Rater training for performance appraisal: A quantitative review

David J. Woehr*

Department of Psychology, Texas A&M University, College Station, TX 77845–4233, USA

Allen I. Huffcutt
Bradley University

A substantial amount of research in the performance appraisal literature has focused on rater training as a means of improving performance ratings. Unfortunately the value of this research is somewhat equivocated by a lack of organization and integration. The present study provides an integration and a quantitative review of the rater training literature. A general framework for the evaluation of rater training is presented in terms of four rating training strategies (rater error training, performance dimension training, frame-of-reference training, and behavioral observation training) and four dependent measures (halo, leniency, rating accuracy and observational accuracy). Finally, a meta-analytic review is presented to assess the effectiveness of the rater training strategies across the four dependent measures.

The measurement of individual work performance has been one of the most prolifically researched areas in the personnel psychology literature. One generally accepted aspect of this literature is that due to difficulties or limitations inherent in applied settings, the vast majority of performance measurement relies on subjective judgemental measures of performance (Guion, 1965; Lacho, Stearns & Villere, 1979; Landy & Rastegary, 1988; Smith, 1986). Another generally accepted conclusion of the performance appraisal literature is that subjective judgements of performance tend to introduce a great deal of distortion into the measurement process. Two strategies have traditionally been advocated to address the problems with subjective performance judgements: rating scale development and rater training. The results of rating scale comparison studies indicate that format modification alone does not result in much improvement in performance ratings (Gomez-Mejia, 1988; Kingston & Bass, 1981; Landy & Farr, 1980). Rater training, however, is an area that has received considerable attention and has shown potential for improving the effectiveness of performance ratings.

The potential value of rater training was recognized as early as 1948 (Bitner, 1948). Subsequently, numerous studies have been conducted evaluating the effectiveness of such training. Previous reviews of this literature have reported equivocal results with respect to rater training (Smith, 1986; Spool, 1978). Spool (1978) concluded that rater training

*Requests for reprints.
is generally effective. However, no conclusions are drawn with respect to the degree of success for different training programs. Smith (1986) presents a more elaborate framework for the evaluation of rater training. Results of the Smith (1986) review suggest that both the method of presentation of the training material as well as the content of training can influence the effectiveness of rater training. Again, however, conclusions with respect to the relative effectiveness of different training strategies on different dependent measures are relatively limited. Thus, the purpose of the present study is to build on previous reviews in presenting a general framework for the evaluation of rater training strategies and to extend previous reviews by providing a preliminary quantitative review of the effectiveness of rater training across rater training approaches and dependent measures.

**Categorization of training approaches**

*Rater error training*

The earliest focus of rater training programmes was on the psychometric properties of subjective performance ratings. The pervasiveness of negatively skewed, range-restricted and moderately to highly intercorrelated performance ratings is well documented (Cooper, 1981; Landy & Farr, 1980; Saal, Downey & Lahey, 1980) and have been interpreted to indicate the presence of leniency, central tendency and halo rating errors respectively (Saal et al., 1980). Performance ratings have also been noted to be relatively unreliable (Pearlman, Schmidt & Hunter, 1980). These psychometric errors have been interpreted as evidence that ratings contain substantial amounts of both systematic and non-systematic error. As a result, many researchers have advocated the elimination of these rating errors through the training of raters to recognize and avoid these problems. This type of training may be broadly categorized as rater error training. The major premise of rater error training is that familiarizing raters with common rating errors (e.g. leniency, halo, central tendency and contrast errors) and encouraging raters to avoid these errors will result in the direct reduction of rating errors and hence more effective performance ratings.

While a few studies have examined the impact of rating error training on rating characteristics such as inter-rater reliability, these studies have generally not found any positive effects for training (e.g. Bernardin & Pence, 1980; Borman, 1975). It is important to note, however, that the primary focus of rater error training is on the reduction of rating errors such as halo and leniency. Consequently, the primary dependent measures in the rater error training literature are some operationalization of the psychometric error(s) that the training programme is designed to reduce. By far, the most widely considered of such errors are halo and leniency. Here it was traditionally assumed that ratings that did not show characteristics such as high dimensional intercorrelations and range-restriction would contain less error and thus be more accurate. This assumption, however, has received somewhat less than unequivocal support. Bernardin & Pence (1980) presented raters with a training programme that involved definitions, graphic illustrations and numerical (i.e. distributional) examples of leniency and halo error. The major focus of this training was to change rater response distributions and presumably improve performance ratings. Paradoxically, results of this training indicated that while ratings did indeed show significantly less rating error, rating accuracy also decreased. Based in large part on the findings of Bernardin & Pence (1980) many researchers concluded that rater error
training was an inappropriate approach to rater training and in fact results in decreased rating accuracy.

