Meaning and Measurement of Performance Rating Accuracy: Some Methodological and Theoretical Concerns

Lorne M. Sulsky and William K. Balzer
Bowling Green State University

We examined methodological and theoretical issues related to accuracy measures used as criteria in performance-rating research. First, we argued that existing operational definitions of accuracy are not all based on a common accuracy definition; we report data that show generally weak relations among different accuracy operational definitions. Second, different methods of true score development are also examined, and both methodological and theoretical limitations are explored. Given the difficulty of obtaining true scores, criteria are discussed for examining the suitability of expert ratings as surrogate true score measures. Last, the usefulness of using accuracy measures in performance-rating research is examined to highlight situations in which accuracy measures might be desirable criterion measures in rating research.

Recently, researchers have raised several methodological and theoretical concerns about the usefulness of rating error indexes such as halo and leniency as surrogate measures of rating accuracy. One methodological limitation noted is the existence of various operational definitions for each of the error measures (Murphy & Balzer, 1981; Saal, Downey, & Lahey, 1980). The different statistical assumptions underlying these multiple definitions have been shown to lead to differences in conclusions drawn depending on the operational definition used (Murphy & Balzer, 1981; Saal et al., 1980).

Theoretical concerns about error measures have also been raised. Cooper (1981) pointed out that patterns of high intercorrelations among dimension ratings, thought to reflect halo error, may not be error at all. For example, if ratesu perform at similar levels across performance dimensions, high rating dimension intercorrelations would actually reflect a correct rank ordering of ratesu's performance. A similar argument may be extended to other rater error indexes as well (Murphy & Balzer, 1981). In addition, Murphy and Balzer (1981) empirically demonstrated across several samples that error measures were at best weakly correlated with direct accuracy measures, providing further support for the notion that error measures may be invalid, indirect indexes of rating accuracy.

As an alternative to the more common rater error criteria, increased attention has been given to the use of accuracy scores (e.g., Borman, 1977; McIntyre, Smith, & Hassett, 1984; see also Vance, Kuhnert, & Farr, 1978, and Dipboye, Stramler, & Fontenelle, 1984, for the use of accuracy scores in selection interview research). The particular advantage of using accuracy scores is that they provide a direct, rather than indirect, measure of accuracy. Furthermore, accuracy scores make no assumptions regarding the actual distribution of ratee performance (i.e., traditional rater error measures are computed under the assumption that performance ratings are normally distributed with zero correlations among rating dimensions; Schwab, Heneman, & DeCotiis, 1975). Given these advantages one might expect accuracy scores to play an increasing role as dependent measures in understanding the determinants of accuracy and error in performance measurement (Dunnette & Borman, 1979; but see Zedeck & Caspio, 1982).

Unfortunately, methodological and theoretical limitations also apply to accuracy scores, requiring caution in the use and interpretation of accuracy scores as dependent measures in performance rating research. Our central thesis is that the potential adoption of accuracy scores over error indexes simply replaces one set of methodological and theoretical concerns with another.

This article addresses a number of different issues relating to the meaning and measurement of performance rating accuracy. First, we review the concept of accuracy and the use of accuracy scores in performance rating research. We argue that various methodologically distinct measures are based on differing conceptualizations of accuracy. Second, in light of the theoretical and methodological issues raised, we report data that examine the relationships among the various measures reviewed. Third, we examine methodological and theoretical limitations with true score procedures presently used to compute accuracy measures. We point out that the different procedures that exist for developing true scores may not individually produce equivalent results. Furthermore, we argue that, at a theoretical level, it is impossible to develop true scores. Given this, we discuss the suitability of expert ratings as an appropriate standard of comparison for subject ratings. Finally, we examine the usefulness of using accuracy measures requiring the laborious and perhaps costly process of obtaining expert ratings. In short, we attempt to answer the question: Are these measures worth the effort and expense involved to produce them in the first place?

Meaning and Measurement of Rating Accuracy
Accuracy of measurement is a term used to describe both the strength and kind of relation between one set of measures and

A version of this article was presented at the annual convention of the Society of Industrial and Organizational Psychology in Atlanta in 1987.
We thank Robert M. Guion, Kevin R. Murphy, and two anonymous reviewers for their helpful comments on an earlier draft of this article.
Correspondence concerning this article should be addressed to William K. Balzer, Department of Psychology, Bowling Green State University, Bowling Green, Ohio 43403.
a corresponding set of measures (e.g., true scores) considered to be an accepted standard for comparison (Guion, 1965). To understand this conceptualization of accuracy, it is useful to draw a distinction between accuracy and the concepts of reliability and validity. Although all three psychometric indexes are functions of the strength of relation between a set of scores and a corresponding set of true scores, accuracy is also a function of the kind of relation between the two score distributions (Gordon, 1970). Thus, reliability and validity are necessary but insufficient conditions for accuracy. As we will see later, however, there are occasions when it may be unnecessary to establish the accuracy of ratings with evidence of reliability and validity being sufficient to judge rating quality.

In general, the accuracy of performance ratings for an individual rater have been computed by comparing the rater’s performance evaluations for n raters on k performance dimensions with corresponding evaluations provided by “expert raters.” Expert ratings are typically obtained by computing (across experts) pooled averages scores, yielding true score ratings for each rater on every performance dimension (cf. Borman, 1977). Individual rater scores (n raters × k dimensions) are then compared with the true scores (n raters × k dimensions); the closer the rater’s ratings are to the true score ratings, the more accurate the ratings are thought to be (Borman, 1977).

