.g., recommended frequency of therapy sessions or extent to which referral for psychotropic medication was indicated), the GAS, and the PS. After rating all five interviews, the participants completed a demographic questionnaire.

Results

Manipulation Checks

A series of preliminary analyses with Interview 1 data served as manipulation checks (see Table 1). As required by the adjustment hypothesis, initial judgments between two groups must be the same. Thus, HE was compared to HEAL, AE to AEHL, and HL to AL because identical information was presented to each set of groups in the first interview. In order to minimize Type II errors, the per comparison alpha was .05. None of the three multivariate t tests revealed significant differences, Vs < .09, Fs(2, 50) < 2.58, ps > .09, $\hat{\rho}_m^2$s < .054; these results thus satisfied the equivalency requirement. An analysis was performed to test the anchoring hypothesis's requirement of differential starting values among comparison groups. Given their equivalence, the respective groups were subsequently aggregated (to maximize statistical power) and then subjected to a one-way MANOVA with Interview 1 data (Healthy = HE + HEAL; Anorexic = AE + AEHL; and Neutral = HL + AL). This omnibus test of differential starting judgments was significant, $V = .153$, $F(4, 308) = 6.37$, $p < .0001$, $\hat{\rho}_m^2 = .064$. Three follow-up multivariate t tests were conducted on the independent variables in order to assess the effectiveness of each manipulation. As expected, two of these analyses attained significance ($\alpha_{pc} = .05 \div 3 = .02$). Starting judgments for the healthy manipulation were significantly less pathological than ratings from the anorexia manipulation, $V = .182$, $F(2, 104) = 11.53$, $p = .0001$, $\hat{\rho}_m^2 = .166$. The group exposed to the healthy manipulation had significantly less pathological starting judgments than did those in the neutral condition, $V = .124$, $F(2, 101) = 7.16$, $p < .001$, $\hat{\rho}_m^2 = .107$. Although in the expected direction, the anorexia manipulation did not yield significantly different initial client pathology ratings than the neutral condition, $V = .019$, $F(2, 100) = 0.98$, $p = .38$, $\hat{\rho}_m^2 = .00$. Thus, the healthy manipulation yielded significantly divergent starting judgments. The anorexia manipulation was effective when compared to the healthy condition, but as expected, the starting judgments were not significantly lower than those for the neutral condition.

Table 1
Means and Standard Deviations for Global Assessment Scale and Prognosis Scale Ratings at Interviews 1 and 5 Across the Six Experimental Conditions

Condition	Interview 1		Interview 5	
	M	*SD*	*M*	*SD*
Global Assessment Scale				
HE	63.80	8.80	65.63	10.10
HL	57.55	6.88	64.09	6.23
AE	58.25	9.62	59.71	6.99
AL	60.39	5.47	59.29	6.38
HEAL	62.92	8.68	59.46	10.66
AEHL	56.45	7.90	62.28	8.58
Prognosis Scale				
HE	2.20	0.81	2.53	1.07
HL	2.68	0.72	2.36	0.73
AE	2.54	0.72	2.58	0.78
AL	2.64	0.68	2.89	0.74
HEAL	2.17	0.76	2.38	0.88
AEHL	3.03	0.82	2.93	0.65

Note. For the Global Assessment Scale (Endicott, Spitzer, Fleiss, & Cohen, 1976), higher scores indicate mental health. The Prognosis Scale (Friedlander & Stockman, 1983) is rated on a scale from 1 = *superior* to 7 = *grossly impaired*. HE = healthy early; HL = healthy late; AE = anorexia early; AL = anorexia late; HEAL = healthy early–anorexia late; and AEHL = anorexia early–healthy late.

Major Analyses

As in Study 1, Interview 5 judgments were tested. A one-way MANOVA (6 groups) yielded significant results, $V = .16$, $F(10, 302) = 2.59$, $p = .005$, $\hat{\rho}_m^2 = .050$, which indicate that the treatments differentially affected ratings of client pathology (Himmelfarb, 1975). Three follow-up multivariate t tests were performed for each of the anchoring and adjustment hypotheses. Even though these were planned comparisons, a Bonferroni procedure was used to establish the familywise Type I error rate for the two sets of analyses ($\alpha_{pc} = .05 \div 3 = .02$).