Latham (1986), however, strongly disagreed with the conclusion that rater error training is an inappropriate approach to rater training. Rather he suggested that the finding of decreased rating accuracy was the result of 'inappropriate response training' as opposed to 'rater error training'. More specifically, rater error training procedures that train raters to recognize and avoid rating errors by identifying 'incorrect' rating distributions (e.g. highly intercorrelated ratings across dimensions) and specifying 'correct' rating distributions (e.g. low or no correlations among ratings across dimensions) replace one inappropriate response set with another and thus detrimentally affect rating accuracy. However, training raters to recognize and avoid rating errors without specifying 'incorrect' rating distributions can be an effective training approach. Similarly, Bernardin & Buckley (1981) recognized the inappropriateness of the Bernardin & Pence (1980) training content. They discussed the fact that high intercorrelations among dimensions are not necessarily an indicator of halo 'error'. Rather, subjects in the Bernardin & Pence (1980) study simply did what they were told to do and hence were inaccurate.

This leads to several questions. First, to what extent does rater error training actually reduce psychometric errors such as halo and leniency? Second, what effect does rater error training have on rating accuracy? In addition, are the effects of rater error training on rating accuracy moderated by the inclusion of inappropriate response training? That is, do rater error training interventions that do more than simply provide inappropriate response training yield different results with respect to rating accuracy?

Performance dimension training

A second type of training deals with the dimensions of performance that will be used in the ratings. In part due to the equivocal results of rater error training with respect to rating accuracy, researchers began to focus on the cognitive processing of information by raters as a key to rater training (Feldman, 1981; Landy & Farr, 1983). The major premise being that an understanding of the way in which raters process information with respect to evaluation will lead to training strategies that improve the effectiveness of performance ratings. One aspect of this research suggests that in many instances people form 'on-line' evaluations of others. That is, people form judgements as behaviour is observed rather than at a later time when an evaluation or rating is required (DeNisi, Robbins & Williams, 1989; Hastie & Park, 1986; Keller, 1987; Lichtenstein & Srull, 1987; Murphy, Philbin & Adams, 1989; Woehr & Feldman, 1993). An implication of these findings is that training raters to recognize and use the appropriate dimensions on which ratings will be required should lead to dimension relevant judgements as opposed to a more global judgement. This in turn should lead to more accurate ratings (Woehr, 1992). This type of training has been categorized as performance dimension training (Smith, 1986). Performance dimension training is based on the premise that the effectiveness of ratings can be improved by familiarizing raters with the dimensions on which performance is subsequently rated prior to the observation of performance. This is typically accomplished by reviewing the rating scale used in evaluations or having raters participate in the actual development of the rating scale.

The primary focus of performance dimension training has been on performance rating
accuracy as opposed to rating error. Performance rating accuracy may be operationalized in a number of ways (see Sulsky & Balzer, 1988). In general, however, the accuracy of performance ratings is computed through some comparison of an individual rater’s ratings across performance dimensions and/or rates with corresponding evaluations provided by expert raters (i.e. ‘true scores’). With these measures then, the closer the raters’ ratings are to the ‘true scores’, the more accurate they are believed to be (Borman, 1977; Cronbach, 1955; Sulsky & Balzer, 1988).

Frame-of-reference training

A more elaborate rater training strategy emerging from the social cognitive approach to performance appraisal focuses on training raters with respect to performance standards as well as performance dimensionality. This strategy, labelled frame-of-reference training (Bernardin & Buckley, 1981), typically involves emphasizing the multidimensionality of performance, defining performance dimensions, providing a sample of behavioural incidents representing each dimension (along with the level of performance represented by each incident) and practice and feedback using these standards to evaluate performance. The primary extension of frame-of-reference training over performance dimension training is a focus on providing raters with appropriate standards pertaining to the dimensions to be rated. More specifically, while both frame-of-reference training and performance dimension training emphasize the multidimensionality of performance and identifying those dimensions, frame-of-reference further attempts to train raters with respect to common evaluative standards. Thus the goal of frame-of-reference training is to train raters to share and use common conceptualizations of performance when making evaluations. Here it is postulated that to the extent that raters evaluate performance in line with dimensions and standards of performance provided by job experts, ratings will be more effective.

As with performance dimension training, the primary focus of frame-of-reference training has been on performance rating accuracy as opposed to rating ‘errors’. However, it should be noted that rating error measures as well as rating accuracy measures are typically collected in evaluations of both performance dimension and frame-of-reference training.