There are, of course, occasions when hard or objective performance criteria (e.g., dollar amount sold, number of accidents) are relied on for evaluating performance (Landy & Farr, 1983; Smith, 1976). The accuracy of these measures may sometimes be questioned; for example, if an accident measure is contaminated or deficient in certain respects (e.g., inaccurate reporting or recording of work-related accidents; see Landy & Farr, 1983, chapter 2, for many excellent examples), performance will not be accurately reflected. However, because the performance accuracy literature has focused on accuracy as it relates to judgmental ratings, we will restrict attention to judgmental performance measures. Nevertheless, the notion of accuracy for objective performance measures is a potentially important issue albeit beyond the scope of this article.

Operational Definitions and Conceptualizations of Accuracy: A Review

Table 1 presents the published performance appraisal research using rater accuracy measures. The table includes articles found from 1970 to the present in the Journal of Applied Psychology, Personnel Psychology, Academy of Management Journal, Academy of Management Review, and Organizational Behavior and Human Performance (after 1984, Organizational Behavior and Human Decision Processes). In addition, known relevant studies from other journals were also included. These studies served as a basis for identifying previously used conceptualizations and operational definitions of accuracy.

Although all of the operational definitions of performance rating accuracy include a comparison of the rater’s ratings with the true scores of ratee performance, a variety of different methods exist for comparing the two sets of ratings. A review of these methods clearly shows, however, that they do not all share a common conceptual base. In this section, we present different existing operational definitions of accuracy and examine their underlying conceptualizations.

Table 1: Performance Appraisal Studies Using Accuracy Measures

<table>
<thead>
<tr>
<th>Measure and citation</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cronbach’s accuracy scores</td>
<td>Also computed correlational components of each score</td>
</tr>
<tr>
<td>Becker &amp; Cardy, 1986</td>
<td>Only computed differential accuracy</td>
</tr>
<tr>
<td>Cardy &amp; Dobkins, 1986</td>
<td>Did not compute elevation</td>
</tr>
<tr>
<td>Murphy &amp; Balzer, 1986</td>
<td>Only computed differential accuracy</td>
</tr>
<tr>
<td>Pulakos, 1986</td>
<td>Dichotomous scoring procedure was used</td>
</tr>
<tr>
<td>Murphy, Balzer, Kellam, &amp; Armstrong, 1984</td>
<td>All studies computed DA separately for each dimension and/or</td>
</tr>
<tr>
<td>Bernardin, Cardy, &amp; Carlye, 1982</td>
<td>computed an overall DA score across dimensions</td>
</tr>
<tr>
<td>Murphy, Garcia, Kerkar, Martin, &amp; Balzer, 1982</td>
<td></td>
</tr>
<tr>
<td>Distance accuracy</td>
<td></td>
</tr>
<tr>
<td>McIntyre, Smith, &amp; Hassett, 1984</td>
<td></td>
</tr>
<tr>
<td>Heneman &amp; Wexley, 1983</td>
<td></td>
</tr>
<tr>
<td>Bernardin et al., 1982</td>
<td></td>
</tr>
<tr>
<td>Bernardin &amp; Pence, 1980</td>
<td></td>
</tr>
<tr>
<td>Gordon, 1970</td>
<td></td>
</tr>
<tr>
<td>Borman’s differential accuracy (DA)</td>
<td></td>
</tr>
<tr>
<td>Becker &amp; Cardy, 1986</td>
<td></td>
</tr>
<tr>
<td>Cardy &amp; Kehoe, 1984</td>
<td></td>
</tr>
<tr>
<td>McIntyre et al., 1984</td>
<td></td>
</tr>
<tr>
<td>Pulakos, 1984</td>
<td></td>
</tr>
<tr>
<td>Bernardin et al., 1982</td>
<td></td>
</tr>
<tr>
<td>Zecek &amp; Cascio, 1982</td>
<td></td>
</tr>
<tr>
<td>Borman, 1979a</td>
<td></td>
</tr>
<tr>
<td>Borman, 1979b</td>
<td></td>
</tr>
<tr>
<td>Borman, 1977</td>
<td></td>
</tr>
<tr>
<td>Halo-type accuracy</td>
<td></td>
</tr>
<tr>
<td>Pulakos, Schmitt, &amp; Ostroff, 1986</td>
<td>Percentile measures used to determine leniency</td>
</tr>
<tr>
<td>McIntyre et al., 1984</td>
<td></td>
</tr>
<tr>
<td>Pulakos, 1984</td>
<td></td>
</tr>
<tr>
<td>Leniency measures</td>
<td></td>
</tr>
<tr>
<td>Farh &amp; Werbel, 1986</td>
<td></td>
</tr>
<tr>
<td>McIntyre et al., 1984</td>
<td></td>
</tr>
<tr>
<td>Pulakos, 1984</td>
<td></td>
</tr>
</tbody>
</table>

Cronbach Accuracy Component Scores

One set of accuracy measures that has been used in performance-rating research was formulated by Cronbach (1955) and is based on the earlier $D^2$ index, an index that represents the squared difference between subject ratings ($x$) and true scores ($t$) averaged across $n$ raters and $k$ dimensions:

$$D^2 = \frac{1}{kn} \sum_{ak} (x_{ak} - t_{ak})^2.$$ (1)

