**Student Perceptions of Fair Grading: A Range-Frequency Analysis**

Douglas H. Wedell; Allen Parducci; Diana Roman

*The American Journal of Psychology*, Vol. 102, No. 2. (Summer, 1989), pp. 233-248.

Stable URL:
http://links.jstor.org/sici?sici=0002-9556%28198922%29102%3A2%3C233%3ASPOFGA%3E2.0.CO%3B2-P

*The American Journal of Psychology* is currently published by University of Illinois Press.

# Student perceptions of fair grading: A range-frequency analysis

DOUGLAS H. WEDELL, ALLEN PARDUCCI, and
DIANA ROMAN
University of California, Los Angeles

University students were instructed to assign grades as fairly as possible to different hypothetical distributions of exam scores (bell, *U*, positively skewed, and negatively skewed). Experiment 1 demonstrated significant distribution effects that were quantitatively consistent with Parducci's (1965) range-frequency theory: Grading reflected a roughly equal compromise between a tendency to assign grades to equal subranges of exam scores (e.g., A's to the top fifth of the range) and a tendency to assign an equal number of scores to each grade (e.g., A's to the top 20% of scores). Individual differences in the relative weighting of these two tendencies were fairly reliable, $r_{tt'}$ = 0.76. Although most students followed a roughly equal compromise, a few favored "grading on a straight scale" (equal subranges) and a few "grading on a curve" (equal frequencies). The results of Experiment 2 supported a range-frequency model in which different grades tend to be used with equal frequency over a modified model in which different grades tend to be used with fixed but unequal frequencies.

The assignment of grades can be a source of frustration for student and teacher alike. Although there are many ways to objectify the assessment of a student's academic performance, the assignment of a corresponding grade necessarily requires a subjective value judgment. The subjective nature of grading raises the issue of fairness. What do students perceive as fair? The present experiments attempt to answer this question by asking students to assign grades themselves for various distributions of exam scores.

By varying the distribution of scores for different exams, a fairness "rule" can be inferred from the grade assignments. For example, using similar manipulations, Mellers (1982, 1986) has shown that judgments of equity as well as salary allocations follow a fairness rule consistent with Parducci's (1965) range-frequency theory of judgment. The aim of the present set of experiments is threefold: (a) to determine whether range-frequency theory characterizes students' conceptions of fair assignment of grades; (b) to test between two competing versions of the range-frequency model; and (c) to investigate individual dif-

AMERICAN JOURNAL OF PSYCHOLOGY
Summer 1989, Vol. 102, No. 2, pp. 233–248

ferences in how subjects weight between range and frequency principles.

Range-frequency theory (see Parducci, 1983, for a more complete presentation of the theory) proposes that value judgments represent a compromise between a range and a frequency principle. According to the range principle, there is a tendency to divide the stimulus range into equal subranges corresponding to the available categories of judgment. For example, if one were to follow the range principle in grading a set of exam scores varying from 50 to 100, each of five letter grades would correspond to an approximate range interval of 10 (e.g., the cutoff would be 90 for A's, 80 for B's, etc.). This method of assigning grades is a special case of a "straight scale," where grade cutoffs are independent of the frequency distribution of scores.

The frequency principle refers to the tendency to assign an equal number of stimuli to each category of judgment. Following the frequency principle, the five letter grades would be used equally often (viz., the highest scoring fifth of students is assigned an A, the next fifth a B, etc.). This method of assigning grades is similar to "grading on a curve," where fixed (but usually unequal) percentages of scores are assigned to each grade. Actual cutoffs are assumed to reflect a compromise between range and frequency principles, represented algebraically as follows:

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

where $J_{ic}$ is the stimulus value corresponding to the cutoff between grades $i$ and $i + 1$ in Context $c$; $R_{ic}$ is the cutoff by the range principle; $F_{ic}$ the cutoff by the frequency principle; and $w$ the relative weighting of the range principle.[1]