Behavioural observation training

A final approach to rater training focuses on raters’ observation of behaviour as opposed to raters’ evaluations of behaviour. Thornton & Zorich (1980) postulate that a distinction between the processes of observation and that of judgement is important for performance ratings. They suggest that judgement processes include the categorization, integration and evaluation of information, while observation processes include the detection, perception and recall or recognition of specific behavioural events. They further argue that rating errors such as leniency and halo are primarily due to a lack of information stemming from problems in observational processes. Consequently, rater training that focuses on strategies to improve behavioural observation should increase the effectiveness of performance ratings. Such training may be categorized as behavioural observation training and includes any methodology that focuses on the observation or recording of behavioural
Evaluation of behavioural observation training interventions typically involves some type of memory measure. The assumption is that better observation of behavioural information will result in higher levels of recall or recognition and thus better ratings. Here observational accuracy can be distinguished from rating or evaluative accuracy (Lord, 1985). A key distinction is that observational accuracy measures are based on some objective quantifiable characteristic of the stimulus material. The most commonly used measure of observational accuracy, for example, is some type of frequency of behaviour scale. More specifically, raters are asked to provide an estimate of the number of times a particular behaviour or event occurred. Observational accuracy is then based on the relationship of this estimate to the number of times the behaviour actually occurred in the stimulus material. Other approaches include the use of behaviour recognition measures (e.g. Sulsly & Day, 1992) in which subjects are asked to indicate which behaviours out of a list of behaviours actually occurred and which did not. Observational accuracy is then assessed through measures based on traditional signal detection theory indices (Lord, 1985). Finally, Thornton & Zorich (1980) developed a questionnaire that included true/false, multiple choice and matching items to assess a sample of behaviours occurring in the stimulus material, with the number of correct answers on the questionnaire as the dependent measure. Again, the common factor across these measures is that subject responses are based on memory for specific events and can be compared to actual events in the stimulus material.

Present paper

Based on the above discussion a conceptual meta-model of rater training can be developed. This model includes four general categories of rater training: (1) rater error training, (2) performance dimension training, (3) frame-of-reference training, and (4) behavioural observation training. Additionally, while each of the rater training approaches tend to focus on a particular dependent measure (i.e. rating errors vs. rating accuracy), the training literature is such that multiple dependent measures are typically collected for each training intervention. Thus the model includes four general measurement constructs most typically used for the evaluation of rater training interventions: (1) halo error, (2) leniency error, (3) rating accuracy and (4) observational accuracy.

As noted, previous reviews of the rater training literature (Smith, 1986; Spool, 1978) have attempted to integrate and summarize findings with respect to the effectiveness of rater training. However, a major shortcoming of these articles is the use of traditional narrative reviews. Because narrative reviews typically do not document how comparisons of studies were made, it is difficult to assess the relative value of different training studies. Alternatively, a number of quantitative procedures, collectively referred to as meta-analysis, have been developed (Glass, McGaw & Smith, 1981; Hunter & Schmidt, 1990). An important advantage of meta-analysis is that it provides explicit information on the decision processes used by the reviewer as well as a quantitative summary of the literature. Thus the primary goal of the present study is to use meta-analysis to provide a quantitative review of the effectiveness of the four rater training approaches with respect to the four dependent measures presented in the conceptual model. A second goal is to provide
an indication of the amount of research focusing on the different approaches as well as the extent to which different combinations of rater training approaches have been studied. Finally, meta-analysis will also be used to address the potential moderating effect of the rater error training strategies as delineated by Latham (1986) on rating accuracy. More specifically, are the effects of rater error training on rating accuracy dependent on whether or not the training procedure is categorized as 'inappropriate response training'?

Method

Literature search

A search was conducted to locate all possible studies which empirically tested the effectiveness of rater training. A literature review was conducted using a number of computerized databases. These databases included PsycLit, Dissertation Abstracts International, PsycFirst, ERIC, ArticleFirst and the Social Science Citation Index. In addition, reference lists from obtained studies were also analyzed to identify additional studies, both published and unpublished. Finally, a number of researchers in the area were contacted in order to try and obtain more unpublished studies.

This search resulted in the location of 37 empirical studies dating from 1949 to 1992. Three decision criteria were used for the inclusion of studies in the final analysis. The first two criteria stem from requirements for the calculation of effect sizes. First, only those studies that provided an empirical comparison of a rater training intervention with a control or comparison group were included. Second, only those studies reporting the necessary statistics for meta-analysis were included (i.e. mean, SD and N for the experimental and control groups, or information necessary for the conversion of test statistics into ds). Finally, only those studies that evaluated at least one of the four general rater training strategies (or some combination of two of the four) and used at least one of the four measurement constructs described above were included. Of the 37 empirical studies, four were excluded from the analysis due to insufficient data for the calculation of effect sizes, three were excluded due to the lack of a control or comparison group, and one was excluded for not using one of the four specified dependent measures. Thus 29 studies were retained for the final meta-analysis. A listing of these studies and an indication of the rater training strategy and dependent measure addressed by each is presented in the Appendix. Of the 29 studies retained, 26 are published articles, two are unpublished papers and one is an unpublished doctoral dissertation. Finally, 19 of the 29 studies included in the analysis compared more than one training approach on multiple dependent measures. Thus the 29 empirical studies resulted in 71 data points (effect sizes) to be analyzed. Although the 71 effect sizes were drawn from only 29 studies, they are independent across training types. That is, each effect size across training type is based on a unique experimental group.