Although the $D^2$ index had been used previously in research investigating interpersonal perception (see Funder, 1987, for a review), Cronbach (1955) argued that the index collapses across potentially important information and that this information may be meaningful for understanding different aspects of accuracy in interpersonal perception. Cronbach proposed one solution to this problem by decomposing the overall distance between rater ratings and true scores into four separable and con-
ceptually independent component accuracy scores. Specifically, these components were derived following the logic of analysis of variance (ANOVA) with each component expressing a different portion of the distance between rater ratings and true scores similar to the partitioning of variance in a two-way ANOVA. These components may therefore be described using ANOVA terminology: (a) elevation (E), the differential grand mean; (b) differential elevation (DE), the differential main effect of raters; (c) stereotype accuracy (SA), the differential main effect of dimensions; and (d) differential accuracy (DA), the Differential Ratee × Dimension interaction.

Following the formulae provided by Cronbach (1955) and Murphy, Garcia, Kerkar, Martin, and Balzer (1982), the squared component scores may be expressed by the following terms, in which smaller values denote greater accuracy:

\[ E^2 = (\bar{x} - \bar{t})^2 \]  

(2)

and

\[ DE^2 = \frac{1}{n} \sum_i [(\bar{x}_i - \bar{x}) - (\bar{t}_i - \bar{t})]^2, \]  

(3a)

or equivalently,

\[ DE^2 = \sigma_{x_i}^2 + \sigma_{t_i}^2 - 2\sigma_{x_i} \sigma_{t_i} r_{x_i t_i}, \]  

(3b)

and

\[ SA^2 = \frac{1}{k} \sum_j [(\bar{x}_j - \bar{x}) - (\bar{t}_j - \bar{t})]^2, \]  

(4a)

or equivalently,

\[ SA^2 = \sigma_{x_j}^2 + \sigma_{t_j}^2 - 2\sigma_{x_j} \sigma_{t_j} r_{x_j t_j}, \]  

(4b)

and

\[ DA^2 = \frac{1}{kn} \sum_{i} [(x_{ij} - \bar{x}_i - \bar{x}_j - \bar{x}) - (t_{ij} - \bar{t}_i - \bar{t}_j - \bar{t})]^2, \]  

(5a)

or equivalently,

\[ DA^2 = \sigma_{a}^2 + \sigma_{b}^2 - 2\sigma_{a} \sigma_{b} r_{ab}, \]  

(5b)

where \( a = x_{ij} - x_i - x_j - \bar{x} \), and \( b = t_{ij} - t_i - t_j - \bar{t} \), \( x_{ij} \) and \( t_{ij} \) = rating and true score for ratee \( i \) on item \( j \); \( x_i \) and \( t_i \) = mean rating and mean true score for ratee \( i \); \( x_j \) and \( t_j \) = mean rating and mean true score for item \( j \); and \( x_{..} \) and \( t_{..} \) = mean rating and mean true score, over all raters and items.

Inspection of Equations 3b, 4b, and 5b illustrates that DE, SA, and DA can be decomposed into separate variance and correlational components. Recently, Becker and Cardy (1986) suggested that the correlational components of DE, SA, and DA may be useful accuracy measures in rating research, reviving Cronbach’s (1955) contention that the separate variance and correlational terms carry individually meaningful information about accuracy.

Note that each of the four component measures is based on the psychometric conceptualization of accuracy noted earlier; the distance between rater ratings and the true scores is considered in each of the component measures. However, the correlational components of the latter three component measures (see Equations 3b, 4b, 5b) are not sensitive to the distances between subject and true score ratings and thus are not conceptually based on the psychometric definition of accuracy outlined earlier. Thus, it would be imprecise to label them as types of accuracy scores, inasmuch as correlations between subject and true scores only provide evidence of rating validity.

**Distance Accuracy**

Despite Cronbach’s (1955) admonition, some researchers have adopted a modified form of the overall \( D^2 \) index, termed distance accuracy (DA), based on the city block metric (Davidson, 1985). Distance accuracy indicates the average absolute deviation of subject ratings from true scores. For example, McIntyre et al. (1984) operationally defined distance accuracy as follows:

\[ DA = \frac{1}{n} \sum_{i} |t_{ij} - r_{ij}| \]  

Distance Accuracy = \[ \frac{1}{n} \sum_{i} \frac{d}{n}, \]  

(6)

where \( k \) refers to the \( k_{th} \) rater; \( n \) is the number of raters; \( d \) is the number of dimensions; \( r \) refers to subject rating; and \( t \) refers to true scores.

The relation between the distance accuracy measure and the overall \( D^2 \) measure is not straightforward at the operational level; inspection of the above formula reveals that the distance accuracy index is not simply equal to the square root of the overall \( D^2 \) index. Because the \( D^2 \) index is computed by squaring the differences between individual ratings (see Equation 1) prior to summing the differences, the two indices are not easily comparable; specifically, \( D^2 \) increases exponentially relative to distance accuracy as discrepancies between individual ratings and corresponding true scores increase.

Theoretically, the distance accuracy measure is similar to the Cronbach component scores and overall \( D^2 \) measures by considering the distances between rater ratings and corresponding true scores. Thus, distance accuracy is conceptually based on the psychometric definition of accuracy as well.