An example will help clarify the relationship between assignment of grades and the distribution of scores described by Equation 1. Let us assume that the subjective range of scores under consideration corresponds to the actual range of scores on a particular exam, in this case from 50 to 100. As described earlier, the range principle places the cutoff scores for this exam at approximately 90, 80, 70, and 60 for A/B, B/C, C/D, and D/F, respectively. However, where the actual cutoffs are made will depend upon the frequency distribution of the scores. If the scores are uniformly distributed, the grade cutoffs based upon the frequency principle will be the same as those based upon the range principle, and thus these will be the actual cutoffs (regardless of the value of $w$). If the scores are normally distributed with a mean of 75 and standard deviation of 10, then the cutoff scores corresponding to the 80th, 60th, 40th, and 20th percentiles will be at approximately 84, 78, 72, and 66. Where actual

cutoffs fall depends on the relative weighting of range and frequency tendencies. For example, if $w = 0.5$, then actual cutoffs will fall halfway between these two sets (i.e., 87, 79, 71, and 63). By having subjects assign grades to a variety of frequency distributions that span the same range, it is possible to estimate $w$ for each subject.

Previous research has shown that ratings of the magnitudes of numbers follow a range-frequency compromise with $w$ close to 0.5 (Birnbaum 1974; Parducci, Calfee, Marshall, & Davidson, 1960). Mellers and Birnbaum (1983) demonstrated similar results for category ratings of performance based upon individual and combined exam scores. However, because letter grades represent a well-established scale of judgment for use in the specific context of academic evaluation, it is unclear whether this relationship will extend to the assignment of conventional grades. Furthermore, one might expect assignment of grades to follow a version of the range-frequency model in which the frequency principle is modified to allow different grade categories to be used with fixed but *unequal* frequencies. The typical "grading curve" is an overt example of this type of modified frequency principle, generally assigning fewer D's and F's than A's and B's.

The present experiments were designed to test between these two models. In addition, by having each student-subject grade each of several different distributions, Experiment 1 allows for the evaluation of individual differences in the weighting of range and frequency tendencies as well as estimates of the reliability of the weighting.

## EXPERIMENT 1

### METHOD

#### Design and subjects

The experiment employed a 4 × 4 × 4 factorial design, with order of presentation (four counterbalanced orders) as the only between-subject factor; distribution (bell, *U*-shaped, positively skewed, and negatively skewed) and grade cutoffs (A/B, B/C, C/D, and D/F) were the two within-subject factors. The basic dependent variable was the score assigned to each grade cutoff (defined here as the lowest score assigned the higher grade).

Subjects were 80 undergraduates enrolled in an introductory psychology course at the University of California, Los Angeles (UCLA), who received credit toward a departmental requirement by participating in the experiment. They were tested in groups of 8 to 12 in a small laboratory room. Data from 4 subjects, who assigned fewer than four different grades, were discarded.

### Exam scores

Each of the four stimulus sets consisted of 100 scores, from 50 to 100, printed in two columns, in order from highest to lowest. Different distributions were created by manipulating the relative frequencies of the scores, but with no score occurring more than five times. Table 1 shows the relative frequencies for the four distributions.

### Instructions

Instructions were printed on the first page of each five-page experimental booklet. Subjects read silently (while the experimenter read aloud) that the study was about how people felt exam grades should be assigned and that the distributions of exam scores printed on each of the following four pages represented scores obtained by 100 students on a 100-question, four-option, multiple-choice exam. The four lists were to be considered independently, as if each represented the test results from a different introductory course in the social sciences at the university. The task was to assign grades *as fairly as possible* for each list of scores. Subjects were encouraged to use all five grades (A, B, C, D, and F). Cutoffs were to be indicated by drawing a line between the scores they wished to separate. An example was given with the A/B cutoff between scores 89 and 90. Subjects were told to work through the four sets of scores in the order in which they appeared. After assigning grades to a set, they were to rate (on the basis of the scores provided) how well that test measured knowledge of the course material. These ratings were made using a five-point rating scale that appeared on each page, from *very poor* (1) to *very good* (5).