Coding procedure

Each study was coded by one of the two authors according to both the type of rater training approach(es) used and the dependent measure(s) used in the evaluation of the training. In addition, eight (approximately 28 per cent) of the 29 studies were randomly selected and independently coded by both authors with respect to rater training strategy and dependent measure. Inter-rater reliability of the coding was then assessed in terms of the percentage agreement between the two coders. The level of agreement with respect to the rater training strategy and dependent measure coding was approximately 98 per cent. Disagreement regarding the other 2 per cent was minor and was resolved by consensus agreement between the coders.

Rater training approach. Information was gathered from each study regarding the type of rater training conducted. Studies were coded as reporting interventions pertaining to at least one of the four general rater training strategies: (a) rater error training (indicated by a focus on the reduction of halo and/or leniency errors), (b) performance dimension training (indicated by a focus on the dimensionality of performance), (c) frame-of-reference training (indicated by a focus on a common set of performance standards corresponding to the
dimensional system used when making performance judgements) and (d) behavioural observation training (indicated by a focus on improving observational skills). In addition, nine of the 29 studies reported interventions that combined elements of two different training approaches. Pulakos (1984), for example, evaluated a training programme that combined elements of rater error training as well as frame-of-reference training. Such studies were coded as representing a combination of the two appropriate training strategies. Data points corresponding to four such combinations were identified and included in the analyses. These combinations were: (a) rater error training and performance dimension training, (b) rater error training and frame-of-reference training, (c) rater error training and behavioural observation training, and (d) behavioural observation training and performance dimension training.

**Dependent measures.** Each study was also coded according to the type of dependent variable(s) assessed. The studies reported data pertaining to at least one of four general categories of dependent measures. These categories were: (a) halo error, (b) leniency error, (c) rating accuracy, and (d) observational accuracy. In addition, as noted above, the majority of studies reported measures of multiple types of dependent measures (e.g. rating accuracy and halo or leniency error) and thus contributed multiple effect sizes to the final analysis.

A potential problem was noted with both the halo error and rating accuracy dependent variables. More specifically, different operationalizations of these constructs are commonly used in the literature. Aggregation across these operationalizations assumes that they are in essence equivalent measures of the same construct. This assumption has been the focus of some concern for measures of both halo and rating accuracy. Fox, Bizman & Hoffman (1989) argue that halo is conceptually not a unitary effect. Similarly, Pulakos, Schmitt & Ostroff (1986) argue that the most common index of halo error, the average standard deviation across rating dimensions, may not be the most conceptually consistent operationalization. However, they found relatively high correlations among the various measures of halo error \( r \approx .80 \). Thus in the present study, studies were coded for the halo error dependent measure based on the intended measurement construct as opposed to specific operationalizations of the construct.

Different operationalizations of rating accuracy are also commonly used in the performance appraisal literature. However, previous studies have reported low correlations at best among these measures indicating that these measures are likely to be non-equivalent measures of rating accuracy (see Sulsky & Balzer, 1988). Sulsky & Balzer (1988) distinguish between two broad types of accuracy measures: distance measures and correlational measures. Both measures are based on the direct comparison of subject ratings with expert-based 'true scores'. Distance accuracy measures provide an index of the average absolute deviation of subject ratings from the true scores. Correlational accuracy measures, however, only provide an index of the similarity of rating patterns and are insensitive to distances between ratings and true scores. Thus in the present analysis only studies that reported distance measures of accuracy were included in the analysis.

**Calculation of effect sizes**

Cumulating effects across studies in a meta-analysis requires that results of all studies be converted to a common metric (Hunter & Schmidt, 1990). For this investigation, the effect size statistic was chosen as the metric of analysis. The effect size statistic, represented by \( d \), is a measure of the strength of a treatment in an experimental design (in this case effects of rater training). In short \( d \) represents the difference between the experimental groups and the control group expressed in standard deviation units. A positive \( d \) value indicates that the experimental group scored higher on the dependent measure than the control group. Similarly, a negative \( d \) value indicates that the control group scored higher than the experimental group and a 0 \( d \) value indicates no difference between the two groups. \( d \) is defined as the difference between the means of the experimental and control groups divided by some measure of the variation (Cohen, 1977; Glass, 1976; Glass, et al. 1981; Hunter & Schmidt, 1990). In accordance with the methods outlined by Hunter & Schmidt (1990, p. 271), the pooled within-group standard deviation was used as the measure of variation for this investigation.

Many of the studies reported the actual means and standard deviations of the experimental (trained) and control (untrained) groups, allowing effect sizes to be calculated directly. In studies where means and deviations were not reported, it was still possible in many cases to calculate effect sizes from \( r \) statistics or univariate two-group \( F \) statistics using standard conversion formulas (Hunter & Schmidt, 1990).
Meta-analysis

Crossing the four rater training approaches with the four dependent measures resulted in a grid containing 16 unique combinations. Each cell in this grid represented the effects of one type of rater training (e.g. rater error training) on one type of dependent measure (e.g. rating accuracy). In addition, a similar categorization was conducted for the four rater training combinations and the four dependent measures.