**Borman's Differential Accuracy**

Borman (1977) was one of the earliest researchers to use the Cronbach component scores in a performance rating context. However, Borman suggested that Cronbach’s differential accuracy is the only conceptually appropriate component score because it is typically important to discern how accurately raters can discriminate among raters on a number of performance dimensions. The formula for Borman’s DA measure may be expressed as follows:

\[ Borman's \ DA = \frac{1}{d} \sum_{j=1}^{d} T_{d}, \]  

(7)

where \( d \) refers to the number of dimensions and \( T_d \) refers to the correlation between ratings and true scores for a particular dimension, transformed to a Z score.

As the formula shows, Borman’s (1977) DA index is computed by correlating a rater’s ratings for each dimension with corresponding true scores across raters, yielding a DA score for each dimension. The overall DA score is computed by averaging
the correlations across dimensions using Fisher's $r$ to $z$ transformation (Borman, Hough, & Dunnette, 1978).

Becker and Cardy (1986) correctly noted that Borman's (1977) operational definition of DA is not equivalent to the Cronbach (1955) DA measure. In contrast to Cronbach's DA score (see Equation 5a), Borman's DA measure provides only correlational information; the actual distances between subject ratings and true scores are not considered (Becker & Cardy, 1986). Furthermore, the correlational information provided by Borman's DA is not equivalent to the correlational component of DA in Cronbach's formulation (see Equation 5b). Although Borman's DA has been termed correlational accuracy by McIntyre et al. (1984), it should not be confused with the correlational component of Cronbach's DA, labeled DAcOR by Becker and Cardy (1986).

Conceptually, as we have seen with the correlational components of Cronbach's scores, Borman's DA measure does not qualify as an index of accuracy because it is insensitive to distances between ratings and true scores (Gordon, 1970; Guion, 1965). Although this measure has been termed correlational accuracy, it is not an accuracy measure based on Guion's (1965) and others' definition of accuracy adopted throughout this article. By providing only correlational information, Borman's DA best qualifies as an index of rater validity. Thus, although this measure can provide important preliminary evidence for accuracy, it should not be labeled as an accuracy measure.

**Halo-Type Accuracy**

Perhaps at least partly motivated by the popularity of halo measures as rating criteria, recent attempts have been undertaken to improve on past operational definitions of halo by acknowledging that correlations among ratings across dimensions for a given rater may partly or wholly reflect actual covariation of performance across dimensions. In other words, these operational definitions incorporate the distinction between true halo and illusory halo (Cooper, 1981, 1983). These new and "improved" halo measures include either averaged true levels of dimensional intercorrelations or true variation in ratings across dimensions (averaged over raters) in the computations (McIntyre et al., 1984; Pulakos, 1984). For example, McIntyre et al. (1984) computed halo as follows:

$$\text{Halo}_k = \frac{r}{k} \sum_{i=1}^{k} \frac{(v_{ij} - v_{ik})}{r}, \tag{8}$$

where $k$ refers to the $k_{th}$ rater; $r$ is the number of raters; $v_{ij}$ is true variance for ratee $i$ across dimensions; and $v_{ik}$ is the rater's variance for ratee $i$ across dimensions.

Operationally, this type of halo measure appears to be similar to accuracy measures to the extent that a comparison takes place between ratings and true scores. Unlike Cronbach's scores and distance accuracy, however, the comparisons that take place are not between individual scores; rather, the comparisons are between correlations or variances. Because the distances between individual ratings and corresponding true scores are not considered, this index does not provide a direct estimate of rating accuracy. The question of whether these halo measures provide some indirect evidence for accuracy will be addressed later.

**Leniency Measures**

Recent operational definitions of leniency have computed both true and illusory leniency to obtain a measure of leniency or severity. For example, McIntyre et al. (1984) computed leniency as follows:

$$\text{Leniency}_k = \frac{d}{n} \sum_{j=1}^{d} \frac{(t_{ij} - r_{jk})}{d}, \tag{9}$$

where $k$, $n$, $d$, $r$, and $t$ are defined as in Equation 6.

Equation 9 provides a leniency (or severity) score for a rater by subtracting the rater's ratings from the true score ratings (averaged over raters and dimensions).

This type of leniency measure appears to be conceptually consistent with the meaning of accuracy; the final estimate is sensitive to the difference between the overall average true score ratings and the overall average rating assigned by a rater. In fact, this leniency measure, when squared, will produce identical results to Cronbach's elevation component score (see Equation 2). The difference between the two measures lies in the fact that the leniency or severity measure is a more descriptive measure inasmuch as it provides information about rater's tendency to be overly lenient or harsh; elevation simply indicates systematic differences between ratings and true scores.

**Summary**

All of the measures reviewed share one common feature: All require the direct comparison of ratings obtained from a single rater with true scores to compute an accuracy index. To properly reflect the psychometric conceptualization of accuracy, this comparison is a necessary feature of any accuracy score operational definition. Given the psychometric definition of accuracy, however, only some of the measures previously discussed can be properly considered accuracy scores.

Based on the previous conclusions, should we attempt to find a common label for these diverse sets of measures? To facilitate discussion about these measures, it is useful to have a common label that is reserved for rating criteria requiring a comparison between ratings and true scores. The term *comparison scores* captures what is common to the different measures and will be used from this point on as a general label.

**Relations Among Operational Definitions of Accuracy**

Given both differing conceptualizations of the accuracy construct and differing operational definitions of accuracy measures, the relations among these measures might not necessarily be strong and positive. Knowing the relations between the measures is important because weak relations imply that research findings may not generalize across the different measures.

