Journal of Personality and Social Psychology
1990, Vol. 58, No. 2, 319-329

0022-3514/90/$00.75

Reducing the Dependence of Clinical Judgment on the Immediate Context: Effects of Number of Categories and Type of Anchors

Douglas H. Wedell
University of Illinois

Allen Parducci and
Michael Lane
University of California, Los Angeles

Two experiments explored methods for standardizing ratings of the psychopathology of clinical case histories. In both experiments, the same case histories were rated as more pathological when mostly mild rather than severe cases were presented as the immediate context. Psychometric analyses demonstrated that this type of contextual effect is a potentially important source of unreliability in clinical judgment. In Experiment 1, increasing the number of points in the rating scale from 3 to either 7 or 100 significantly reduced the effects of the immediate context. Ratings were parsimoniously modeled by Parducci's (1983) range-frequency theory. In Experiment 2, providing verbal anchors in the form of either detailed *DSM* descriptions for each rating category or sample case histories for the two end-categories increased the reliability of the ratings by reducing the effects of the immediate contexts; however, these reductions occurred only when the ranges of the immediate contexts had been severely restricted. According to the range-frequency analysis, verbal anchors served to equate the endpoints of the subjective range for the different contextual conditions. Comparison with previous research suggests that the anchors also reduced the effects of the sequential position in which clinical cases appear. We therefore recommend that studies of the reliability of behavioral assessment techniques take into account the effects of differences in context.

Although reliability does not imply validity, it is a necessary prerequisite. Unfortunately, clinical assessment in psychology is notoriously unreliable. Eysenck, Wakefield, and Friedman (1983) have documented the lack of interrater reliability. Goldberg (1968, 1970) concluded that clinical judgments tend to be unreliable and only minimally related to the degree of experience of the person making the assessment. Arnoff (1954) reported an actual decrease in reliability with greater clinical experience.

One generally accepted method for improving clinical prediction has been the use of statistical models. In particular, there is overwhelming evidence for the superiority of simple linear models over clinicians in combining information to form a diagnosis (Dawes, 1979; Dawes & Corrigan, 1974; Einhorn, 1972; Goldberg, 1968, 1970; Meehl, 1954). But as Sawyer (1966) has pointed out, combining information is only the second half of the problem, the first half being the problem of measurement. Although there are a large number of objective measurement devices currently in use in clinical psychology, it is doubtful that the clinician will be supplanted in the near future as a primary source of information used in prediction. Therefore, development of greater accuracy of clinical judgments is of fundamental importance. Regardless of whether the judgments are combined intuitively by the clinician or statistically by linear regression, unreliability of the information limits diagnostic validity.

The two experiments reported here attempt to reduce differences in ratings resulting from the use of different clinical standards by different judges. A key to the consistent use of any measuring instrument is that it is calibrated in the same way on each application: Reliability of judgment is maximized when different judges apply the same standards. However, there is overwhelming evidence in the literature of psychophysics and social judgment (e.g., Eiser & Stroebe, 1972; Johnson, 1972; Parducci, 1983; Poulton, 1979; Upshaw, 1969; Wedell & Parducci, 1988) that standards are largely determined by characteristics of the particular set of stimuli being judged. Consequently, persons judging different sets of stimuli tend to employ different standards.

Clinical judgment is subject to these same effects of context. In an extensive unpublished study, Perrett (1971) presented a series of abstracted case histories for assessment of psychopathology. Moderate test cases were judged more severely when they appeared among milder cases than when they appeared among more severe cases. This type of shift of ratings away from the values of contextual stimuli is typically referred to as a contrast effect and has been demonstrated repeatedly in judgments of clinical materials (e.g., Bieri, Orcutt, & Leaman, 1963; Campbell, Hunt, & Lewis, 1957; Manis & Paskewitz, 1984a, 1984b; Manis, Paskewitz, & Cotler, 1986). Despite their greater familiarity with clinical case histories, Perrett found that clinicians were just as susceptible to contextual effects as were untrained undergraduate students. This suggests that the same case will be judged as more pathological by a clinician who typically encounters relatively mild problems (e.g., an outpatient

Preparation of this article was carried out while Douglas H. Wedell was at the University of Illinois on a postdoctoral traineeship, Alcohol, Drug Abuse, and Mental Health Administration National Support Award MH14257, Lawrence Jones, training director.

We thank James Austin and Ulf Bockenholt for their helpful comments.

Correspondence concerning this article should be addressed to Douglas H. Wedell, who is now at Department of Psychology, University of South Carolina, Columbia, South Carolina 29209.

therapist) than by one who is exposed to severe, chronic cases (e.g., an inpatient therapist).

In light of the potentially deleterious consequences of shifting standards within a clinical setting, it seems important to investigate ways in which effects of the immediate context can be minimized. Development of judgmental measures that are stable across different contextual settings would enhance the reliability, and potentially the validity, of clinical judgment. Given the ubiquitous use of rating scales across a wide spectrum of applications, including survey research, performance appraisals in industry, and personality assessment, techniques for reducing unwanted contextual shifts may prove generally useful. Toward this end, Experiment 1 investigated whether the effects of context on ratings of psychopathology can be reduced by increasing the number of judgmental categories. For psychophysical stimuli, contextual contrast has been found to decrease with increasing numbers of rating categories (Parducci, 1982; Parducci & Wedell, 1986). The same *category effect* has been demonstrated using more complex social stimuli (Wedell & Parducci, 1988). These studies suggest that increasing the number of categories used to judge psychopathology should reduce interrater unreliability resulting from the different standards that are evoked by different sets of cases.

Another way to reduce contextual effects might be through instructional manipulations designed to firmly anchor the response categories to particular case histories. A variety of techniques for anchoring rating categories with behavioral descriptions have been developed in organizational psychology. Although these techniques often lead to higher reliability (Bernardin & Beatty, 1984), there is no indication that they reduce the effects of differences of the immediate context for judgment. In previous research on clinical judgment, use of detailed descriptions of each category has met with little success in reducing contextual effects (Arnoff, 1954; Perrett, 1971); however, the results of Experiment 1 suggested circumstances under which anchoring might prove effective. Experiment 2 investigated these conditions using two different techniques for anchoring the scale of judgment.

The strong contextual effects observed in both experiments are modeled using Parducci's (1965) range–frequency theory of judgment. Thus, before proceeding, we first present an overview of that theory and how it predicts that judgments of psychopathology will be affected by contextual manipulations.

Range–Frequency Theory

Contextual effects have been studied most extensively using psychophysical stimuli where the judgment of each stimulus appears to be determined by the entire set of stimuli presented in the experimental session. Parducci's (1965, 1983) range–frequency theory proposes that judgments reflect two principles of judgment. According to the range principle, the value of a stimulus is determined by the proportion of the subjective range of stimuli lying below it (cf. Volkmann, 1951). The range value, R_{ic}, of stimulus i in context c, can be expressed algebraically as:

$$R_{ic} = (S_i - S_{\min})/(S_{\max} - S_{\min}), \quad (1)$$

where S_i is the subjective value of the stimulus and $S_{\max}$ and $S_{\min}$ are the maximum and minimum subjective values defining context c. The range principle accounts for differences in judgments when contexts have different end-stimuli. For example, a clinician who has recently experienced a set of very disturbed individuals (as in an inpatient setting) would be expected to judge a moderate case history as exhibiting only mild pathology because that case history is among the most mild in the experienced range. Conversely, a clinician in an outpatient setting who works mostly with individuals exhibiting mild pathologies would be expected to judge the same moderate case as fairly extreme in pathology because it is among the most extreme cases he or she experiences.