For each rater training type and dependent measure where data were available (i.e. each cell), a sample weighted mean effect size and corresponding variance was calculated. Essentially, effect sizes within a given cell were multiplied by their respective sample size, summed, and then divided by the total sample size for that cell. Multiplying by sample size is done to give studies with larger sample sizes more weight, since sampling error generally decreases as the sample size increases (see Hunter & Schmidt, 1990).

Results

The results of the meta-analysis for each of the four rater training approaches by each of the four dependent measures are reported in Table 1. The first and second columns of the table are the number of data points and overall number of subjects respectively across all studies for each training type and dependent measure. The third and fourth columns are the sample weighted mean effect size and standard deviation of the effect sizes respectively for each training type and dependent measure. Finally, the last two columns present the minimum and maximum effect size for each training type and dependent measure. As a rough guideline for the interpretation of effect sizes, Cohen (1977) suggests an effect size of .2 represents a small effect, an effect size of .5 represents a medium effect, and an effect size of .8 represents a large effect.

Several interesting results emerge with respect to the individual training approaches (Table 1). As would be expected, rater error training is the most frequently evaluated training strategy with nearly double the number of data points (28) as frame-of-reference training (15), the next most frequently evaluated approach. Rater error training alone appears to be moderately effective at reducing halo error (d = .33) and somewhat less effective with respect to leniency (d = .21). In addition, contrary to current accepted belief, rater error training does not result in decreased rating accuracy but rather in a modest increase (d = .26).

Although rater error training did not result in decreased rating accuracy, the four rater error training programmes providing a measure of rating accuracy were further categorized as one of two types. The error training in two of these studies (Bernardin & Pence 1980; McIntyre, Smith & Hassett, 1984) is based on the training paradigm in which numerical (i.e. distributional) examples of rating errors are used to train raters to recognize particular distributions of ratings as an indication of rating error and to ensure that their ratings are not distributed in a similar fashion (e.g. ratings with low dimensional intercorrelations are always desirable). In essence, this training replaces one response set with another (Bernardin & Pence, 1980; Latham, 1986). The error training in the remaining two studies (Fay & Latham, 1982; Heneman, 1988), however, is based on the paradigm originally presented by Latham, Wexley & Pursell (1975) in which common rating errors are demonstrated and explained to raters and ways to avoid these errors are discussed. More specifically, no reference is made to ‘correct’ or ‘incorrect’ rating distributions.

Mean effect sizes were calculated for each of the two groups of studies. This analysis indicates that the rater error training focusing on rating distributions actually leads to a
Table 1. Results of meta-analysis for performance appraisal rater training

<table>
<thead>
<tr>
<th>Training type</th>
<th>No. of data points</th>
<th>N</th>
<th>Mean $d$</th>
<th>Var $d$</th>
<th>Min. $d$</th>
<th>Max. $d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RET</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dependent measure:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Halo</td>
<td>13</td>
<td>1330</td>
<td>.33</td>
<td>.17</td>
<td>0</td>
<td>1.27</td>
</tr>
<tr>
<td>Leniency</td>
<td>9</td>
<td>2255</td>
<td>.21</td>
<td>.48</td>
<td>-.25</td>
<td>2.18</td>
</tr>
<tr>
<td>Rating accuracy</td>
<td>4</td>
<td>249</td>
<td>.26</td>
<td>.27</td>
<td>-.54</td>
<td>.82</td>
</tr>
<tr>
<td>Observational accuracy</td>
<td>2</td>
<td>51</td>
<td>-.17</td>
<td>.05</td>
<td>-.31</td>
<td>.01</td>
</tr>
<tr>
<td>PDT</td>
<td>8</td>
<td>776</td>
<td>.18</td>
<td>.12</td>
<td>-.19</td>
<td>.93</td>
</tr>
<tr>
<td>Dependent measure:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Halo</td>
<td>3</td>
<td>328</td>
<td>.30</td>
<td>.16</td>
<td>-.01</td>
<td>.93</td>
</tr>
<tr>
<td>Leniency</td>
<td>2</td>
<td>100</td>
<td>-.14</td>
<td>.008</td>
<td>-.19</td>
<td>-.02</td>
</tr>
<tr>
<td>Rating accuracy</td>
<td>3</td>
<td>348</td>
<td>.13</td>
<td>.12</td>
<td>-.03</td>
<td>.70</td>
</tr>
<tr>
<td>Observational accuracy</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>FOR</td>
<td>15</td>
<td>914</td>
<td>.45</td>
<td>.21</td>
<td>-.43</td>
<td>1.33</td>
</tr>
<tr>
<td>Dependent measure:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Halo</td>
<td>4</td>
<td>234</td>
<td>.13</td>
<td>.18</td>
<td>-.43</td>
<td>.72</td>
</tr>
<tr>
<td>Leniency</td>
<td>3</td>
<td>174</td>
<td>.15</td>
<td>.13</td>
<td>-.17</td>
<td>.66</td>
</tr>
<tr>
<td>Rating accuracy</td>
<td>6</td>
<td>365</td>
<td>.83</td>
<td>.13</td>
<td>.17</td>
<td>1.33</td>
</tr>
<tr>
<td>Observational accuracy</td>
<td>2</td>
<td>141</td>
<td>.37</td>
<td>.08</td>
<td>.01</td>
<td>.55</td>
</tr>
<tr>
<td>BOT</td>
<td>4</td>
<td>224</td>
<td>.59</td>
<td>.18</td>
<td>0</td>
<td>1.09</td>
</tr>
<tr>
<td>Dependent measure:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Halo</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Leniency</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Rating accuracy</td>
<td>2</td>
<td>78</td>
<td>.77</td>
<td>.18</td>
<td>.26</td>
<td>1.09</td>
</tr>
<tr>
<td>Observational accuracy</td>
<td>2</td>
<td>146</td>
<td>.49</td>
<td>.12</td>
<td>0</td>
<td>.63</td>
</tr>
</tbody>
</table>