## RESULTS AND DISCUSSION

### Assignment of grades

As shown in Figure 1, the assignment of grades depends upon the distribution of test scores. Considering that the mean of the standard errors for the cutoffs is only 0.484 exam score units (less than 1/100 of the range of exam scores), it is obvious that the effects of distribution and also their interaction with grade cutoff are highly reliable. This was confirmed by a 4 × 4 × 4 analysis of variance (ANOVA): Main

Table 1. Frequency distribution of scores for each set

| | Subranges of scores | | | | |
|---|---|---|---|---|---|
| Sets | 50–59 | 60–69 | 70–79 | 80–89 | 90–100 |
| Bell | 11 | 17 | 42 | 18 | 12 |
| *U* | 31 | 13 | 11 | 13 | 32 |
| Positive | 42 | 28 | 14 | 10 | 6 |
| Negative | 5 | 10 | 14 | 27 | 44 |

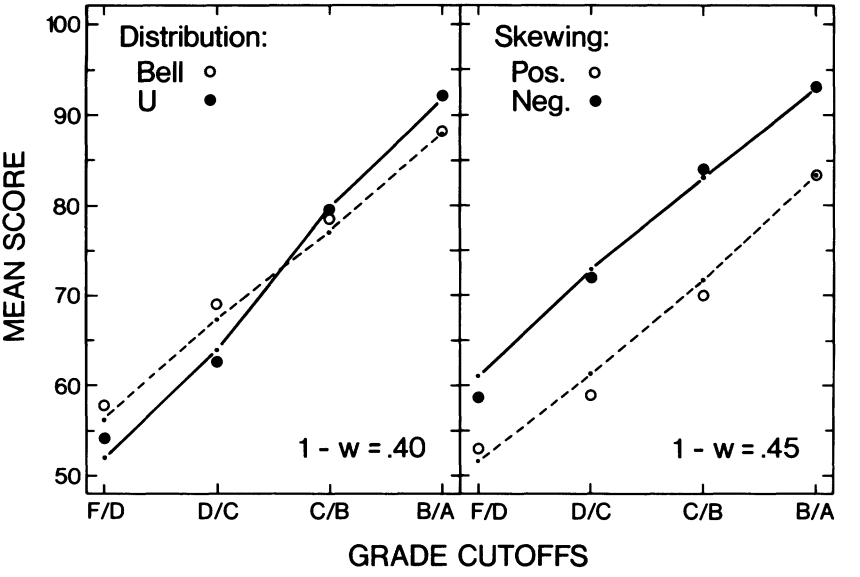


Figure 1. Mean cutoff scores for grading different sets of test scores in Experiment 1 (lines represent fits of range-frequency model)

effect of distribution, $F(3, 216) = 165.6$, $p < .0001$, and interaction between distribution and grade cutoffs, $F(9, 648) = 63.4$, $p < .0001$. Planned comparisons were used to evaluate the effects of distribution for positive/negative and bell/*U* sets separately because different distribution effects (as shown in Figure 1) are predicted by range-frequency theory for the two pairs of complementary sets. For the positive/negative sets, the effects of skewing are described primarily by the main effect of distribution, which was highly significant, $F(1, 72) = 233.3$, $p < .0001$. For the bell/*U* sets, the effects of distribution are described by the interaction between distribution and cutoffs, which was also highly significant, $F(3, 216) = 90.24$, $p < .0001$. Neither the main effect of order of presentation (viz., order of distributions in the booklet) nor any of the interactions with order were significant, $p > .10$. The only other statistically significant effect was the main effect of grade cutoffs, $p < .0001$.

**Fit of the range-frequency model**

Figure 1 also shows how well the grade assignments reflect a range-frequency compromise. The cutoff scores entailed by the frequency principle alone were first calculated for each distribution (viz., the D/F cutoff was the score at the 20th percentile, the C/D cutoff was the score at the 40th percentile, etc.).

Estimation of the frequency weighting ($1 - w$) requires a comparison

of judgments obtained from at least two different distributions. Transposing Equation 1 for each distribution and subtracting yields:

$$J_{i1} - J_{i2} = w(R_{i1} - R_{i2}) + (1 - w)(F_{i1} - F_{i2}), \quad (2)$$

where the 1 and 2 indicate different distributions. Because extreme scores are the same for all exams, range values for the different distributions are assumed to be equal and thus drop out so that

$$1 - w = (J_{i1} - J_{i2})/(F_{i1} - F_{i2}). \quad (3)$$

Estimates of $1 - w$ were calculated separately for the two pairs of sets, bell/*U* and positive/negative, by averaging the estimates from application of Equation 3. The inferred weightings, shown in Figure 1, are close to the typical value of 0.5, indicating a roughly equal compromise between the range and frequency principles.