Anchoring hypothesis analyses. Three analyses were performed to test the hypothesis that anchoring errors occur in clinical judgments. That is, two groups with significantly

different starting judgments ought to have significantly different final judgments even though both were presented the same information. If anchoring occurred, then the HEAL group would have significantly better client pathology ratings than the AEHL group, HE better than HL, and AE worse than AL. The first multivariate t test found that the HEAL group judged the client as significantly less pathological than the AEHL group, $V = .16$, $F(2, 50) = 4.82$, $p = .012$, $\hat{\rho}_m^2 = .128$. It is important, however, to note the pattern of these results (see Table 1), which is contrary to an anchoring or mitigation effect. Initially the HEAL group rated the client significantly less pathological (GAS) than the AEHL group, yet at the final judgments the pattern was reversed (HEAL > AEHL). The multivariate tests to assess the predicted differences between the HE and HL groups, $V = .03$, $F(2, 49) = 0.84$, $p = .44$, $\hat{\rho}_m^2 = .00$, and the AE versus the AL groups, $V = .04$, $F(2, 49) = 1.08$, $p = .35$, $\hat{\rho}_m^2 = .0004$, were both nonsignificant.

Adjustment hypothesis analyses. Three analyses were conducted to test the hypothesis that practitioners adequately adjust their judgments. Specifically, this required that two groups presented the same information early (i.e., same starting judgments) and different information later would have significantly different final judgments (i.e., appropriate adjustments were made). Thus, the HL group ratings were predicted to be less pathological (better) than AL group judgments, HE better than HEAL, and AE worse than AEHL. As expected, the HL group had significantly better ratings than the AL group, $V = .20$, $F(2, 47) = 5.95$, $p = .005$, $\hat{\rho}_m^2 = .167$. One marginally significant difference was found: More client pathology was reported by the HEAL group than the HE group, $V = .13$, $F(2, 51) = 3.64$, $p = .033$, $\hat{\rho}^2 = .096$. The predicted effects between the AE and AEHL groups did not achieve statistical significance, $V = .098$, $F(2, 50) = 2.70$, $p = .077$, $\hat{\rho}_m^2 = .062$. As can be seen in Table 1, however, the pattern of GAS means reversed from Interview 1 to 5, which indicates an adjustment in the judgments of the AEHL group.

Mitigation hypothesis analyses. Assessing the mitigation hypothesis entailed specifying a set of predicted effects on the basis of a series of comparisons and then measuring the degree of association between the predicted effects and the observed effects. For the purposes of this post hoc analysis, we adopted "+" to indicate psychological health and "−" to indicate psychopathology. Furthermore, coefficients were assigned and effects predicted according to the mitigation hypothesis (see Tables 2 and 3; the choice of 10 was arbitrary as any real number would yield identical results). For example, AE presented anorexia information in Interview 1 (−10), followed by four neutral sessions (0 + 0 + 0 + 0), which when summed yield a −10 coefficient for this condition. AEHL had initial anorexic information (−10), followed by two neutral (0 + 0), one healthy reduced by 50% (10 ÷ 2 = +5), and one neutral (0) interview, which gives a coefficient of −5. Because AEHL was expected to be greater than AE, the predicted effect value was obtained by subtracting the coefficient for AE from the AEHL coefficient [−5 − (−10) = 5]. Because we pursued a conservative data analysis approach, Kendall's tau corrected for ties (Marascuilo & McSweeney, 1977, pp. 444–446) was computed, and the results yielded a strong correlation between the predicted effects and the observed effects: $\tau = .93$, $p < .01$.