A second way in which judgment can be affected by the contextual set is described by the frequency principle, according to which the value of a stimulus is determined by the proportion of the total number of contextual stimuli lying below it on the dimension of judgment (i.e., its percentile rank in the contextual set). Expressed in parallel form to Equation 1, the frequency value of stimulus i in context c is then:

$$F_{ic} = (r_{ic} - 1)/(N_c - 1), \quad (2)$$

where r_{ic} is the rank of stimulus i in context c and N_c is the total number of stimuli in the context. The frequency principle accounts for differences in judgment when contexts are defined by the same end-stimuli but differ in the relative frequency or spacing of the stimuli. Thus, even if two clinicians experience the same range of psychopathology, their judgments may still differ systematically. If one works with predominantly mild cases, a moderate case may rank among the most severe experienced and hence be judged as severely disturbed. If the other works with predominantly severe cases, that same moderate case may rank among the most mild experienced and hence be so judged.

Range–frequency theory proposes that internal judgments represent a compromise between range and frequency principles. The judgment, J_{ic}, of stimulus i in context c may be described algebraically as a weighted average of range and frequency values:

$$J_{ic} = wR_{ic} + (1 - w)F_{ic}, \quad (3)$$

where w is the relative weighting of the two values. The overt category ratings are assumed to be linearly related to internal judgments:

$$C_{ic} = bJ_{ic} + a, \quad (4)$$

where C_{ic} is the numerical rating, b is the range of categories (i.e., 5 for a 6-category scale), and a is the number representing the lowest category.

The range–frequency model has provided good fits to data from a wide variety of psychophysical experiments (for a review see Parducci, 1983). Good fits have also been obtained for judgments of various social dimensions, such as ratings of performance (Mellers & Birnbaum, 1983; Wedell, Parducci, & Roman, 1989), equity (Mellers, 1983, 1986), physical attractiveness (Wedell, Parducci, & Geiselman, 1987), and happiness (Smith, Diener, & Wedell, 1989; Wedell & Parducci, 1988).

Implications for Reliability of Judgment

There are many senses of reliability that have been described in the psychometric literature (for a discussion of the multiple levels at which reliability may be assessed, see Cronbach, Gleser,

& Rajaratnam, 1963). One sense of reliability is captured in terms of the correlation between repeated assessments of the same individuals. Because correlation is invariant under linear transformations of scale, reliability in this sense should be relatively insensitive to range–frequency effects. For example, even though the range values for the same stimuli may vary greatly across contexts, they will be perfectly correlated. Furthermore, although frequency values derived from different contexts will typically be nonlinearly related, they will always be monotonically related, and hence reliability in the correlational sense will be relatively unaffected (see Labowitz, 1970, for a demonstration of the stability of correlation under monotonic transformations).

In a diagnostic setting, however, judgmental categories are being applied in an absolute rather than a relative sense. It is not acceptable for one judge to rate a person's behavior as *moderately maladaptive* and another judge to rate that same behavior as *severely maladaptive,* even though the judges' ratings may be linearly related. Thus, traditional measures of interrater reliability, such as coefficient kappa (Cohen, 1960), measure the degree to which raters assign the same stimulus to the same category. Following this definition, it is possible for the ratings of two judges to be perfectly correlated and yet the interrater reliability be at zero. It is this sense of reliability (i.e., interrater agreement) that can be drastically attenuated by the judgmental shifts described in range–frequency theory.

If one assumes that range–frequency theory provides a fairly accurate description of the judgment process, then one must reject the notion that judgments can be made context free. Instead, the key to eliminating differences in standards is through equating the subjective contexts on which the judgments are based. One trivial way to do so is to equate the overt set of contextual events experienced by different judges. In many ways, traditional studies of reliability do just that by having the different judges rate the same set of stimuli. However, in real-world settings, clinicians will be exposed to widely differing sets of individuals and yet they must be able to accurately communicate with one another.

A range–frequency analysis suggests that a major step toward standardizing judgment could be achieved if the judges under different contextual conditions evoked the same extreme end-stimuli that define the subjective range. Experiments 1 and 2 explored three different techniques for equating the range. First, the range of cases presented was directly manipulated so that some subjects judged case histories at both extremes while others judged restricted sets. Second, the number of categories was varied to investigate whether judgments would be less tied to the immediate range of stimuli when using a large number of categories, as has been demonstrated in previous research (Wedell & Parducci, 1988). Finally, detailed descriptions were provided for different rating categories to determine whether these could effectively define the endpoints of the range.

Even if the subjective range were equated for all judges, effects of the immediate context would still occur via the frequency principle. These effects could be eliminated if the weighting of the frequency principle were zero (i.e., $w = 1.0$). Thus, according to range–frequency theory, procedures that fix the range at predetermined values and eliminate the weighting of the frequency principle should produce the same judgmental standards, regardless of the immediate context for judgment. One factor known to affect the weighting of frequency values is the number of judgmental categories (Parducci & Wedell, 1986; Wedell & Parducci, 1988). Experiment 1 explored effects of manipulating this factor.

Experiment 1: Testing for the Category Effect

Within the framework of range–frequency theory, the reduction in the effects of context with increasing number of categories (i.e., the *category effect*) is most simply described in terms of changes in the weighting of the frequency principle, $1 - w$, which has varied from 0.87 for two categories to 0.07 for a 100-point scale (Parducci & Wedell, 1986). In terms of actual ratings of stimuli common to different distributions, the effects of skewing the distribution of contextual stimuli were highly significant, more than half the range of the scale, when subjects used just two categories, but were scarcely discernible and not statistically significant with a 100-point scale. Thus, increasing the number of categories in clinical judgment may significantly reduce unreliability resulting from use of different, contextually generated standards.

However, an increase in the number of rating categories does not always result in a decrease in the effects of context. In psychophysical research, the category effect is found only when the contextual distribution consists of stimulus values occurring with unequal frequencies. The category effect disappears when stimuli are presented with equal frequency (i.e., when skewing of the contextual distribution is manipulated by altering the spacing of the stimuli along the dimension of judgment, but not their frequencies). Thus, increasing the number of categories is like reducing the differences between the stimulus frequencies (Parducci & Wedell, 1986; Wedell & Parducci, 1985, 1988).

The dependence of the category effect on unequal stimulus frequencies raises some doubts about its applicability to ratings of more complex stimuli. For example, in evaluating different case histories, it is doubtful that any two cases would occupy exactly the same position on the dimension of judgment, and hence the contextual distribution could be characterized as a set of different stimulus values, each occurring with equal frequency. However, given the limitations on human information processing (e.g., Miller, 1956), it might be expected that similar cases would be grouped together to create, in effect, a distribution with unequal frequencies and thus conducive to the category effect. This latter expectation is consistent with the recent demonstration of the category effect for ratings of complex social stimuli (the happiness of life situations; Wedell & Parducci, 1988). Therefore, it seems likely that increasing the number of rating categories should at least partially reduce contextual effects for judgments of clinical case histories, even when no case is repeated.