Cutoffs following the range principle were then derived by substituting the empirical values, frequency values, and the derived weighting value into Equation 1. On the basis of previous work with numerals (Parducci et al., 1960), these range values were assumed to be linear to numerical values. Consequently, the inferred range values from the four distributions were fit by linear regression to obtain a single range function for all four sets. The subjective range of scores inferred in the range-frequency fit to the students' ratings extends down to 36.78, about halfway between pure guessing (25) and the lowest actual score (50). The upper end of this subjective range is 102.81, only slightly above the upper value in the experimental set.

Predicted judgments were then obtained by substituting these inferred range values, the weighting value, and the a priori frequency values into Equation 1. The fit of the range-frequency model captures the general form of the grading scales quite well using only four fitted parameters, the two estimates of $w$ and the slope and origin of the single linear range function. Although the predicted points deviate significantly from the data points (the mean of absolute deviations is 2.39 standard errors), the significance of these deviations seems attributable to the power of the study. The average absolute deviation of predicted cutoffs from actual cutoffs is only 1.16 exam score units.

**Individual differences**

To evaluate individual differences in the relative weighting of range and frequency tendencies, $1 - w$ was determined separately for each subject by dividing differences in the cutoff scores by the differences predicted given $1 - w = 1.0$ for each of the two pairs of sets and then averaging these two estimates. A plot of the cumulative percentage of subjects at each value of $1 - w$ was sigmoidal in shape,

indicating that frequency weightings for individuals approximate a normal distribution around a central value of $1 - w$ slightly below 0.5. A few subjects fall at either extreme, 8% not shifting their grading cutoffs at all and 3% assigning grades with approximately the same frequencies for the different distributions (grading on a curve); however, most subjects compromise between these extremes (the frequency weighting for 75% of the subjects was between .25 and .75).

How consistent are individual subjects in weighting range and frequency tendencies? The correlation between the two sets of weighting values inferred for each subject is 0.61. If each pair of distributions is considered as a parallel item from a single test, the reliability of $w$ based on all four distributions is estimated by the Spearman-Brown formula:

$$r_{tt'} = 2r_{aa'}/(1 + r_{aa'}). \quad (4)$$

The obtained reliability coefficient, $r_{tt'} = 0.76$ indicates some consistency of individual differences, but it also suggests that particular subjects may be affected differently by the differences between distributions.

**Test ratings**

Mean ratings of how well the test measured knowledge were 2.9, 3.2, 3.3, and 3.6, for positive, *U*, negative, and bell sets, respectively. These differ significantly, $F(3, 225) = 6.01$, $p < .001$. Least significant difference tests (conducted at $p < .05$) showed that the positive set was rated significantly lower than the other three sets and that the bell set was rated significantly higher than the positive and *U* sets.

One way in which the experimental situation differed from assignment of grades in a real course situation is that the subject had no contact with the materials (or students) responsible for generating the distributions of scores. In the absence of such information, the judge must make inferences about the test materials from the scores themselves. Because actual exam scores tend to be normally distributed for well-constructed, norm-referenced, multiple-choice tests, students' preference for the test producing the bell distribution (rating it as a better test of knowledge) seems encouragingly appropriate.

Table 2 shows the percentage use of each grade for the different distributions. It is interesting that students assign grades for the bell set that approximate cutoffs actually recommended by their (UCLA) Psychology Department (A, 15%; B, 25%; C, 45%; D, 10%; F, 5%). Surprisingly, where the students differed from the recommended curve was in their greater willingness to assign D's and F's (with fewer C's).

Table 2. Experiment 1: Percentage of scores assigned to each grade

| | Grade | | | | |
|---|---|---|---|---|---|
| Sets | F | D | C | B | A |
| Bell | 8.01 | 18.42 | 34.46 | 23.86 | 15.25 |
| *U* | 14.34 | 18.76 | 21.68 | 19.68 | 25.53 |
| Positive | 13.75 | 24.03 | 29.41 | 20.09 | 12.72 |
| Negative | 5.29 | 12.55 | 21.38 | 26.45 | 34.33 |
| Mean | 10.35 | 18.44 | 26.73 | 22.52 | 21.95 |

Still, the overall percentages indicate that students are more reluctant to assign D's and F's than higher grades.

