Rater training revisited: An updated meta-analytic review of frame-of-reference training

Sylvia G. Roch1*, David J. Woehr2, Vipanchi Mishra1 and Urszula Kieszczynska1

1University at Albany, State University of New York, USA
2University of North Carolina Charlotte, USA

The present study updates Woehr and Huffcutt’s (1994) rater training meta-analysis and demonstrates that frame-of-reference (FOR) training is an effective method of improving rating accuracy. The current meta-analysis includes over four times as many studies as included in the Woehr and Huffcutt meta-analysis and also provides a snapshot of current rater training studies. The present meta-analysis also extends the previous meta-analysis by showing that not all operationalizations of accuracy are equally improved by FOR training; Borman’s differential accuracy appears to be the most improved by FOR training, along with behavioural accuracy, which provides a snapshot into the cognitive processes of the raters. We also investigate the extent to which FOR training protocols differ, the implications of protocol differences, and if the criteria of interest to FOR researchers have changed over time.

The evaluation of individual work performance is a long-standing and ubiquitous concern for both organizational researchers and practitioners alike. To date, there have literally been thousands of published articles focusing on performance appraisal (PA) (a cursory search of SSCI using ‘performance appraisal’ as the search term revealed 2,887 hits). Moreover, an overwhelming proportion of this literature has focused on ways to improve the quality of evaluations provided by individual raters. In general, two approaches have been pursued with respect to improving performance ratings: developing better rating scales and training raters. Of these two approaches, rater training is by far the most successful and widely accepted. In fact, Landy and Farr (1980) called for a moratorium on research investigating rating scales because scale differences accounted for only a small percentage of the variance in ratings. On the other hand, training raters can result in large improvements in rating accuracy (e.g., Woehr & Huffcutt, 1994). Moreover, even though very little empirical research has investigated the benefits of rater training

An earlier version of this study was presented at the 2010 annual meeting of the Society for Industrial and Psychology.

*Correspondence should be addressed to Sylvia G. Roch, Department of Psychology, University at Albany, State University of New York, Albany, NY 12222, USA (e-mail: sroch@albany.edu).

DOI:10.1111/j.2044-8375.2011.02045.x
beyond rating accuracy, rater training has the potential to improve ‘buy-in’ from both the rater and ratee, which may in turn increase rater motivation to provide accurate ratings as well as ratee motivation to use the feedback provided to improve performance.

While a number of rater training approaches (e.g., rater error training, performance dimension training, behavioural observation training) have emerged from the performance appraisal literature, the most widely cited is frame-of-reference (FOR) training. As originally proposed by Bernardin and Buckley (1981), the goal of FOR rater training is to train raters to use a common conceptualization (i.e., frame-of-reference) of performance when observing and evaluating ratees (Athey & McIntyre, 1987; Woehr, 1994). Typical FOR rater training includes emphasizing the multidimensionality of job performance, concretely defining performance dimensions, providing sample behavioural incidents indicative of each dimension and corresponding evaluative standards, along with practice and feedback using these standards to evaluate performance. However, it is not clear to what extent researchers’ FOR training protocols differ and to what extent FOR training is successful. The last meta-analysis investigating rater training was conducted in 1994 (Woehr & Huffcutt) and provides little, if any, information regarding training protocols. Thus, our purpose is not only to update the Woehr and Huffcutt meta-analysis, concentrating on FOR training, but also to provide a snapshot of FOR training protocols and to explore potential moderators of FOR training effectiveness based on protocol differences. Another purpose is to increase our understanding of the cognitive mechanisms that may contribute to FOR training effectiveness. Investigating the extent to which different operationalizations of accuracy are influenced by FOR training and whether behavioural accuracy, in addition to rating accuracy, is influenced by FOR training can help shed light on the cognitive mechanisms contributing to FOR training effectiveness.

Frame-of-reference training

It may be noted that FOR training incorporates many aspects of other approaches to rater training. For example, the emphasis on performance dimensionality is similar to that of performance dimension training and the inclusion of specific behavioural incidents overlaps with behavioural observation training. The primary extension of FOR training over previous rater training paradigms is a focus on providing raters with appropriate standards pertaining to the dimensions to be rated and its emphasis on practice and feedback. Moreover, FOR training emerged out of the literature on rater cognitive processing of performance information. Specifically, FOR training is designed to influence how raters encode, represent, organize, and recall information. Conceptually, FOR training can improve rating accuracy through two processes: (1) by helping raters understand what behaviours constitute specific levels of performance on specific dimensions, and (2) by establishing performance prototypes that allow raters to counteract normal information loss (i.e., forgetting) by categorizing ratee performance based on these performance prototypes presented during the training (Hauenstein & Foti, 1989; Ilgen & Feldman, 1983; Sulsky & Day, 1992, 1994; Woehr, 1994).