Recent experimental research on social judgments has also demonstrated a second way in which the number of categories moderates contextual effects (Wedell & Parducci, 1988; Wedell et al., 1987). Because of the relative familiarity of social stimuli, it seems only natural that their context should cover a range that extends beyond the restricted set actually presented for judgment. However, when limited to just two or three rating categories, reserving end-categories for stimuli of more extreme value than those actually presented limits the ability of subjects to discriminate between stimuli within the immediate set. This

Table 1
Number of Cases at Each Prescaled Level of Psychopathology

Condition	Prescaled level[a]						
	1	2	3	4	5	6	7
Restricted range							
Mild	7	5	4	2	2	0	0
Severe	0	0	2	2	4	5	7
Full range							
Mild	6	4	3	2	2	0	3
Severe	3	0	2	2	3	4	6

[a] Based on median of ratings by five clinicians (Perrett, 1971).

suggests that fewer categories should result in greater matching of the range of response categories to the range of experimental stimuli, a prediction confirmed for ratings of physical attractiveness (Wedell et al., 1987) and ratings of happiness of facial expressions and of life events (Wedell & Parducci, 1988). Thus, increasing the number of categories used to rate clinical case histories might also reduce the effects of manipulating the range of stimuli actually presented for judgment.

Experiment 1 varied the number of categories for ratings of psychopathology. In order to explore the two effects related to the number of categories, two contextual manipulations were employed: (a) The relative frequencies of cases in different parts of the range were manipulated so that either mild or severe cases predominated in contexts with identical endpoints, or (b) endpoints were also varied to add to the effects of varying frequencies. Because the experimental design required a large number of subjects and previous research using the same case histories (Perrett, 1971) failed to find a significant difference in the magnitude of contextual effects between experienced clinicians and college undergraduates, the present study sampled only the latter population.

Method

Design and subjects. The experiment used a $3 \times 2 \times 2 \times 6$ factorial design, with three between-subjects factors: number of categories (3, 7, or 100), context (mild or severe), and range (full or restricted). The one within-subject factor was test cases (six moderate case histories common to all conditions). The ratings of the six test cases constituted the dependent variable.

Subjects, 503 undergraduates from the University of California, Los Angeles, participated to fulfill a course requirement. After first participating in one of a number of short psychophysical experiments, subjects were randomly assigned to 1 of the 12 experimental conditions and tested in groups of 8 to 12.

Case histories. Thirty-four condensed case histories of psychiatric patients in actual treatment were taken from the Perrett (1971) study (see Appendix A for the six test cases). Perrett had selected these 34 from a set of 80 case histories initially rated by five clinical psychologists. The six test cases had received median ratings between 3 and 5 on a 7-point scale during the initial scaling. Table 1 summarizes the different contextual conditions. It should be noted that endpoints and relative frequencies were both manipulated in the restricted-range conditions, but only the relative frequencies were manipulated in the full-range conditions.

Case histories were numbered and typed on three pages of an experimental booklet. To maximize contextual effects, the 14 cases constituting the manipulated context appeared first. In the full-range condition, the three extreme cases of opposite value appeared in positions 4, 8, and 11. The six test cases were randomly assigned to positions 15 through 20 (the same order occurring in all conditions).

Instructions. Instructions were printed at the top of a separate response sheet and were read aloud by the experimenter while subjects followed along. These instructions stated that the experiment was concerned with how people rate degree of mental disturbance or behavioral maladjustment on the basis of short case histories. A rating scale was printed at the top of the response sheet with the lowest category (1 or 0) labeled *very, very mild disturbance* and the highest category (3, 7, or 100) labeled *very, very severe disturbance.* Subjects were instructed to read each case history and mark the number (no fractions allowed) corresponding to their rating on the response sheet. Subjects were also asked not to compare case histories but rather to apply their own standards. Following Perrett (1971), instructions stated:

> You may find it difficult to rate some items because of a lack of information. However, make a quick assessment even when you are reluctant to do so. All relevant symptoms for the rating task have been included in each synopsis so that if a particular symptom is not included in the history, please assume that it is not present. None of these cases is complicated by mental deficiency or any known organic condition. (p. 73)

Results

Fit of the range–frequency model. Details of how the data were fit by the range–frequency model are presented in Appendix B. These procedures are similar to those used in previous experiments (Wedell & Parducci, 1988; Wedell et al., 1987). The major assumptions made in fitting the data were as follows: (a) Frequency values were based on a single ordering of cases averaged across conditions; (b) a single best-fit value of w was calculated for each number of categories on the basis of judgments of the six test cases in the full-range conditions, and (c) a single set of scale values for the 34 cases was calculated using inferred range values from 7- and 100-point scales (the 3-point estimates were too unreliable because of the low value of w).

The top two rows of Figure 1 show the mean ratings of the case histories along with the fit of the range–frequency model. The range functions and the values of the frequency weighting $(1 - w)$ used to generate the fits are shown in the bottom panels. The generally close adherence of the data points to the theoretical functions indicates that the model fits the data well. The squared correlation between the predicted values and the mean ratings is 0.966. Approximately 82% of the means deviate from the predicted values by less than 2 standard errors.

As predicted by range–frequency theory, the large differences in the mean ratings of the case histories are all in the direction of contrast: Mean ratings of the six test cases are higher when the contextual set consists mostly of milder cases. This effect is very large for restricted-range sets (top row of Figure 1). For example, the most severe case in the restricted-range, mild context (Test Case 6) is rated 78 on a 100-point scale; however, the same case is rated 33 in the corresponding severe context, a difference of nearly half the range of the rating scale. As implied by range–frequency theory, these effects of immediate context are greatly reduced (less than half as large) when the full range is presented (second row). For example, the ratings of Test Case 6 on the 100-point scale in the full-range condition were 60 and 40 in mild and severe contexts, respectively. Thus, merely