**A modified range-frequency model**

An alternative interpretation of the frequency principle would assign grades with fixed but unequal frequencies; for example, grade assignments might reflect a compromise between following a "straight" scale (with cutoffs 90, 80, 70, 60) and following the recommended departmental curve described above. However, fitting the modified range-frequency model to the data involves estimation of additional parameters for the relative frequencies of four of the five categories (for a total of eight parameters to fit 16 data points). Not only would fitting parameters for *both* range and frequencies be unparsimonious (conceptually and statistically), but there is no guarantee of a unique solution. Thus, in fitting this model we chose to estimate only the two values of $w$ (for bell/$U$ and positive/negative sets) and the four parameters for frequencies of category use. The range was equated a priori with the actual range of scores (50–100).

The fitting procedure for this model parallels that for the standard model except that the unit of analysis is at the frequency of category usage rather than the score assigned to each cutoff. Equation 1 can be rewritten in terms of frequencies of category usage as follows:

$$f[\mathrm{J}_{ic}] = wf[\mathrm{R}_{ic}] + (1 - w)f[\mathrm{F}_{ic}], \tag{5}$$

where the function $f$ describes the relative frequency of use of each category ($i$). Transposing Equation 5 for each distribution and subtracting yields:

$$f[\mathrm{J}_{i1}] - f[\mathrm{J}_{i2}] = w(f[\mathrm{R}_{i1}] - f[\mathrm{R}_{i2}]) + (1 - w)(f[\mathrm{F}_{i1}] - f[\mathrm{F}_{i2}]), \tag{6}$$

but because the modified frequency principle requires the same set of unequal frequencies for all distributions, the equation reduces to:

$$w = (f[\mathrm{J}_{i1}] - f[\mathrm{J}_{i2}])/(f[\mathrm{R}_{i1}] - f[\mathrm{R}_{i2}]). \tag{7}$$

Estimates of $w$ were calculated separately for the two pairs of sets, bell/$U$ and positive/negative, by averaging the estimates from application of Equation 7. The inferred frequency weightings, 0.48 and 0.55, are slightly higher than those inferred for the equal frequency model.

The grade frequencies dictated by the frequency principle, $f[F_{ic}]$, were then derived by substituting the values for $f[J_{ic}]$, $f[R_{ic}]$, and $w$ into Equation 5. A single set of frequencies was determined by averaging the estimates from each distribution. These represent the fixed, but unequal frequencies, dictated by the frequency principles. Predicted grade cutoffs were determined by substituting the weighting value ($w$) and the score cutoffs dictated by range and frequency principles into Equation 1.

The best-fitting frequencies per grade were as follows: 19.9% A's, 26.0% B's, 31.9% C's, 18.5% D's, and 3.6% F's, almost as skewed as the grading curve actually used by UCLA instructors. This model fits slightly better than the equal-frequency model (used to fit Figure 1); the mean of the absolute differences between theoretical and obtained cutoffs is 1.82 standard errors. However, this relatively modest increase in goodness of fit is at the cost of estimating two additional parameters. Because the frequency weighting for this fixed-frequency model is still very close to 0.5, the grading behavior of subjects in the present experiment can be characterized as an equal compromise between using a "straight scale" and grading on the "curve," regardless of which interpretation of the frequency principle is applied. Experiment 2 is designed to test between the equal- and unequal-frequency models.

## EXPERIMENT 2

Modifying the frequency principle so that some grade categories are used more frequently than others has an intuitive appeal—there is widespread reluctance to fail people. This version of the frequency principle has been explored by Birnbaum (1974) for nine-category ratings of numerical magnitude; but even though additional parameters were estimated to fit the empirical data, the advantage of assuming the unequal frequencies was small. Experiment 2 attempts to distinguish between the different range-frequency models for grade assignment by increasing the range of scores so that the lowest of the presented scores is the value expected for pure guessing. This value is assumed to represent a natural lower boundary on the range. The fit of the range-frequency model (assuming an equal frequency prin-

ciple) to the data of Experiment 1 required that the subjective range extend well below the lowest score in the set. If the inferred subjective range for Experiment 2 should extend well below pure guessing, the assumption of equal frequencies would become less attractive. On the other hand, if the particular unequal frequencies required to fit the data of Experiment 2 (under the assumed matching of subjective to objective range) should be radically different from those derived in Experiment 1, the assumption of fixed but unequal frequencies would become less attractive.