Table 2
Coefficients (Weights) Assigned to Conditions for the Mitigation Hypothesis

Condition	Coefficient
Anorexia information	−10
Neutral information	0
Healthy information	10
AE	−10 + 0 = −10
AL	0 + (−10) = −10
HE	10 + 0 = 10
HL	0 + 10 = 10
AEHL	−10 + 0.5(10) = −5
HEAL	10 + 0.5(−10) = 5

Note. The coefficient for a condition was obtained by summing weights for each of the five interviews in the case material. For AEHL and HEAL, the weight for the late manipulation was mitigated by 50%. AE = anorexia early; AL = anorexia late; HE = healthy early; HL = healthy late; AEHL = anorexia early–healthy late; and HEAL = healthy early–anorexia late.

Table 3
Mitigation Hypothesis Predicted Effects and Observed Effects

Comparison	Predicted effect	Observed effect ($\hat{\rho}_m^2$)
HL > AL	10 − (−10) = 20	.17
HEAL > AEHL	5 − (−5) = 10	.13
HE > HEAL	10 − 5 = 5	.10
AEHL > AE	−5 − (−10) = 5	.06
AE = AL	−10 − (−10) = 0	.001
HE = HL	10 − 10 = 0	.00

Note. Predicted effects were obtained by taking the difference between the coefficients corresponding to the two conditions being compared. $\hat{\rho}_m^2$ = estimated multivariate population effect size (shrunken η_m^2) from data analysis. HL = healthy late; AL = anorexic late; HEAL = healthy early–anorexic late; AEHL = anorexic early–healthy late; HE = healthy early; and AE = anorexic early.

General Discussion

The larger question underlying the two studies presented in this article was: How does a client's disclosure of salient information affect the practitioner's judgments of client pathology? Study 2 was designed to test multiple hypotheses about clinical judgments. So what reasonable inferences can be drawn from the results of Studies 1 and 2? Two conclusions seem justifiable.

When taken collectively, virtually no evidence was found in either study to support the anchoring hypothesis as explicated by Tversky and Kahneman (1974). In fact, a good deal of evidence accrued to disconfirm the anchoring hypothesis in clinical judgment (Klayman & Ha, 1987). Specifically, one significant (HL > AL) and one marginally significant (HE > HEAL) adjustment effect emerged, and although one anchor-

ing hypothesis comparison attained statistical significance (HEAL > AEHL), the reversal in the GAS judgments seemed more indicative of an adjustment effect. That is, the inversion in judgments was opposite to what might be expected had anchoring or mitigation occurred. With Study 2 as a more powerful, more rigorous, yet unsuccessful replication, the cumulative evidence (Friedlander & Phillips, 1984; Levin, 1984) suggested that Friedlander and Stockman's (1983) anchoring effect was most likely spurious and attributable to Type I error or sampling vicissitudes.

The second tentative conclusion concerns the newly formulated adjustment and mitigation hypotheses (recall that the mitigation hypothesis differs from the adjustment hypothesis only when two sets of salient information are presented within a case). Although the results tended to confirm both hypotheses, some evidence was found to disconfirm the adjustment hypothesis. That is, much of the data suggested that counselors did adjust their clinical judgments appropriately in response to one set of salient client information, but when two sets of contrasting salient information were presented, the counselors' judgments were more often consistent with the mitigation hypothesis than the adjustment hypothesis. The adjustment effect of the second set of salient information appeared to be mitigated by an anchoring bias toward the first set (cf. Bieri et al., 1963). This in conjunction with the strong correspondence between the predicted and observed effects seems to support the mitigation hypothesis as viable and parsimonious.

Indeed, the adjustment mitigation hypothesis has intuitive appeal. Counselors are expected to form hypotheses about clients and adjust them as new information is acquired (Pepinsky & Pepinsky, 1954; Strohmer & Chiodo, 1984). A counselor's failure to take into account new salient client information (e.g., psychological health) can be potentially hazardous to the client (e.g., Rosenhan, 1973), and thus adjustment in this context is appropriate and desirable. Furthermore, when new client information contrasts with previous information, the counselor apparently integrates the new data into the existing conceptualization of the client and judges the client accordingly. Hence, the conceptualization of the client is not altered dramatically with each piece of new information (Klayman & Ha, 1987).