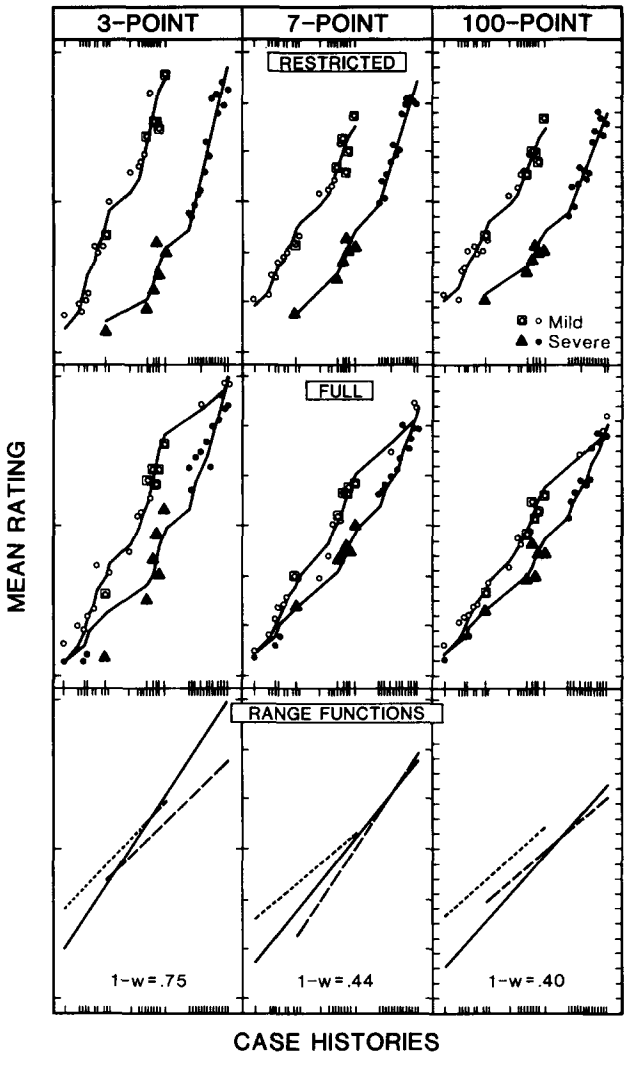

Figure 1. Contextual effects on ratings of psychopathology (Experiment 1). Contrast is represented by the higher mean ratings of case histories when context is mild (open symbols) than when context is severe (closed symbols). Target cases are enclosed with a square (mild context) or solid triangle (severe context). Lines represent the range–frequency model predictions for restricted- (top row) and full-range (middle row) conditions. Bottom panels show values of $1 - w$ and inferred range functions: restricted-range, mild context (dotted lines); restricted-range, severe context (dashed lines); and full-range mild and severe contexts (solid lines).

equating the range of stimuli in the immediate context can greatly reduce the magnitude of contextual effects.

The differences in the fitted values of the frequency weighting ($1 - w$) indicate the presence of the category effect. In particular, frequency values appear to receive much greater weight when ratings are made using three categories than when 7- or 100-point scales are used. Although the frequency weighting ($1 - w$) decreases with more categories, it is still quite substantial for the 100-point scales.

As with previous research on social judgments (Wedell et al., 1987; Wedell & Parducci, 1988), the slopes of the range functions generally decrease with increase in number of categories, reflecting a corresponding tendency to extend the range of subjective values beyond the immediate set. This tendency is greater when the range is restricted, as indicated by reduced slopes for restricted-range sets. The bottom panels of Figure 1 graphically illustrate how the differences in the range functions for mild and severe sets add to the greater contextual effects for restricted-range conditions.

Statistical analysis. Analysis of variance (ANOVA) was performed on ratings of the six test cases after linearly transforming them to a common 7-point scale (following Equation 4). The results of the analysis generally confirm the range–frequency effects described above. The main effect of context was highly significant, $F(1, 491) = 355.46$, $p < .0001$, as was the Context × Range interaction, $F(1, 491) = 37.97$, $p < .0001$, which indicates that contextual effects were greater for restricted-range conditions. These large effects of the contextual manipulations imply that the reliability would be higher for ratings of the test stimuli calculated separately within each context than if calculated for all contexts combined in one analysis.

The category effect, represented by the interaction between context and number of categories, was also statistically significant, $F(2, 491) = 8.78$, $p < .001$. Planned comparisons at $p < .05$ showed that the effects of context were significantly reduced from three to seven categories; however, the difference between 7- and 100-point scales was not statistically significant.

Evidence for greater adjustment to the experimental range with fewer categories is given by the significant interaction between test cases and categories, $F(10, 2455) = 4.79$, $p < .0001$. The decrease in the slopes of the rating functions with increase in number of categories is indicative of an increasing tendency to reserve extreme categories for stimulus values more extreme than those presented in the experimental set. The only other significant effects were the Context × Cases and the Context × Range × Cases interactions ($p < .01$). As shown by the close fit of the model to the ratings, these effects reflect the expected differences in range–frequency effects for case histories at different locations on the dimension of judgment.

Psychometric analysis. The repeated-measures design of Experiment 1 allows for the estimation of coefficients of internal reliability based on the error terms from the ANOVA. In order to estimate the reliabilities for the different rating scales, separate one-way ANOVAs were run for each. Following Winer (1971, pp. 283–296), the reliability of the judgments of a single subject is given by:

$$r_1 = \frac{\theta}{\theta + 1}, \tag{5}$$

where θ is the ratio of true score variance to error score variance. θ can be estimated from the ANOVA as follows:

$$\theta = \frac{MS_{b,\text{cases}} - MS_{w,\text{cases}}}{(n)MS_{w,\text{cases}}} = \frac{F - 1}{n}, \tag{6}$$

where n is the number of subjects and F is the ratio of between- and within-cases variance. An alternative estimate would substitute the MS_{residual} for the $MS_{w,\text{cases}}$ term, thereby eliminating from the error term differences in how subjects centered (or anchored) their scales. However, use of the within-cases term is consistent with the present focus on the effects of systematic shifts of scale and is closer in spirit to measures of interrater

Table 2
Reliability Indices r_1 *for Experiment 1*

	Restricted range			Full range		
Context	3 pt.	7 pt.	100 pt.	3 pt.	7 pt.	100 pt.
Mild	.40	.42	.39	.45	.53	.48
Severe	.48	.48	.42	.53	.52	.54
Combined	.31	.35	.30	.43	.50	.50

reliability that are based on the relative frequency of a match between judges. All $MS_{w,\text{cases}}$ terms were taken from ANOVAs run on the six test cases. The $MS_{b,\text{cases}}$ terms were taken from ANOVAs run on all 20 cases. Because different sets of cases were judged in the combined conditions, these between-cases terms were estimated by simply adding the component mean squared terms.

Table 2 presents the reliability coefficients as calculated using Equation 5. The generally low values for the reliability coefficients indicate the use of different judgmental standards by different individuals. The effects of the contextual manipulations are reflected in the lower reliabilities for the combined versus the separate (mild or severe) contexts. The decrement in reliability is greatest for the restricted-range sets. Although increasing the number of categories resulted in a decrease in the error terms, this decrease was accompanied by a corresponding decrease in the between-cases variance so that reliability estimates were nearly the same across different numbers of categories. Thus, the psychometric analysis shows no particular advantage from increasing number of categories on reliability.

Discussion

The large contextual effects demonstrated in Experiment 1 represent an important potential source of unreliability in clinical assessment. Although one might assume that professional clinicians would be less susceptible to these effects, Perrett (1971) found no significant difference in the magnitude of contextual effects between experienced clinicians and naive subjects. Because clinicians in her study were more reliable within contextual conditions, the relative impact of the contextual manipulation was actually greater for them.

Increasing the number of categories from three to seven did decrease the magnitude of contextual effects, although there was no further reduction for the 100-point scale. Because the contextual effects for the 7- and 100-point scales were still substantial, increasing the number of categories in the rating scale does not appear to be the answer to eliminating unwanted contextual dependencies in clinical judgment. However, the greater contextual dependency of the three-category scale should serve as a warning to clinicians against making even casual covert assessments on a simple two- or three-category scale. Our results suggest that judgments based on a broad categorization scheme (*well-adapted* versus *poorly adapted*) would be heavily dependent on the set of cases recently encountered by the clinician.