## METHOD

Experiment 2 differed from Experiment 1 in that the 100 numerals ranged from 25 to 100 instead of 50 to 100 (25 being the expected score for pure guessing on these "four-choice" tests). The four stimulus sets constructed for Experiment 2 were in other respects as similar as possible to those of Experiment 1. The relative frequencies of scores for the five subranges (25–39, 40–54, 55–69, 70–84, 85–100) were the same as shown for the five subranges in Table 1. Because the order of the four sets had no effect in Experiment 1, this was not analyzed in Experiment 2. Instead, data from eight counterbalanced orders were grouped solely by set.

The instructions were essentially the same as in Experiment 1 except that they emphasized that a score of 25 was what would be expected by chance. Ratings of the tests were not obtained in Experiment 2.

Subjects were 40 undergraduate students drawn from the same population and tested in the same manner as in Experiment 1. Data from one subject, who failed to indicate some of the grade cutoffs, were dropped from the experiment.

## RESULTS AND DISCUSSION

Figure 2 presents the mean scores for each grade cutoff along with predictions and $1 - w$ values for the equal-frequency version of the range-frequency model. Again, the effects of distribution and cutoff are obviously highly reliable (average standard error is only 1.083). The fitting procedure was the same as that described for Experiment 1. The new derived range function varied from 26.34 (close to guessing level, 25) to 109.98 (somewhat higher than the observed maximum of 100). This model fits the data well: The mean of the absolute differences between theoretical and empirical points was only 1.11 standard errors, with only two points erring by more than 2 standard errors.

Table 3 shows the percentage use of each grade for the four distributions. As compared with Experiment 1 (Table 2), the largest

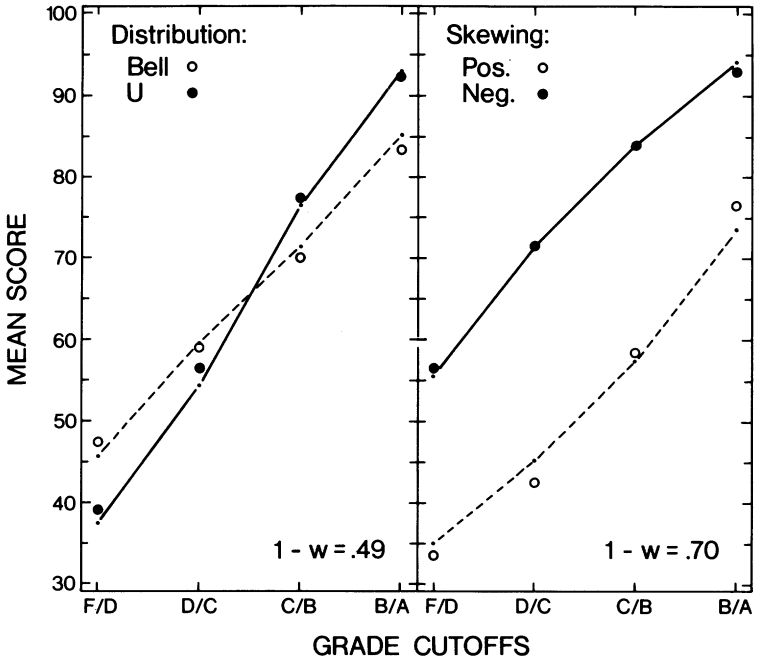


Figure 2. Grades and range-frequency fits as in Figure 1 (Exp. 2)

Table 3. Experiment 2: Percentage of scores assigned to each grade

| | Grade | | | | |
|---|---|---|---|---|---|
| Sets | F | D | C | B | A |
| Bell | 19.85 | 18.95 | 29.31 | 18.23 | 13.67 |
| *U* | 27.62 | 16.33 | 17.13 | 16.64 | 22.28 |
| Positive | 23.95 | 21.90 | 25.85 | 16.23 | 12.08 |
| Negative | 17.10 | 15.39 | 21.59 | 21.92 | 24.00 |
| Mean | 22.13 | 18.14 | 23.47 | 18.26 | 18.00 |

difference is in the higher proportion of F's. Extending the experimental range tends to equalize the empirical category frequencies.