Because the mitigation hypothesis was developed post hoc, it may be considered tentative and needs further testing. That is, given that counselors are exposed potentially to multiple sets of salient information for a client during the course of treatment (e.g., test data, client statements, or counselor observations), why and how may subsequent sets of information affect clinical judgments? The results of Study 2 suggest incorporating repeated measures or within-subjects designs to clarify the nature of mitigation effects. For instance, mitigation effects may be attributable partially to a confirmatory bias (Lopez, 1989; Snyder, 1981). Although considered conceptually distinct, mitigation effects also resemble assimilation effects (Bieri et al., 1966; Bieri et al., 1963; Campbell et al., 1957; Sherif & Hovland, 1961), and assimilation effects yield biased judgments. The extent to which adjustment mitigation effects yield biased judgments, however, remains open for investigation.

The conclusions based on the research presented here have limitations. For example, though Lopez (1989) argued convincingly for the necessity of analogue studies of clinical judgment, our conclusions are tempered by the analogue design. Whereas no evidence was found to indicate that our sample was not representative of psychologists listed in the *National Register,* the 21% response rate suggests caution in generalizing to this population. Of course, the inferences in regard to the adjustment and mitigation hypotheses require empirical verification.

While Study 2 was underway, two articles that investigated anchoring errors in clinical judgments appeared in the literature (Pain & Sharpley, 1988, 1989). In both studies Pain and Sharpley used client cases in which each case consisted of four reports, two neutral (N), one good (G), and one bad (B), presented in four different orders (i.e., valence order; GNNB, NGBN, NBGN, and BNNG). Thus, for each case the G and B reports were each rated at the four times by different participants. Although not readily obvious, the data for both studies were apparently incorrectly analyzed; within each case GAS ratings of only the four G (or B) reports (Times 1–4) were compared (i.e., a confounding of the valence order and time factors) rather than testing Time 4 judgments across the valence factor. Hence, the results were confounded not only by the prior amount of information presented (i.e., each group was exposed to different types and amounts of information) but also by the prior number of ratings. Without a complete reanalysis of the data, these confounds cannot be disentangled. As such, the interpretations and conclusions that Pain and Sharpley derived from their results were highly suspect. Therefore their results and conclusions were not applicable to the types of anchoring we investigated.

As in any study, rival explanations emerge that may account for the pattern of results that were found (Cook & Campbell, 1979). For Studies 1 and 2, several rival hypotheses seem sufficiently plausible to warrant investigation (Serlin & Lapsley, 1985). Bieri et al. (1963) reasoned "that differences in findings and thus of [anchoring] theory may be . . . a function of method differences" (p. 616). Given the numerous definitions and operationalizations of anchoring (e.g., Bieri et al., 1966; Sherif & Hovland, 1961), perhaps different types of anchoring lead to biased clinical judgments (e.g., contrast and assimilation effects). Rather than bipolar descriptors (e.g., Bieri et al., 1966), the dependent measures were behaviorally anchored at each level, which may reduce the degree of uncertainty of judgments and avert a stronger bias toward the anchor. In addition, using single item dependent variables (as in Studies 1 and 2) can also hinder the ability to detect anchoring errors. Another possibility is that counseling situations entail more risk associated with clinical judgments not only for the client but for the practitioner as well (e.g., Friedlander & Stockman, 1983). Thus, the sheer nature of the manipulations and their consequences may be so salient that they mitigate or even prohibit anchoring errors (see Klayman & Ha, 1987). These explanations seem readily testable and worthy of study. Indeed, continued research endeavors to understand the process by which practitioners make judgments about clients remains a potentially fruitful field of inquiry.

References

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (Rev. 3rd ed.). Washington, DC: Author.

Arkes, H. R. (1981). Impediments to accurate clinical judgment and possible ways to minimize their impact. *Journal of Consulting and Clinical Psychology, 49,* 323–330.