The greatest reduction in contextual effects found in Experiment 1 was achieved by simply presenting subjects with the full range of case histories. The range–frequency analysis attributes this reduction to elimination of the differences in the range values for the common cases and to decreased differences in the frequency values. Thus, including both extremely mild and extremely severe cases might go a long way toward standardizing clinical judgments. However, it should be noted that the contextual effects for the full-range conditions of Experiment 1 were still quite substantial, on the order of a full category-step difference for the seven-category scales. In a situation in which a clinician must decide whether or not to commit an individual to a state institution, such differences could have important consequences.

The seven-category, restricted-range conditions can be directly compared with corresponding conditions of Perrett's (1971) experiment because the case histories and instructions were virtually the same. It is therefore striking that the present contextual effects for these conditions are approximately twice as great as Perrett's. This difference may be attributed to a difference in the order of presentation of the case histories: In our experiment, the 6 test cases were presented last, whereas Perrett interspersed them among the 14 contextual cases. Positional effects on clinical ratings have been reported in previous studies (e.g., Bieri et al., 1966; Campbell et al., 1957), and Perrett (1971) noted that the effects of context were greatest for test cases occurring near the end of the experimental series.

The range–frequency analysis presented in Figure 1 suggests that putting the test cases last in the present study enhanced the effects of context by shifting the subjective range, that is, context effects were much greater when endpoints varied (top panels) than when they were constant across conditions (middle panels). Perrett (1971) found no reduction in contextual effects when the rating categories were anchored by detailed verbal descriptions; however, if such descriptions could establish a common range for the different contexts, they might thereby reduce effects of the immediate context. Experiment 2 explores this possibility using two different types of anchors.

Experiment 2: Anchoring the Scale of Judgment

An anchor refers to a stimulus or verbal description that stabilizes the scale of judgment by prescribing a fixed correspondence between a stimulus value and a particular category. In general, stimulus anchors tend to reduce the variability in responses to nearby stimuli (Johnson, 1972). Of more relevance to this discussion is the possibility that providing a stimulus anchor for each of the two endpoints of the rating scale may help to establish a common range of stimuli and hence eliminate contextual effects due to differences in range values. Prescribing anchors may also reduce the effects of manipulating the relative frequencies of stimuli: Subjects' repeated referrals to anchors may function like repeated presentations of their values within the distribution of contextual stimuli, resulting in a leveling out of contextual frequencies (which would produce a reduction in effects of the immediate context). Finally, providing stimulus examples or well-defined verbal descriptions for each category may reduce contextual effects by emphasizing the nature of the task as one of absolute identification. Range–frequency theory describes how the judgment locates the stimulus within the distribution of contextual stimuli; clinical assessment attempts to assign the individual's behavior to a well-defined, absolute category.

Although the use of descriptive labels has been demonstrated

Table 3
DSM Descriptions Used in Experiment 2

1 = SUPERIOR:	Unusually effective in social relations, occupation, and use of leisure time.
2 = VERY GOOD:	Better than average in job, leisure time, and social functioning.
3 = GOOD:	No more than slight impairment in either occupational or social functioning.
4 = FAIR:	Moderate impairment in either social or occupational functioning or both.
5 = POOR:	Marked impairment in either social or occupational functioning or moderate impairment in both.
6 = VERY POOR:	Marked impairment in social and occupational functioning.
7 = GROSSLY IMPAIRED:	Gross impairment in virtually all areas of functioning.

to increase reliability of judgment (Bernardin & Beatty, 1984), these demonstrations have not included manipulations of context, so there is no basis for concluding that anchors reduce this source of unreliability. In experiments using clinical judgment, descriptive anchors have not succeeded in reducing contextual effects (Arnoff, 1954; Perrett, 1971). However, comparison between the results of Experiment 1 and Perrett's (1971) study suggested that providing scale anchors might reduce the effects of placing the target stimuli last for restricted-range conditions. To test this hypothesis, the 7-point rating scales of Experiment 2 employed two different types of anchors: (a) A case history was provided for each endpoint of the scale, or (b) a detailed description from the third edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-III;* American Psychiatric Association, 1980) was provided for each category.

Method

Design and subjects. The 3 × 2 × 2 × 6 factorial design of Experiment 2 closely paralleled that of Experiment 1, except that anchoring (instead of number of categories) was manipulated at three levels (unanchored, exemplar anchors, and descriptive anchors). All ratings were made on a 7-point scale. Each Context × Range condition used the same case histories presented in the same order as in Experiment 1. An additional 335 subjects, sampled from the same population as Experiment 1, participated in the exemplar and descriptive anchor conditions; data from the 163 subjects in the 7-point conditions of Experiment 1 were used for the unanchored conditions.

Instructions. Instructions were made as similar as possible for the different anchoring conditions. For *exemplar anchoring,* mild and severe examples corresponded to the first case histories presented in the mild and severe contexts. These were printed on the instruction sheet and labeled 1—*Very, Very Mild Disturbance* and 7—*Very, Very Severe Disturbance,* respectively. In Experiment 1, the mild example ranked 2 in the overall set of 34 case histories and the severe case example ranked 31. The two example case histories read:

> *1—very, very mild disturbance.* A 24-year-old mother of one child came for treatment asking for marriage counseling. Although she is quite an adequate housekeeper and mother, she is feeling a lack of fulfillment in her life. She feels that she and her husband do not communicate well and the problem is compounded by her husband's busy schedule.
>
> *7—very, very severe disturbance.* A 44-year-old housewife has been hearing a voice telling her to shoot her husband and stab herself in the heart for the past year. She claims that a neighbor, whom she believes is a witch, is responsible for this command and she has almost acted on the command of the voice.

For *descriptive anchoring,* descriptions from the *DSM-III* Axis V were used. These are reproduced in Table 3.

Results and Discussion

The differences in the mean ratings of the case histories shown in Figure 2 (top two rows of panels) demonstrate a clear contrast effect, as in Experiment 1. However, both exemplar and descriptive anchoring reduced the contrast for the restricted range (top row), and less so for the full range (second row). Anchoring also affects the general slopes of the rating functions: The rating scales anchored at the endpoints by case histories show a slightly steeper slope; those anchored by the *DSM-III* descriptions show a markedly reduced slope and a much higher intercept on the ordinate.