In general, the results of Experiment 2 favor assumption of an equal-frequency principle over the assumption of fixed but unequal frequencies for assignment of grades. Assuming a range of 25 to 100, the best-fitting frequencies per grade are 15.2% A's, 19.0% B's, 24.2% C's, 19.0% D's, and 22.5% F's for the fixed-frequency model. These inferred frequencies are much more equal and skewed in the opposite direction than those inferred for the data of Experiment 1. When

frequencies from Experiment 1 are used to fit Experiment 2 data, the fit is very poor (the mean of the absolute deviations from predicted points is 5.22 standard errors) and the inferred subjective minimum is an implausible 95.5. Not surprisingly, when the more equal inferred frequencies from Experiment 2 are used to fit Experiment 1 data, the results are very similar to those for the equal-frequency model.

Finally, a nonlinear optimization procedure was used to fit the data from the combined set of experiments allowing the subjective ranges, $w$ values, and frequencies per grade category to vary for a total of 12 free parameters to fit 32 data points (as opposed to the 8 free parameters of the equal-frequency model).[2] Table 4 compares the fits of the two models. Although freeing four additional parameters provides a slightly better fit, the inferred frequencies for the different grade categories are nearly equal, except for the somewhat higher frequency of C's. Interestingly, the subjective ranges inferred for this model do not vary much across the rather strong manipulation of the experimental range. Although the greater proportional use of C's in these particular experiments may be due to the use of a modified frequency principle, we would argue that the more parsimonious fit of the equal-frequency model suggests that it provides a better approximation of student grading behavior.

A surprising finding of Experiment 2 is the large discrepancy in the value of the frequency weighting for bell/$U$ versus positive/negative sets (0.49 vs. 0.70). In Experiment 1 the difference in frequency weightings for the two pairs of sets was in the same direction but not nearly so great, possibly because the range cutoffs (90, 80, 70, 60) are so familiar to students. The greater frequency weighting for skewed sets may be due to their frequencies being more salient and hence easier to apprehend (cf. Parducci & Wedell, 1986). The shift in the mean frequency weighting also results in less stable individual differences. The correlation between the two frequency weighting estimates calculated for each student ($r$ = .43, $r_{tt'}$ = .60) is significantly lower than the correlation from Experiment 1 ($p < .01$, two-tailed test).

**Implications**

Experiments 1 and 2 suggest that the use of absolute standards would be perceived as unfair by students. When asked to assign grades as fairly as possible, students are extremely sensitive to percentile ranks of exam scores. Would instructors show this same type of relativism? Those who adhere strictly to a traditional grading curve are conforming to an unequal-frequency principle (with a range-frequency value of $w = 0$). The present experiments suggest that students regard this formula as less appropriate when scores do not approximate a

Table 4. Fit of equal- and unequal-frequency models to data from Experiments 1 and 2

| Model | Frequencies | | | | | Subjective range | | $w_{pos/neg}$ | $w_{bell/U}$ | Mean of absolute deviations |
|---|---|---|---|---|---|---|---|---|---|---|
| | A's | B's | C's | D's | F's | Minimum | Maximum | | | |
| Equal frequency | | | | | | | | | | |
| Exp. 1 | 20.00 | 20.00 | 20.00 | 20.00 | 20.00 | 36.78 | 102.81 | 0.546 | 0.602 | 1.155 |
| Exp. 2 | 20.00 | 20.00 | 20.00 | 20.00 | 20.00 | 26.34 | 109.98 | 0.303 | 0.494 | 1.202 |
| Unequal frequency | | | | | | | | | | |
| Exp. 1 | 17.69 | 20.00 | 25.70 | 18.36 | 18.25 | 38.06 | 101.48 | 0.525 | 0.602 | 1.105 |
| Exp. 2 | 17.69 | 20.00 | 25.70 | 18.36 | 18.25 | 34.16 | 102.22 | 0.239 | 0.390 | 1.059 |

*Note.* Fitted parameters are underlined; pos/neg = positive/negative.

normal distribution. In such cases, the test itself is seen as a poorer measure of students' knowledge, and students appear to favor a range-frequency compromise over more traditional methods of grade assignment.