Key: RET = Rater error training, PDT = Performance dimension training, FOR = Frame-of-reference training, BOT = Behavioural observation training.

small decrease in rating accuracy ($d = -.20$, total $N = 130$). Alternately, the rater error training focusing on understanding and avoiding common rating errors without emphasizing rating distributions resulted in a large increase in rating accuracy ($d = .76$, total $N = 119$).

Further support for the moderating effect of the type of error training is provided by four additional studies (i.e. Borman, 1975, 1979; Hedge & Kavanagh, 1988; Pulakos, 1984). These studies were included in the primary analysis for a dependent measure other than rating accuracy (i.e. halo, leniency and/or behavioural accuracy). They were not, however, included with respect to rating accuracy because they only report correlation-based measures (as opposed to distance measures) of rating accuracy. However, two of the studies (i.e. Borman, 1975; Hedge & Kavanagh, 1988) evaluate rating distribution based training programmes and the other two evaluate error avoidance based programmes. Thus
the mean effect sizes, based on the correlational accuracy measures, for each of the two sets of studies were computed and compared. This comparison indicates almost no effect for the distribution based programmes (d = .08, total N = 207) and a small positive effect for the error avoidance programmes (d = .21, total N = 174).

Performance dimension training appears to be moderately effective at reducing halo error (d = .30) and less effective with respect to increasing rating accuracy (d = .13). Performance dimension training also resulted in a small overall increase in leniency (d = −.14).

Frame-of-reference training appears to be the most effective single training strategy with respect to increasing rating accuracy. Results indicate a large mean effect size for interventions on rating accuracy (d = .83). In addition, small to moderate effects with respect to increases in observational accuracy (d = .37) and decreases in both halo (d = .13) and leniency (d = .15) are indicated.

Behavioural observation training is the least frequently evaluated of the four training strategies with a total of only four data points. Based on these studies, however, behavioural observation training appears to have a medium to large positive effect on both rating and observational accuracy (d = .77 and .49 respectively). No data were available with respect to the effect of behavioural observation training on either halo or leniency rating errors.

Analogous results for the combinations of rater training strategies are reported in Table 2. Relatively few studies have examined the effects of combinations of the different rater training approaches. As a result the number of data points available is extremely small and the results should be interpreted with caution.

The combination of rater error training and performance dimension training involved error training along with some form of performance dimension definition and identification. With respect to this combination, effect sizes for both halo and leniency measures (d = .38 and .27 respectively) were similar to those for rater error training alone. No data were available for either rating accuracy or observational accuracy for rater error training along with performance dimension training. Results pertaining to a combination of rater error training and frame-of-reference training indicate that this combination resulted in moderate positive effect sizes for decreases in both halo and leniency error (d = .43 and .67 respectively) as well as increases in rating accuracy (d = .52). Three studies reported results for a combination of rater error training and behavioural observation training. These results indicate a small positive effect for the reduction of halo error (d = .10), a small negative effect for leniency (d = −.08), and a large positive effect for observational accuracy (d = 1.27). Finally, a combination of behavioural observation training and performance dimension training resulted in a large negative effect size for halo error (d = −1.03) and a moderate positive effect for leniency (d = .35). In addition, large positive effects for both rating and observational accuracy (d = 1.14 and 1.10 respectively) were also indicated.

Discussion

Several conclusions are suggested by the results of the present study. First, each of the four rater training strategies appears to be at least moderately effective in addressing the aspect of performance ratings that it was designed to address. That is, rater error training reduces
Table 2. Results of meta-analysis for performance appraisal rater training