Correlational evidence already exists suggesting that various comparison measures are not highly related. Low and nonsignificant correlations among the various Cronbach (1955) component scores have been found across a number of studies (e.g., Becker & Cardy, 1986; Cline, 1964; Murphy & Balzer, 1981; Murphy et al., 1982), supporting Cronbach's claim that each component score reflects an independent aspect of rating accu-
racy. Becker and Cardy (1986) demonstrated both empirically and theoretically that different relationships among halo and accuracy can be obtained depending on the particular halo and accuracy measures used. They also found a weak nonsignificant relation between Cronbach’s differential accuracy measure and Borman’s differential accuracy measure, supporting their claim that the two measures are not equivalent.

Although a number of relations among various comparison scores have been examined previously, we extend this research by examining additional relations among different comparison measures using specific operational definitions of halo and leniency not previously considered (Fisicaro, 1988; McIntyre et al., 1984; Pulakos, Schmitt, & Ostrom, 1986). Two previous studies, differing in both ratee and rater populations, were included to enhance generalizability. The first study was originally designed to examine the effects of behavioral diary keeping on rating accuracy (Sulsly & Balzer, 1986). A total of 90 undergraduate subjects each viewed four tapes randomly sampled from a population of eight tapes, with each tape depicting a teacher giving a brief lecture (see Murphy et al., 1982, for a detailed description of the tapes). Each subject rated each of the four teachers on eight performance dimensions using 5-point Likert-type scales.

The second data set was randomly sampled from a larger data set originally used to evaluate the impact of various training strategies on rating accuracy (Ruddy & Kavanagh, 1986). Ratings were collected posttraining across various experimental conditions. In all, 85 undergraduate subjects each viewed four tapes depicting a manager in a problem-solving situation (see Borman, 1977, for a full description of the tapes) and rated each manager performance on seven dimensions using 7-point scales.

For both data sets, true scores were available based on procedures developed by Borman (Borman, 1977; Murphy et al., 1982), and comparison scores were computed for each subject. The Cronbach component scores, distance accuracy, and Borman’s DA were operationally defined following Equations 2–7, respectively. Halo and leniency/severity were measured using modified versions of Equations 8 and 9, respectively. Specifically, the absolute values of the resulting estimates were taken to obtain final estimates for both measures. This was done so that any negative values for halo or leniency (i.e., severity) would be treated conceptually as error when correlations are performed and interpreted (Fisicaro, 1988). The operational definition of halo was further modified by first standardizing ratings to eliminate the effect of dimension mean differences, making the measure consistent with the conceptual definition of halo (Pulakos et al., 1986). Finally, following Murphy et al. (1982), we took the square roots of the Cronbach component scores as final estimates. Because this transforms the scale of measurement, we also analyzed the scores without the transformation. The effects of the transformation on the correlations were negligible, and thus we report only the square root measure results.

Correlations among the measures described earlier are presented in Table 2. Although several pairs of correlations between the two studies differ in magnitude, the differences are not significant ($p > .05$). Correlations across both data sets support the general claim that the relationships among the various measures are weak ($F = .19$). One notable exception is the statistically significant and relatively large correlation between leniency and distance accuracy measures. Also, significant correlations were obtained between Borman’s DA measure and distance accuracy, suggesting that there is at least some degree of relation between the ability to produce accurate ratings and the ability to correctly rank-order rates. This should not be surprising because the ability to correctly rank rates will often imply some level of rating accuracy (although this is not necessarily true). Statistically significant correlations were also found between Borman’s DA and differential elevation, and the magnitude of the correlation suggests that the relation has some practical significance. The common variance in these two measures is probably due to the common information about ratee rank order contained in both measures. Finally, of interest are results showing generally weak relations between halo and accuracy measures (i.e., based on distance information), with the few significant correlations suggesting a positive relation between accuracy and halo (these correlations are not reported in Table 2 because they are data dependent, as discussed later). Because previous research has revealed a paradoxical positive relation between halo and accuracy (Cooper, 1981; Fisicaro, 1988; Murphy & Balzer, 1986), this result may be expected. However, it can be argued at a theoretical level that either positive, negative, or zero relations between the two types of measures are possible in a given data set. This is because halo measures based on interdimensional correlations are insensitive to the distance between the rater’s ratings and true score estimates. Hypothetical data sets may be created to show any form of relation between halo and accuracy. Therefore, the relation between halo and accuracy is likely to be data dependent and unstable across data sets, even when levels of true halo are available.

Overall, the results show weak relationships among various measures based on differing conceptualizations of accuracy, differing operational definitions of accuracy, or both, replicating and extending earlier work. Taken together, these findings suggest that different measures may tap different facets of rating ability. This implies that one should choose comparison scores carefully because a given study may yield quite different results depending on the particular measure(s) chosen. Thus, the nature of any differences found among raters or experimental groups may depend on the comparison score measures used. We believe this to be a critical point because the ability to meaningfully integrate research across studies requires an understanding of the differences among comparison score measures used in performance rating research. One implication, for example, is that if a training program increased rating accuracy, this finding may not be meaningful without clearly specifying the specific type(s) of comparison scores used. In sum, these findings suggest that comparison score values may not only be influenced by the manipulation of independent variables, but also by the particular operational definition used when computing them.