The analysis of individual differences produced two revealing results. First, the fact that nearly all subjects follow a compromise between range and frequency principles indicates that the "averaging" process described by the model takes place within subjects and is not just the result of averaging across subjects using different strategies. Second, the reliability of these individual differences suggests that susceptibility to shifts in contextual stimuli may represent a type of "cognitive style," perhaps related to field dependence and leveling-sharpening distinctions raised in the literature (Hettema, 1968). However, some caution should be exercised in generalizing these results beyond the particular task employed. In assigning grades, subjects may have deliberately made an effort to follow a consistent rule, thus enhancing the reliability of this measure. Furthermore, the discrepancy between inferred values of $w$ for bell/$U$ versus positive/negative distributions in Experiment 2 implies that subtle situational factors may strongly influence $w$, reducing the proportion of variance attributable to individual differences.

Finally, Experiments 1 and 2 present a strategy for determining whether the equal- or unequal-frequency version of the model is most appropriate. The greater number of parameters used by the unequal-frequency model will typically provide a better fit to the data, but at the cost of parsimony. Beyond simply cross-validating those fitted frequencies on a new sample, our strategy is to utilize a stimulus range that can be equated with the subjective range. This is done by choosing extreme values that lie at or near natural boundaries. For exam scores, we chose the values for "pure guessing" and 100% correct. For psychophysical judgments of line lengths, the natural boundaries would correspond to the longest and shortest lines that could possibly be presented. By controlling the subjective range, a more powerful test of the nature of the frequency principle is possible. The results of Experiment 2 imply that the general tendency to avoid use of the lower grade categories (D and F) in Experiment 1 was not due to use of an unequal-frequency principle, but rather to the extension of the subjective range below the lowest score. This result is consistent with results of other recent experiments using social stimuli in which we have found a similar tendency to extend the range of values considered at the time of judgment beyond the set actually presented (Wedell & Parducci, 1988; Wedell, Parducci, & Geiselman, 1987).

## Notes

Preparation of this article was completed in part while Douglas H. Wedell was at the University of Illinois on a postdoctoral traineeship, ADAMHA National support award MH14257, Lawrence E. Jones, Training Director.

Requests for offprints should be addressed to Douglas H. Wedell, Department of Psychology, University of Illinois, 603 E. Daniel St., Champaign, IL 61820. Received for publication January 6, 1988; revision received May 25, 1988.

1. Although recent formulations of range-frequency theory have been in terms of predicting the mean judgment of each *stimulus,* the present article follows Parducci's (1965) formulation in terms of predicting the stimulus value that defines each *cutoff.* This approach was chosen because it facilitates analysis of individual judgment functions, which tend to be stepwise in form.

2. D. H. Wedell combined the two fitting procedures (Equations 1, 2, 3, 4, 5, 6, 7) in an iterative program with the loss function defined by the mean of absolute deviations from empirical cutoffs.

## References

Birnbaum, M. H. (1974). Using contextual effects to derive psychophysical scales. *Perception & Psychophysics, 15,* 89–96.

Hettema, J. (1968). Cognitive abilities as process variables. *Journal of Personality and Social Psychology, 10,* 461–471.

Mellers, B. A. (1982). Equity judgments: 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, 8,* 242–270.

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

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

Parducci, A. (1983). Category ratings and the relational character of judgment. In H. G. Geissler & V. Sarris (Eds.), *Modern trends in perception* (pp. 89–105). Berlin: VEB Deutscher Verlag der Wissenschaften.

Parducci, A., Calfee, R. C., Marshall, L. M., & Davidson, L. P. (1960). Context effects in judgment: Adaptation level as a function of the mean, midpoint, and median of the stimuli. *Journal of Experimental Psychology, 60,* 65–77.

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.

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.