<table>
<thead>
<tr>
<th>Training type</th>
<th>No. of data points</th>
<th>N</th>
<th>Mean $d$</th>
<th>Var $d$</th>
<th>Min. $d$</th>
<th>Max. $d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RET &amp; PDT</td>
<td>4</td>
<td>244</td>
<td>.32</td>
<td>.03</td>
<td>.15</td>
<td>.48</td>
</tr>
<tr>
<td>Dependent measure:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Halo</td>
<td>2</td>
<td>122</td>
<td>.38</td>
<td>.008</td>
<td>.26</td>
<td>.44</td>
</tr>
<tr>
<td>Leniency</td>
<td>2</td>
<td>122</td>
<td>.27</td>
<td>.03</td>
<td>.15</td>
<td>.48</td>
</tr>
<tr>
<td>Rating accuracy</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Observational accuracy</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RET &amp; FOR</td>
<td>4</td>
<td>326</td>
<td>.50</td>
<td>.11</td>
<td>0</td>
<td>1.08</td>
</tr>
<tr>
<td>Dependent measure:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Halo</td>
<td>2</td>
<td>136</td>
<td>.43</td>
<td>.30</td>
<td>0</td>
<td>1.08</td>
</tr>
<tr>
<td>Leniency</td>
<td>1</td>
<td>54</td>
<td>.67</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Rating accuracy</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Observational accuracy</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RET &amp; BOT</td>
<td>3</td>
<td>469</td>
<td>.31</td>
<td>.37</td>
<td>-.08</td>
<td>1.27</td>
</tr>
<tr>
<td>Dependent measure:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Halo</td>
<td>1</td>
<td>178</td>
<td>.10</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Leniency</td>
<td>1</td>
<td>178</td>
<td>-.08</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Rating accuracy</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Observational accuracy</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BOT &amp; PDT</td>
<td>5</td>
<td>194</td>
<td>.75</td>
<td>.76</td>
<td>-1.03</td>
<td>1.81</td>
</tr>
<tr>
<td>Dependent measure:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Halo</td>
<td>1</td>
<td>25</td>
<td>-1.03</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Leniency</td>
<td>1</td>
<td>25</td>
<td>.35</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Rating accuracy</td>
<td></td>
<td></td>
<td>1.14</td>
<td>.44</td>
<td>.47</td>
<td>1.81</td>
</tr>
<tr>
<td>Observational accuracy</td>
<td></td>
<td></td>
<td>1.10</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Key: RET = Rater error training, PDT = Performance dimension training, FOR = Frame-of-reference training, BOT = Behavioural observation training.

rating error, performance dimension training and frame-of-reference training increase rating accuracy, and behavioural observation training increases observational accuracy. In addition, in most cases, each of the four training strategies resulted in positive effects on all of the four dependent measures. Exceptions to this were that rater error training appeared to lead to decreased observational accuracy and rater error training lead to increased leniency.

One somewhat surprising finding, given the current 'doctrine' with respect to rater error training, was that rater error training did not lead to a decrease in rating accuracy across studies. Rather the effects of rater error training on rating accuracy appear to be moderated by the nature of the specific error training approach. A focus on 'correct' and 'incorrect' rating distributions tends to decrease rating accuracy, while a focus on under-
standing and avoiding errors without focusing on specific distributions increases rating accuracy. These findings support Latham's (1986) contention that rater error training can increase rating accuracy as well as decrease rating errors. These findings also reinforce the contention that traditional 'error' measures are not necessarily direct measures of rating quality. More specifically, the extent to which moderately to highly intercorrelated or negatively skewed rating distributions represent 'rating errors' depends on the actual distributions of the performance being evaluated. Thus any rater training approach that seeks simply to eliminate these characteristics from ratings will be problematic. Similarly any research that attempts to evaluate rating quality solely on the basis of these characteristics will be tentative at best.

Not surprising was the finding that frame-of-reference training leads to the largest overall increase in rating accuracy. It is not surprising that raters trained to evaluate performance using the same standards as 'expert raters' will produce ratings more like the 'expert ratings' (the operationalization of rating accuracy). This finding, however, suggests that raters can be trained based on a specific theory of performance and rating accuracy increases when this theory is applied to the evaluation task. A vital aspect of this training approach pertains to how the specific theory of performance ought to be determined. To date most researchers have simply used 'expert ratings' in a laboratory context to define performance standards. Thus an important concern for the application of frame-of-reference training in field settings is that performance standards as reflected in the frame-of-reference training reflect organizational goals and values with respect to performance.

Another important finding of the present study is that the combination of various aspects of the different rater training strategies can increase the effectiveness of rater training. Unfortunately, relatively few studies have examined the impact of combined training strategies. Thus the meta-analytic data presented here are based on an extremely small number of data points and must be interpreted cautiously. However, many of the results suggested in this analysis appear to be theoretically consistent. That is, given the theoretical underpinnings of the different training strategies it is entirely logical that combinations of these strategies should enhance performance ratings. For example, a focus on understanding and avoiding errors such as halo or leniency that involve distinguishing among performance dimensions (i.e. rater error training) in conjunction with an identification of the relevant dimensions of performance prior to performance observation (i.e. performance dimension training) should result in greater reductions in rating errors and more accurate ratings than either approach alone.