**True Score Development: Methodological Issues**

Although all comparison score measures use true scores as the accepted standard for comparison, a review of the performance-rating articles using accuracy scores shows considerable variation in the procedures used to obtain true scores. Probably the simplest and most questionable procedure is to use the aver-
Table 2  
Correlations Among Comparison Measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Elevation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Differential elevation</td>
<td>-13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Stereotype accuracy</td>
<td>0.4</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Differential accuracy</td>
<td>0.06</td>
<td>-0.05</td>
<td>0.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Borman's differential accuracy</td>
<td>0.21</td>
<td>0.008</td>
<td>0.27**</td>
<td>0.31**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Distance accuracy</td>
<td>-0.06</td>
<td>-0.48</td>
<td>0.04</td>
<td>0.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Halo</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Leniency</td>
<td>1.00**</td>
<td>-0.13</td>
<td>0.16</td>
<td>0.03</td>
<td>0.30**</td>
<td>0.74**</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.00**</td>
<td>0.04</td>
<td>0.02</td>
<td>0.21*</td>
<td>0.06</td>
<td>0.58**</td>
<td>0.08</td>
<td></td>
</tr>
</tbody>
</table>

\* = denotes correlations not computed because the relationships are data dependent.  
\*\* = p < .05.  \*\*\* = p < .01.

The first entry is from Sulsky & Balzer (1986), n = 90; the second entry is from Ruddy & Kavanaugh (1986), n = 85. Correlations between Borman’s differential accuracy and other measures were reverse coded so that positive correlations reflect positive relations between the measures.

age of the scores provided by all subjects or raters as the true score because information about rater agreement or disagreement is lost (e.g., Bernardin & Pence, 1980). Also, one might legitimately question the use of undergraduates as expert raters, given their lack of training and experience with performance rating tasks.

A second procedure for developing true scores involves averaging the ratings of previously scaled written incidents included as information in raters’ performance profile (Becker & Cardy, 1986; Borman, 1975; Cardy & Keohoe, 1984; Zedeck & Cascio, 1982). Assuming the scaling of incidents was adequate (i.e., following the Smith and Kendall, 1963, procedures; this evaluation was not conducted here because the behaviorally anchored rating scales [BARS] scaling references used are not identified or are unpublished sources), the use of an average critical incident rating is still questionable because each incident may be differentially important in computing a true score across incidents.

A third strategy for computing true scores is to use a group of expert raters to provide the true scores. Expert raters often refer to graduate students in industrial–organizational (I–O) psychology with an implied expertise in the conceptual issues of performance rating (e.g., coursework in performance ratings) and experience as a rater (e.g., having viewed and evaluated a number of classroom instructors; Borman, 1977, 1978; McIntyre et al., 1984; Murphy et al., 1982), but may include practicing I–O psychologists with varying years of experience (Borman, 1977, 1978).

Different strategies for obtaining true scores from expert raters, however, have been used. McIntyre et al. (1984), for example, used graduate students’ consensus ratings to obtain performance-rating true scores. Although this strategy does result in an agreed-upon performance true score for rater on dimension, the consensus procedure does not address whether, or to what degree, individual expert judges disagreed. The negotiated final rating could be the product of either small or large initial disagreements among experts; large initial disagreements prior to the consensus process, however, question whether the final rating is an appropriate true score. McIntyre et al. (1984) did report low agreement among their six expert judges, but it is impossible to assess the impact of this disagreement on the resulting true scores.

The most involved strategy for obtaining true scores was developed by Borman (1977, 1978) and used by Murphy and colleagues (cf. Murphy et al., 1982). Expert raters are made thoroughly familiar with the rating task by, for example, previewing the rating scales used in the studies, reading raters’ scripts, and taking notes during multiple observations of ratee performance. This extended opportunity to review the relevant performance-related behavior is expected to lead to highly informed and reliable expert ratings (Borman, 1977). In addition, an indirect validation approach using an ANOVA framework (Kavanaugh, MacKinney, & Wolins, 1971) is used to assess whether the expert raters show adequate agreement among their ratings (i.e., a ratee main effect), whether experts can appropriately discriminate rates on different dimensions (i.e., a ratee by dimension interaction), and whether the expert ratings correlate with the intended true scores portrayed in the scripted performance (Borman, 1977).

Despite these elaborate procedures, questions may be raised regarding the precision of the true scores. First, there is no appropriate criterion for evaluating whether interexpert agreement is adequate; intraclass estimates of agreement show less than perfect agreement among expert judges. And as Borman (1978) pointed out, even under ideal conditions, experts often do not totally agree in their ratings of others. The strategy of correlating true score averages with intended ratee performance levels, although it provides some indication of similarity in rank ordering between raters and experts, also does not provide strong evidence for the precision of the true scores. First of all, this Pearson product–moment correlation fails to take into account mean differences between the two score vectors. In addi-
tion, the vector of true scores provided by the experts is a vector of average expert scores; the variability among experts in each score is not considered in this analysis (although it is evaluated in the main effect and interaction intraclass indexes).

In summary, various procedures for obtaining true scores exist, and each procedure may produce inadequate measures of performance true scores. Furthermore, any operational definition of comparison scores comparing subject ratings with these true scores may produce varying results depending on the particular procedure used. The different procedures for obtaining true scores may further contribute to the lack of convergence among comparison scores and make it difficult to determine the extent to which unique variance in a given score is due to the specific true score procedure involved. Thus, the understanding of the relations among the various comparison scores will be limited until we standardize true score procedures.