A four-way ANOVA was performed on the mean ratings of the

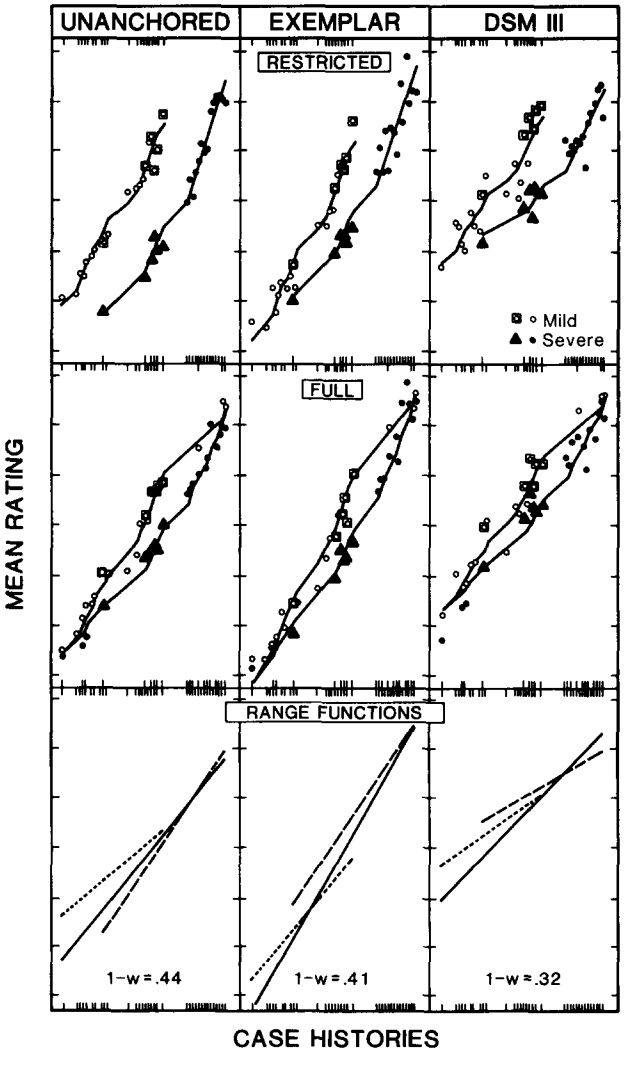

Figure 2. Reduction in contextual effects for anchored scales of Experiment 2. (Data for unanchored condition from Figure 1, 7-point data.)

six test cases. As in Experiment 1, the main effect of context, $F(1, 496) = 293.3, p < .0001$, and the interaction of context and range, $F(1, 496) = 29.9, p < .0001$, were highly significant and in accordance with the range–frequency model. However, the interaction between anchor and context, of particular concern in this study, was only marginal, $F(2, 496) = 2.7, p < .10$. Although the three-way Anchor × Context × Range interaction was not statistically significant, $p > .25$, separate analyses for the two levels of range revealed a significant Anchor × Context interaction for the restricted-range sets, $F(2, 496) = 3.1, p < .05$, but not for the full-range sets, $p > .50$. Specific comparisons for restricted-range sets revealed that the effects of context were significantly greater in the unanchored condition than in either of the anchored conditions ($p < .05$); effects of context did not differ significantly between exemplar and descriptive anchoring conditions ($p > .50$). Thus, Experiment 2 provides some evidence that anchoring the clinical scale of judgment can reduce the effects of restricting the range of cases presented for judgment, in this case by 29% for exemplar anchors and 25% for *DSM* anchors. Moreover, the failure of anchoring techniques to produce similar reductions when cases from the same restricted sets were not reserved until the final presentations (Perrett, 1971) suggests that the anchors serve to reduce the effects of the position in which the test cases are presented.

Fit of the range–frequency model. The range–frequency model allows for a more detailed analysis of the effects of anchoring the clinical rating scale. The fit to the data, shown in Figure 2, was obtained in the same manner as Experiment 1 (except that order and spacing of the cases along the abscissa were taken directly from Experiment 1). The essential difference between anchored and unanchored scales can be seen in the inferred range functions for the restricted sets (bottom panels). For the unanchored scales, the inferred range function for the mild context lies above that for the severe context so that differences in the range values add to the overall contrast effect; however, for the anchored scales, the inferred range function for the mild context lies below that for the severe context so that differences in the range values subtract from the observed contextual effects.

The use of different anchors had a strong effect on the slope of the rating functions as revealed in the ANOVA by an interaction between anchors and case histories, $F(10, 2480) = 5.27, p < .0001$. When exemplar anchors were used, the rated differences among the test cases increased (i.e., slope of the rating function increased), and when *DSM* anchors were used, the rated differences among the test cases were reduced. These effects on the slopes of the rating functions are explained by the range–frequency model in terms of changes in the subjective range. Anchoring the scale with case histories restricted the subjective range of judgment to the overall set of cases from which they were drawn, as indicated by range functions falling near the diagonal (bottom row, middle panel of Figure 2). Thus, subjects in these endpoint-anchoring conditions apparently succeeded in following instructions, generally anchoring the endpoints of their rating scale to the specific examples provided.

The *DSM* anchors extended the subjective range to include cases much milder (healthier) than any of those presented in the experimental sets, as indicated by the reduced slopes and higher intercepts on the ordinate for these range functions (bottom right panel of Figure 2). Once again, subjects apparently followed instructions, because the descriptions for the first two categories (i.e., *superior* and *above-average* functioning) were not applicable to any of the case histories presented. The hypothesized differences in the intercepts of the range functions across anchor conditions are supported by the significant main effect of anchor in the ANOVA, $F(2, 496) = 66.5, p < .0001$.

Table 4
Reliability Indices r_1 *for Experiment 2*

	Restricted range		Full range	
Context	Exemplar	*DSM*	Exemplar	*DSM*
Mild	.45	.54	.66	.63
Severe	.57	.49	.69	.66
Combined	.44	.38	.64	.62

The range–frequency analysis indicates that the use of descriptive anchors may reduce the weighting of the frequency principle, $1 - w$. Although there was no significant interaction between anchor and context for the full-range conditions, the value of $1 - w$ inferred for these conditions was reduced by approximately one fourth for the descriptive versus exemplar or unanchored scales. One interpretation of this apparent reduction in the frequency weighting is that it is artifactual, that is, a consequence of restricting the relevant range of judgmental categories to only the top five rather than the full range of seven categories (because the first two categories were reserved for individuals functioning at above-average levels). This restriction limits the degree to which a low frequency value can pull the rating of a case history down.

Finally, the closeness of empirical data points to the model's predictions can be considered, in part, as a test of the cross-validity of the scale values derived from the model's fit to data from Experiment 1. Scale values appear to have cross-validated well for exemplar–anchor conditions, as indicated by the good fit of the model. However, the deviation of data points from model predictions is much greater for the *DSM*-anchor conditions, where ordering of the case histories is less consistent with the spacing of the stimuli along the abscissa. One explanation of this poorer fit is that the *DSM* response scale emphasizes different aspects of the case histories (behavioral maladaptiveness rather than mental disturbance), and thus it might be represented more appropriately as a vector (through the multidimensional space defining the stimuli) that is highly correlated with, but not identical to, the vector defining the other scales.

Psychometric analysis. As in Experiment 1, separate one-way ANOVAs were run on the data from the two different anchor conditions to generate reliability coefficients. Table 4 presents the results of this analysis. Overall, reliabilities are higher for these anchored scales than for the unanchored scales (cf. Table 2). The higher reliabilities were achieved in different ways by the two types of anchors. Greater reliability for the exemplar anchors resulted from a moderate (21%) reduction in the within-cases error variance combined with a moderate (26%) increase in the between-cases variance. Greater reliability for the *DSM* anchors resulted from a large (52%) reduction in variance within cases that was partially offset by a moderate (19%) reduction in variance between cases. As in Experiment 1, the effects of the contextual manipulations are reflected in reduced

reliabilities for the combined contexts as compared with the separate contexts. These effects were greater for restricted-range conditions.