Bieri, J., Atkins, A. L., Briar, S., Leaman, R. L., Miller, H., & Tripodi, T. (1966). *Clinical and social judgment.* New York: Wiley.

Bieri, J., Orcutt, B. A., & Leaman, R. (1963). Anchoring effects in sequential clinical judgments. *Journal of Abnormal and Social Psychology, 67,* 616–623.

Campbell, D. T., Hunt, W. A., & Lewis, N. A. (1957). The effects of assimilation and contrast in judgments of clinical materials. *American Journal of Psychology, 70,* 347–360.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston, MA: Houghton-Mifflin.

Council for the National Register of Health Service Providers in Psychology. (1985). *National register of health service providers in psychology.* Baltimore: Author.

Endicott, J., Spitzer, L., Fleiss, J. L., & Cohen, J. (1976). The Global Assessment Scale: A procedure for measuring overall severity of psychiatric disturbance. *Archives of General Psychiatry, 33,* 766–771.

Friedlander, M. L., & Phillips, S. D. (1984). Preventing anchoring errors in clinical judgment. *Journal of Consulting and Clinical Psychology. 52,* 366–371.

Friedlander, M. L., & Stockman, S. J. (1983). Anchoring and publicity effects in clinical judgment. *Journal of Clinical Psychology, 39,* 637–643.

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

Haase, R. F., & Ellis, M. V. (1987). Multivariate analysis of variance. *Journal of Counseling Psychology, 34,* 404–413.

Haase, R. F., Ellis, M. V., & Ladany, N. (1989). Multiple criteria for evaluating the magnitude of experimental effects. *Journal of Counseling Psychology, 36,* 511–516.

Hayes, S. C., & Haas, J. R. (1988). A reevaluation of the concept of clinically significance: Goals, methods, and methodology. *Behavioral Assessment, 10,* 189–196.

Himmelfarb, S. (1975). What do you do when the control group doesn't fit into the factorial design? *Psychological Bulletin, 82,* 363–368.

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94,* 211–228.

Levin, R. (1984). *Differential anchoring effects in clinical judgment.* Unpublished doctoral dissertation, State University of New York at Albany.

Lopez, S. R. (1989). Patient variable biases in clinical judgment: Conceptual overview and methodological considerations. *Psychological Bulletin, 106,* 184–203.

Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences.* Monterey, CA: Brooks/Cole.

Meehl, P. E. (1954). *Clinical versus statistical prediction.* Minneapolis: University of Minnesota Press.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Pain, M. D., & Sharpley, C. F. (1988). Case type, anchoring errors, and counselor education. *Counselor Education and Supervision, 28,* 53–58.

Pain, M. D., & Sharpley, C. F. (1989). Varying the order in which positive and negative information is presented: Effects on counselors' judgments of clients' mental health. *Journal of Counseling Psychology, 36,* 3–7.

Pepinsky, H. B., & Pepinsky, P. (1954). *Counseling theory and practice.* New York: Ronald Press.

Rosenhan, D. L. (1973). On being sane in insane places. *Science, 179,* 250–258.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40,* 73–83.

Sherif, M., & Hovland, C. I. (1961). *Social judgment.* New Haven, CT: Yale University Press.

Snyder, M. (1981). On the self-perpetuating nature of social stereotypes. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 183–212). Hillsdale, NJ: Erlbaum.

Strober, M. (1986). Anorexia nervosa: History and psychological concepts. In K. D. Brownell & J. O. Foreyt (Eds.), *Handbook of eating disorders: Physiology, psychology, and treatment of obesity, anorexia, and bulimia* (pp. 231–246). New York: Basic.

Strohmer, D. C., & Chiodo, A. L. (1984). Counselor hypothesis testing strategies: The role of initial impressions and self-schema. *Journal of Counseling Psychology, 31,* 510–519.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.

Received June 23, 1989
Revision received November 22, 1989
Accepted November 29, 1989 ■