True Score Estimation: Theoretical Issues

Although we have argued that deficiencies exist with current true score procedures at the methodological level, it appears that there is a more serious theoretical problem that calls into question the veracity of any true score procedure. Specifically, the problem stems from a lack of congruence between the ways in which true scores are obtained and the definition of true scores.

By definition, a true score represents the mean of an infinite number of scores across parallel measures of a test (Allen & Yen, 1978). Therefore, true scores are theoretical constructs that cannot be obtained and can only be estimated. Although accuracy researchers have never directly stated that actual true scores are obtainable, the term true score is used loosely. Replacing the term true score with true score estimate is suggested because it is more precise and conceptually correct.

Although the distinction between true score values and estimates is merely semantic at one level, the distinction takes on importance when it is recognized that we can question whether we have obtained suitable estimates. Of course, the meaning of suitable is critical here and requires elaboration. According to R. M. Guion (personal communication, May 5, 1987), a carefully developed and generalizable standard is necessary if the terms accuracy and inaccuracy are to be meaningful. The procedures that have been used to estimate true scores all produce standards, but are they generalizable and carefully developed? It seems that none of the procedures explore the generalizability of the expert ratings and develop a standard that is fixed and reliable but perhaps numerically arbitrary. In the following paragraphs, we describe three important criteria for establishing the suitability of expert ratings as a standard for comparison.

Generalizability

The first criterion for investigating suitability of the expert ratings is their consistency or generalizability across various measurement conditions. This is a problem of reliability, and Cronbach, Gleser, Nanda, and Rajaratnam (1972) generalizability model of reliability is a useful tool in this regard. In this model, facets or variables are chosen by the researcher to create a universe of facets that may each account for rating variance. Each facet can be individually examined in a generalizability study to determine whether the resulting ratings can be generalized across values or conditions of the variable not included in the study. Thus, for example, we may examine expert raters by randomly sampling from the population of available experts (called the universe of conditions) and investigate whether experts explain significant rating variance. If experts do not account for significant variance, the model states that we may generalize the findings to the population of rating experts.

To illustrate how the generalizability conceptualization can be useful for thinking about expert ratings, we refer the reader to Figure 1.

The three dimensional figure illustrates a hypothetical universe with three variables (information source, sample of experts, time of measurement) in which each slab represents an individual element or condition for one of the variables (e.g., a specific sample of experts or information source). (Although other variables might be included, resulting in an n dimensional universe, we have restricted discussion to three variables to facilitate discussion.) Because a typical study purporting to estimate true scores will likely consider only one information source, one sample of experts, and one time of measurement, the universe associated with the study is illustrated by the shaded portion of Figure 1, which represents a universe consisting of one element for each of the three variables. To expand the scope of generalizability to include all possible information sources, for example, we would conduct a generalizability study.
If sources do not explain significant rating variance, we may then generalize the rating estimates to the population of sources by collapsing across sources, resulting in a two dimensional universe (Expert Sample × Time of Measurement).

Validity

Although we can examine the reliability of expert ratings using the generalizability model, establishing generalizability does not imply that valid ratings are produced. For example, we can establish that ratings are generalizable across raters, but shared systematic biases across raters (e.g., similar stereotypes held by a homogeneous group of experts) may exist that enhance reliability while lowering validity. Although Borman’s (1977) true score procedure is the only one that examines expert rating validity, the convergent validity technique used in the Borman procedure is not sensitive to the possibility that convergence among raters involves converging biases.

Establishing the construct validity of the expert ratings is important to ensure that an appropriate standard for comparison is being used. It makes little sense to compare subject ratings to an invalid standard because the resulting comparison measure becomes meaningless. Furthermore, as expert ratings become less valid as indicators of the performance construct of interest, the relation outlined at the outset between validity and accuracy no longer holds when one compares subject and expert ratings. Distances between the two sets of ratings become difficult to interpret in terms of accuracy; discrepancies amount to differences in scaling or calibration. Said another way, without validity, accuracy is not a meaningful construct as it has been defined in this article. Techniques for examining rater validity exist (cf. Dickinson, 1986) and should be used to bolster claims of validity.

Calibrating the Experts

The final criterion is the degree to which the experts are calibrated with respect to the rating instrument and to each other. Although the numbers on the scale are arbitrary, a certain level of performance should be calibrated to a specific scale point, and experts should be calibrated to each other so that a discrete sample of work performance is judged similarly across experts. This criterion is important because the resulting ratings should be fixed with respect to a standard of work performance. Thus, for example, being a 3 versus a 5 on communication takes on meaning with respect to performance standards.

To address this calibration issue, it might be useful to consider a strategy for developing expert ratings that is based on frame of reference training (Bernardin & Buckley, 1981; Pulakos, 1984). First, experts would be required to watch videotaped performances of the target rates (e.g., an actual work sample or a performance in assessment center exercises). Next, experts would attempt to reach consensus about what aspects of performance are important for evaluation, what constitutes effective and ineffective performance, and how performance information should be combined when evaluating a ratee on a specific rating scale. If experts can agree on what the organization considers important and what constitutes success or failure in terms of meeting these goals, experts may operate from a common frame of reference when discussing and evaluating performance.