General Discussion

There is ample evidence that reliability of clinical judgment is unacceptably low. Our focus in this study was to develop a better understanding of unreliability that results from the use of different contextually generated standards. The large contextual effects obtained here reinforce previous findings that clinical judgment is subject to the same principles demonstrated for perceptual and social judgments. In Experiments 1 and 2, independence from the immediate set of case histories was facilitated by using at least seven categories and by anchoring the rating scale with detailed descriptions or examples. However, the strongest decrease in contextual effects was created by equating the range of cases within the immediate context for judgment (i.e., reducing the differences between the stimulus sets that were judged). Because undergraduates served as judges in these experiments, one may question whether our findings extend to trained clinicians. There are two good reasons to believe that they should. First, using these same stimulus materials, Perrett (1971) found no difference in the magnitude of contextual effects between the 342 clinicians and the 155 undergraduates in her study. Second, more generally, there is considerable evidence that accuracy of clinical judgment does not differ between clinicians and laypersons (for a review, see Faust & Ziskin, 1988).

Because shifts in context constitute a potential source of unreliability, failure to completely control them in past and present studies suggests the need for further research in this area. One avenue for future exploration would be to vary the procedure for eliciting judgments. Instead of making a single absolute evaluation for each case, the judge might make a series of comparative judgments to established standards or example cases. Although a comparative judgment procedure may be implied by the use of detailed descriptive anchors, it seems unlikely that subjects spontaneously follow this time-consuming procedure. In an analysis of the loci of contextual effects, Mellers and Birnbaum (1982) demonstrated that difference ratings of stimulus pairs varying along the same dimension were less governed by the immediate context than ratings of single stimuli. The same is true when stimuli vary on multiple dimensions (Corter, 1987; Jones & Wedell, 1987). Better understanding of these differences might provide a basis for control of the context, which in turn could produce substantial increase in the reliability of clinical assessments.

In addition to the ambitious task of controlling the context for judgment, it seems important to develop a more precise understanding of just what constitutes the context in a clinical situation. The present study demonstrated that the set of case histories recently encountered has a strong impact on judgment. This type of social comparison context might be expected to have greatest impact when a clinician encounters a new client's problems for the first time (i.e., the client's behaviors are compared to those of other clients recently encountered). As therapy progresses, the effective context on which judgment is based may switch to the set of behaviors exhibited by the same client on previous occasions, an intrapersonal context. Smith et al. (1989) have demonstrated that judgments based on both social comparisons and intrapersonal comparisons follow range–frequency principles. Because these contexts may differ markedly, judgments should depend on which context is brought to mind at the time of the evaluation.

The rating task employed in the present experiments is most similar to the *DSM-III* evaluation task along Axis V (i.e., "Highest Level of Adaptive Functioning Past Year"). It would be of interest to determine whether similar contextual effects occur for evaluations along Axis IV (i.e., "Severity of Psychosocial Stressors") and, more important, for evaluations along the first three axes, which together constitute the official diagnostic assessment. These three axes differ from Axes IV and V in that their categories are not ordered along dimensions but rather represent patterns of symptoms. As such, the range–frequency model might not be directly applicable. However, some features of the model, such as the tendency to use categories with equal frequencies, might explain the general tendency to ignore base rates (Balla, Elstein, & Gates, 1983; Casscells, Schoenberger, & Grayboys, 1978; Kahneman & Tversky, 1973; Meehl & Rosen, 1955). Furthermore, insofar as these three more specific axes tend to tap more global evaluative dimensions, similar contextual effects should be expected, and range–frequency theory would be directly applicable.

Finally, because a general goal of personality assessment is the standardization of judgment across potentially discrepant contextual settings, more attention should be paid to the effects of manipulating context. In a typical reliability study, judges rate a common set of stimuli, and the degree of interrater agreement is evaluated. Because the immediate stimulus context is the same for all judges, it is difficult to determine how susceptible these scales are to contextual effects. Within the framework of generalizability theory (Cronbach et al., 1963), the pertinent universe of generalization is one that includes different judges and different contexts. As demonstrated in our studies, a reliability coefficient generated within a constant context will overestimate the reliability of the measure across contexts. These results have general application beyond clinical assessment. For example, contextual effects have been recognized by organizational psychologists as a potential source of error in appraising performance (Bernardin & Beatty, 1984; Borman, 1979), and efforts have been made to train judges to adopt the same standards (Bernardin & Buckley, 1981). However, the assessment of the effectiveness of any such training procedure requires estimation of reliability across manipulated contexts. A failure to address the issue of systematic shifts of standards across different judges and different contextual settings will result in the overestimation of both the reliability and validity of the assessment procedures.

References

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

Arnoff, F. N. (1954). Some factors influencing the unreliability of clinical judgments. *Journal of Clinical Psychology, 10,* 272–275.

Balla, J. I., Elstein, A., & Gates, P. (1983). Effects of prevalence and test diagnosticity upon clinical judgments of probability. *Methods of Information in Medicine, 22,* 25–28.

Bernardin, H. J., & Beatty, R. W. (1984). *Performance appraisal: Assessing human behavior at work.* Boston: Kent.

Bernardin, H. J., & Buckley, M. R. (1981). A consideration of strategies in rater training. *Academy of Management Review, 6,* 205–212.

Bieri, J., Atkins, A. L., Briar, S., Leaman, R. L., Miller, H., & Tripoldi, T. (1966). *Clinical and social judgment: The discrimination of behavioral information.* 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.

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology, 64,* 410–421.

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.

Casscells, B. S., Schoenberger, A., & Grayboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine, 229,* 999–1000.

Cohen, J. A. (1960). Coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Corter, J. E. (1987). Similarity, confusability, and the density hypothesis. *Journal of Experimental Psychology: General, 116,* 238–249.

Cronbach, L. J., Gleser, G. C., & Rajaratnam, N. (1963). Theory of generalizability. A liberalization of reliability theory. *British Journal of Mathematical and Statistical Psychology, 16,* 137–173.

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

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81,* 95–106.

Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance, 13,* 171–192.

Eiser, J. R., & Stroebe, W. (1972). *Categorization and social judgment.* London: Academic Press.

Eysenck, H. J., Wakefield, J. A., Jr., & Friedman, A. F. (1983). Diagnosis and clinical assessment: The DSM-III. *Annual Review of Psychology, 24,* 167–193.

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

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

Goldberg, L. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin, 73,* 422–432.

Johnson, D. M. (1972). *A systematic introduction to the psychology of thinking.* New York: Harper & Row.

Jones, L. E., & Wedell, D. H. (1987, July). *Contextual effects in multidimensional scaling: Variations in stimulus densities.* Paper presented at the fifth annual European meeting of the Psychometric Society, Enchede, Netherlands.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251.

Labowitz, S. (1970). The assignment of numbers to rank order categories. *American Sociological Review, 35,* 515–524.

Manis, M., & Paskewitz, J. R. (1984a). Specificity in contrast effects: Judgments of psychopathology. *Journal of Experimental Social Psychology, 20,* 217–230.

Manis, M., & Paskewitz, J. R. (1984b). Judging psychopathology: Expectation and contrast. *Journal of Experimental Social Psychology, 20,* 363–381.

Manis, M., Paskewitz, J. R., & Cotler, S. (1986). Stereotypes and social judgment. *Journal of Personality and Social Psychology, 52,* 663–676.

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

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficacy of psychometric signs, patterns or cutting scores. *Psychological Bulletin, 52,* 194–216.

Mellers, B. A. (1983). Equity judgment: A revision of Aristotelian views. *Journal of Experimental Psychology: General, 111,* 242–270.

Mellers, B. A. (1986). "Fair" allocations of salaries and taxes. *Journal of Experimental Psychology: Human Perception and Performance, 12,* 80–91.