Finally, the data indicate a surprising lack of studies focusing on the effectiveness of behavioural observation training and the observational accuracy dependent measure in general. The data that were available suggest that behavioural observation training may be a very effective approach for increasing rating accuracy as well as observational accuracy. However, very few studies have directly examined the effect of behavioural observation training on rating accuracy and even fewer have examined the effect of behavioural observation training on traditional 'rating errors'. In addition, the importance of the effect of any rater training strategy on observational accuracy should not be overlooked. More specifically, given that a primary purpose of much performance appraisal is to provide specific feedback for training and development, observational accuracy is likely to be at least as important a criteria as evaluative rating accuracy.
As with any study, there are a number of limitations in the present study that should be noted. First, the primary goal of this study was to examine the effectiveness of relatively broad categories of rater training with respect to four general dependent measures. It is important to note that the focus of the present study is primarily on 'what' is trained (i.e. the different information presented) as opposed to 'how' (i.e. the process by which) the information is conveyed. This is not to imply that such a process orientation is not important. There are undoubtedly a number of other moderators of rater training effectiveness that were not considered. These may include process characteristics of the training program itself. Smith (1986), for example, indicates that the method of presentation of the training material (e.g. lecture, group discussion, practice and feedback) may influence the effectiveness of the training programme. Other variables such as the type of rating scale used (Borman, 1979; Fay & Latham 1982), the nature of the rating task (Pulakos, 1979; Warmke & Billings, 1979) or ratee characteristics such as race or sex (Norton, Gustafson & Foster, 1977) may also influence training effectiveness. Other characteristics such as laboratory vs. field settings and length of time between training and evaluation were also not considered. All of these pose important questions for rater training and deserve the attention of future research. However, given the objective of this study and the relatively small number of data points, a finer subgrouping than that presented here was impractical.

Other potentially important moderators of the effectiveness of the rater training approaches are the dependent measures themselves. It is likely that the various operationalizations of the dependent measures for a given construct may also moderate the effectiveness of each of the training strategies. Again, given the relatively small number of data points, a finer subgrouping than that presented here was impractical in the present study. However, while a finer subgrouping of dependent measures might have allowed a more detailed examination of rater training effects, it does not invalidate the information obtained at the current level of aggregation.

Finally, a limitation common to all meta-analytic studies is that the results are heavily dependent on the representativeness of the studies included in the analysis. The vast majority of the studies included in the present analysis are published articles. There are undoubtedly other studies, especially unpublished studies, that were not included. However, every attempt was made to locate as many empirical studies pertaining to rater training as possible. Thus we believe that the studies analysed in this review present an accurate overview of the rater training literature.

It may be argued, however, that given the relatively small number of studies at this time, the rater training literature is not conducive to a meaningful quantitative review. We disagree. It should be noted that the current review incorporates more studies than either of the two previously published rater training reviews. In addition, while the outcome of any meta-analysis based on a small number of studies (even if these are all the studies that exist at the time) depends on which studies randomly happen to be available, a quantitative review still offers many advantages in terms of information yield over either narrative reviews or individual primary studies (Hunter & Schmidt, 1990). Thus the present review builds on and extends previous reviews in this area. It also highlights the fact that there are fewer primary studies focusing on rater training than might be expected. From this perspective this review should be considered preliminary and more primary studies should be conducted.
Acknowledgements

Special thanks to Winfred Arthur, Jr. and Bob Pritchard for their helpful comments on earlier drafts of this paper.

References


Hastie, R. & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. Psychological Review, 93, 258–268.


Appendix 1

Studies in performance appraisal rater training meta-analysis

Study No.

Received 23 August 1993; revised version received 7 January 1994
David J. Woehr and Allen I. Huffcutt


### Appendix 2

Studies contributing effect sizes to each rater training type by dependent measure category

<table>
<thead>
<tr>
<th>Training type</th>
<th>RET &amp; PDT</th>
<th>FOR</th>
<th>BOT</th>
<th>RET &amp; PDT</th>
<th>FOR</th>
<th>BOT</th>
<th>RET &amp; PDT</th>
<th>FOR</th>
<th>BOT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent measure:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Halo</td>
<td>3,5,6,7,8,10,12,15,16,17,20,27,29</td>
<td>1,27,28</td>
<td>1,3,17</td>
<td>—</td>
<td>4,14</td>
<td>17,20</td>
<td>9</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Leniency</td>
<td>2,3,8,10,12,18,20,24,27</td>
<td>1,27</td>
<td>1,3,20</td>
<td>—</td>
<td>4,14</td>
<td>20</td>
<td>9</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Rating accuracy</td>
<td>3,11,13,17</td>
<td>1,19,28</td>
<td>1,3,17,21,22,23</td>
<td>13,19</td>
<td>—</td>
<td>17</td>
<td>—</td>
<td>19,21</td>
<td></td>
</tr>
<tr>
<td>Observational accuracy</td>
<td>13,26</td>
<td>—</td>
<td>21,23</td>
<td>13,25</td>
<td>—</td>
<td>25</td>
<td>21</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Key.** RET = Rater error training, PDT = Performance dimension training, FOR = Frame-of-reference training, BOT = Behavioural observation training.

**Note.** Numbers refer to study no. in Appendix 1.