Once consensus is reached on performance issues, experts would again view the videotaped performances and individually rate each ratee on each performance dimension. Discrepancies among expert ratings would then be discussed by all the experts to reach consensus. This discussion should require the experts to justify their ratings and could provide a useful cross check on whether the experts share a common frame of reference. Overall, such a procedure may be effective for calibrating expert raters; it remains, however, an empirical question that must be addressed in future research (e.g., Can an expert with an established frame of reference modify his or her performance standard?).

Perhaps the biggest obstacle to such a procedure (beyond practical considerations) is our current lack of a theory of performance. Although Smith’s (1976) seminal article provides a good starting point by emphasizing the importance of organizational goals when attempting to describe effective performance, we need a comprehensive theory that can be used to guide the task of rating performance. This is an important point because it forces us to consider the calibration issue more critically. Each expert may have a different yet equally valid definition of effective or ineffective performance, thus making the task of calibration difficult or even impossible. Without a common theory of performance shared by the experts to guide the calibration process, expert ratings will likely amount to summaries or averages of divergent opinions. The challenge, then, is to arrive at shared definitions of performance that reflect organizational needs and goals. If the organization does not specify what it means by performance and what behaviors are or are not expected to achieve varying levels of growth and prosperity, calibration becomes less meaningful and accuracy becomes a moot issue.

In summary, if sufficient levels of reliability and validity are obtained (according to the researchers and research consumers), and if experts are properly calibrated, we may have faith in the expert ratings as a suitable standard of comparison. If accuracy is to have any meaning in an appraisal context, the rating standards should be carefully developed and examined with respect to the aforementioned criteria for suitability.

Is Accuracy Important?

Historically, it appears that the criteria used to evaluate ratings have shifted away from traditional error measures to measures purporting to directly assess rating accuracy. Validity, however, has been relatively neglected as a criterion for evaluation (Bernardin & Beatty, 1984). All of this raises a basic question: What should be our evaluative criteria?

Unfortunately, there is no simple answer to this question. What seems clear is that establishing rating accuracy may not always be necessary. For example, if employees must be promoted from a pool of employees within the organization, validity is sufficient to establish the quality of the ratings. In fact, using certain accuracy measures can be misleading. For instance, although it has been suggested that differential elevation may be useful when ratings are used for the purpose of making promotion decisions (Dickinson, 1986; Murphy et al., 1982), it can be shown that a low score on differential elevation (i.e., high accuracy) does not necessarily mean that the rater did a good
job of ranking ratees. This is because the final estimate confounds rank order information with distance information.

There do exist occasions when information about rater accuracy might be important. For example, if a cutoff score is used to determine promotions, inaccurate ratings imply more than a lack of calibration with expert ratings—it implies that promotion decisions will not be maximally fair and that false positives and false negatives may result. If accuracy is considered important, the only way to properly assess accuracy is by obtaining a standard of comparison for computing accuracy scores. Although there may be problems and costs associated with generating accuracy measures in the field (e.g., a limited timeframe for the lengthy process of developing suitable expert ratings), an organization may nevertheless decide that assessing accuracy is useful.

A number of different uses for generating accuracy scores in the field may be identified. For example, accuracy scores may be used in rater training programs to screen out inaccurate raters, provide feedback to inaccurate raters, or evaluate the impact of a rater training intervention. Ratee performance stimuli may include actual samples of employee behavior (e.g., videotaped performance, personnel records, and measures of work output) or simulate samples of ratee performance based on job analyses or typical employee performance profiles. Standards of comparison can be developed by experts (e.g., current supervisors and human resource consultants) or may be established by requiring management to provide objective or subjective weights for specific rater behaviors according to some formal organizational policy (Campbell, 1983; Smith, 1976). Obtaining a standard may be further simplified if accuracy scores are computed simply on the occurrence or nonoccurrence of particular behaviors rather than on the evaluation of particular behaviors (Gordon, 1970, 1972). Feedback and training can potentially improve the level of rating accuracy on the job and help provide raters with a similar frame of reference when appraising ratees in field settings. Finally, prospective candidates for promotion to management might also be evaluated for rating accuracy, and this information could be incorporated into the candidates' overall rating (assuming ability to rate is an important task on the job in question).

Overall, rater accuracy can be evaluated in the field, suggesting that accuracy scores are not solely limited to laboratory studies. Furthermore, numerous standards beyond expert ratings are available, including hard performance criteria that may be useful, provided they are subject to the same scrutiny as are experts' judgmental ratings. When effective decision making requires accurate ratings, it might be useful to examine rater accuracy as described earlier. The ultimate criterion is effectiveness—when inaccurate ratings have no implications for correct or effective appraisal decisions, accuracy is not important.

Concluding Remarks

Although we have focused on the accuracy of performance ratings in this article, note that researchers have recently become interested in the inaccuracies that occur at various stages of cognitive processing and have proposed cognitive models of the rating process (e.g., DeNisi, Cafferty, & Meglinn, 1984; Landy & Farr, 1983). These theories and research offer an entirely different perspective on the study of accuracy by considering the accuracy of cognitive processing rather than the accuracy of performance judgments. Because the linkages between the accuracy of processing and the effectiveness of decisions based on ratings may not be straightforward (see Funder, 1987), understanding how different cognitive errors relate to effective decision making in an appraisal context represents one avenue for future research.

References


Received May 18, 1987
Revision received November 25, 1987
Accepted December 10, 1987

LORNE M. SULSKY AND WILLIAM K. BALZER