Mellers, B. A., & Birnbaum, M. H. (1982). Loci of contextual effects in judgment. *Journal of Experimental Psychology: Human Perception and Performance, 8,* 582–601.

Mellers, B. A., & Birnbaum, M. H. (1983). Contextual effects in social judgment. *Journal of Experimental Social Psychology, 19,* 157–171.

Miller, G. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72,* 407–418.

Parducci, A. (1982). Category ratings: Still more contextual effects. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement* (pp. 262–282). Hillsdale, NJ: Erlbaum.

Parducci, A. (1983). Category ratings and the relational character of judgment. In H. G. Geissler, H. F. J. M. Buffort, E. L. J. Leeuwenberg, & V. Sarris (Eds.), *Modern issues in perception* (pp. 89–105). Berlin: VEB Deutsher Verlag der Wissenschaften.

Parducci, A., & Wedell, D. H. (1986). The category effect with rating scales: Number of categories, number of stimuli, and method of presentation. *Journal of Experimental Psychology: Human Perception and Performance, 12,* 496–516.

Perrett, L. F. (1971). *Immediate and background contextual effects in clinical judgment.* Unpublished doctoral dissertation, University of California, Los Angeles.

Poulton, E. C. (1979). Models for biases in judging sensory magnitude. *Psychological Bulletin, 86,* 777–803.

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

Smith, R. H., Diener, E., & Wedell, D. H. (1989). Intrapersonal and social comparison determinants of happiness: A range-frequency analysis. *Journal of Personality and Social Psychology, 56,* 317–325.

Upshaw, H. S. (1969). The personal reference scale: An approach to social judgment. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 4, pp. 315–371). New York: Academic Press.

Volkmann, J. (1951). Scales of judgment and their implications for social psychology. In J. H. Rohrer & M. Sherif (Eds.), *Social psychology at the crossroads* (pp. 273–294). New York: Harper & Row.

Wedell, D. H., & Parducci, A. (1985). Category and stimulus effects: A process model for contextual memory in judgment. In G. d'Ydewalle (Ed.), *Cognition, information processing, and motivation* (pp. 55–70). New York: Elseveir.

Wedell, D. H., & Parducci, A. (1988). The category effect in social judgment: Experimental ratings of happiness. *Journal of Personality and Social Psychology, 55,* 341–356.

Wedell, D. H., Parducci, A., & Geiselman, R. E. (1987). A formal analysis of ratings of physical attractiveness: Successive contrast and simultaneous assimilation. *Journal of Experimental Social Psychology, 23,* 230–249.

Wedell, D. H., Parducci, A., & Roman, D. (1989). Student perceptions of fair grading: A range-frequency analysis. *American Journal of Psychology, 102,* 233–248.

Winer, B. J. (1971). *Statistical principles in experimental design.* Tokyo: McGraw-Hill Kogakusha.

Appendix A

Six Test Cases

The scale value (S) accompanying each case history is derived from the fit of the range–frequency model (reflected in the spacing along the abscissa in Figures 1 and 2). These values range from 0 = *mildest case* in the set of 34 to 1.0 = *most severe case* in the set.

1. This 32-year-old divorced woman came for advice on vocational guidance. She had difficulties in meeting people and feels inadequate and insecure in relationships with others, particularly with those of her own age. She asked for some kind of course to "get me out of myself." (S = .26)

2. The patient is a single 22-year-old woman who is finding difficulty in separating from her parents, with whom she lives. She relates to people with considerable anxiety, feels angry toward men, and is ambivalent about physical contact. (S = .51)

3. This 29-year-old married man was discharged from the service, where he had been very tense and nervous. A few months after discharge, he experienced such anxiety that he was unable to hold his job. A physical examination did not reassure him that there was nothing wrong with his heart. He is very dependent on his mother and not very sociable. There is a marked lack of spontaneity in his speech and a poverty of ideation. (S = .55)

4. The patient is a 47-year-old housewife who has experienced severe headaches for 25 years and abdominal pains for 12 years—although no organic basis for the symptoms has been found. She does not talk freely about her personal life and states that her home life is satisfactory. She considers emotional problems to be without importance and refused to try to relate them to her present symptoms. (S = .56)

5. This 41-year-old man feels his marriage is dissolving. He has been expressing increasing anger and lack of satisfaction with his wife, who is 14 years his senior. He has no sexual relations with her. Rather, he has had frequent homosexual relationships for the past 10 years. In addition, he is unable to bring home enough money to maintain their standard of living. (S = .58)

6. This 55-year-old mother took an overdose of pills. She feels hopeless and depressed. Her appetite is poor and she has numerous physical symptoms. She has been feeling increasingly alienated and alone since her husband died 4 years ago. She never attempted to cultivate any close friendships following the death of her husband. (S = .62)

Appendix B

Fitting the Range–Frequency Model

To fit the range–frequency model to the data of Experiment 1, a single ordering of the 34 case histories (from most mild to most severe) was first established on the basis of an average of the relative ranking of the cases within the different contextual conditions. This ordering of the stimuli was used along with Equation 2 to generate frequency values for the cases in each of the four contextual conditions. Next, the frequency weighting, $1 - w$, was inferred. Substituting in Equation (2) for each distribution and subtracting yields:

$$J_{im} - J_{is} = (wR_{im} + (1 - w)F_{im}) - (wR_{is} + (1 - w)F_{is}). \quad (1A)$$

When the extreme stimuli are the same, range values for mild (R_{im}) and severe (R_{is}) contexts are assumed to be equal and so drop out; thus, transposing for the full-range conditions yields:

$$(1 - w) = (J_{im} - J_{is})/(F_{im} - F_{is}). \quad (1B)$$

A single weighting value was then determined for each number of categories by averaging the $1 - w$ values calculated for the six test cases of the full-range condition. These values are shown in the bottom panels of Figure 1 and were assumed to be the same for corresponding restricted-range conditions.

Range values were inferred by substituting frequency values, $1 - w$, and empirical ratings (transformed via Equation 4) into Equation 3. Range–frequency theory assumes that scale values (S_is of Equation 1) are invariant across contextual conditions and monotonically related to the ranks of the stimuli. Because of the high frequency weighting for the 3-point scales, range values inferred for these conditions are less reliable and hence were not used in the determination of the spacing of scale values. Therefore, to establish the relative spacing of scale values for the 34 case histories (as shown on the abscissa of the bottom panels of Figure 1), range values for 7- and 100-point scales were averaged together under the condition of preserving the ordering of the stimuli. Range values were then linearly regressed onto the scale values to derive the best-fit range functions. Because range–frequency theory assumes that the range function is influenced by the contextual endpoints, and because previous research has demonstrated that the range function is affected by the number of categories (Wedell & Parducci, 1988; Wedell, Parducci, & Geiselman, 1987), three separate range functions were inferred for each number of categories (one for each of the restricted-range conditions and one for the combined full-range conditions). These inferred range functions are shown in the bottom panels of Figure 1. Altogether, 54 estimated parameters (33 for the spacing of the 34 cases, a slope and an intercept parameter for each of the range functions, and 3 values of w) were used to fit the 240 data points.

Received October 6, 1988
Revision received April 11, 1989
Accepted July 26, 1989 ■