The creation of performance prototypes shared by all raters may be the most important contributing factor to FOR training effectiveness. Person perception theories postulate that people in general are ‘cognitive misers’ and will categorize others (e.g., Fiske & Taylor, 1984). In general, objects, including people, are placed into categories, sometimes conceptualized as bins (Murphy & Cleveland, 1995), based on how well the features of the object match the prototype for the category (Feldman, 1981; Rosch,
Once a person is categorized, information is stored in terms of the category and not the person (see Hamilton & Trolier, 1986, for a review). Consequently, recall of information tends to be biased towards the prototypical features of the category (Lord, 1985). Thus, one of the primary benefits of FOR training is the creation of relevant prototypes representing different levels of performance shared by all raters. These shared prototypes should facilitate raters’ common classification of ratees into performance categories. To the extent that FOR training results in performance categories shared by all raters, it should lead to increased rating accuracy in comparison with untrained raters who recall information from their own idiosyncratic categories, most likely based on social information (Fiske & Neuberg, 1990).

Even though FOR training was developed within the context of performance appraisal, the importance of raters sharing a common conceptualization of performance categories, that is, a shared FOR, is not only relevant to performance appraisal but for all human resource functions that rely on raters. Researchers have shown that FOR training is directly applicable to a variety of evaluative contexts including assessment centres (e.g., Goodstone & Lopez, 2001; Lewis, 2002; Schleicher, Day, Mayes, & Riggio, 2002), selection test cut scores (Fehrmann, Woehr, & Arthur, 1991), employment applications (Lundstrom, 2008), competency modelling (Lievens & Sanchez, 2007), job analysis (Aguinis, Mazurkiewicz, & Heggestad, 2009), interviews (e.g., Hauenstein, Facteau, & Schmidt, 1999; Hessler, 1992), and even therapy (Angkaw, Tran, & Haaga, 2006).

Certainly one of the primary reasons for the popularity and expanded use of FOR training is the evidence demonstrating its positive impact on performance rating quality. To date, quite a few empirical studies have examined the impact of FOR training, and this body of evidence continues to grow. Yet, one of the most frequently cited references pertaining to the evaluation of FOR training is Woehr and Huffcutt (1994). Indeed, according to Google Scholar (searched May 17, 2011), Woehr and Huffcutt (1994) has been cited 180 times, nearly as many as the 201 citations associated with the original Bernardin and Buckley (1981) article developing FOR training. Woehr and Huffcutt (1994) present a meta-analytic review of the rater training literature and report a relatively large average effect size ($d$) of .83 representing the effect of FOR training on rating accuracy.

Given the impact of the Woehr and Huffcutt (1994) meta-analysis, it is surprising to note that the findings are based on a relatively small number of studies. Despite an extensive literature search, Woehr and Huffcutt (1994) were only able to include six rating accuracy effect sizes representing 365 participants. Woehr and Huffcutt (1994) directly acknowledge this issue. Specifically, they conclude their review by stating that the result “. . . highlights the fact that there are far 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” (p. 201). Thus, one of our goals is to revisit the findings of Woehr and Huffcutt (1994) with respect to FOR rater training.

**Performance rating criteria**

The identification of appropriate criteria for the evaluation of performance ratings has long been a key concern in the performance appraisal literature. Woehr and Huffcutt (1994) examined the impact of rater training with respect to four dependent measures frequently used in the performance appraisal literature prior to their review. Specifically, they examined the effects of training on two types of psychometric rating ‘errors’
(i.e., halo and leniency), as well as two types of accuracy (i.e., rating accuracy and observational accuracy). Rating ‘errors’, one of the earliest operationalizations of rating quality, involve an examination of the psychometric and/or distributional properties of ratings. The pervasiveness of negatively skewed, range-restricted, and moderately to highly inter-correlated performance ratings is well documented (Cooper, 1981; Landy & Farr, 1980; Saal, Downey, & Lahey, 1980) and in the past had been interpreted to indicate the presence of leniency, central tendency, and halo ‘rating errors’, respectively (Saal et al., 1980). Yet, it is now widely recognized that these psychometric characteristics do not necessarily reflect ‘error’ (Murphy & Balzer, 1989). That is, to the extent that rating distributions reflect actual performance distributions, these psychometric characteristics may not be indicative of rating error but of true performance. Consequently, rating error criteria such as measures of halo and leniency are viewed by key performance appraisal researchers as inappropriate criteria for the evaluation of rating quality (Murphy & Balzer, 1989; Murphy, Jako, & Anhalt, 1993) and have largely been abandoned in performance rating research. It is also not surprising that rater training focusing on reducing these psychometric errors, rater error training, has also fallen out of favour. We could only find five studies investigating rater error training since 1994; four of which combined rater error training with either behaviour observation training or performance dimension training.

Woehr and Huffcutt (1994) also examined the impact of rater training in regard to rating accuracy. In general, the purpose of rater training is to align raters with respect to a common set of evaluative standards – a necessity when comparing ratings across raters for human resource practices, such as promotion, training, downsizing, etc. Thus, rating or evaluative accuracy, which assesses the extent to which raters provide ratings that match an established standard (e.g., ‘expert ratings’ or ‘true scores’) represents an appropriate criterion.

Woehr and Huffcutt (1994) differentiated between rating/evaluative accuracy and observational accuracy. Observational accuracy is typically operationalized as improved recall or recognition of specific behavioural incidents, which has also been referred to in the literature as behavioural accuracy. Rating/evaluative accuracy and observational/behavioural accuracy are the two major classifications of accuracy (Murphy, 1991; Padgett & Ilgen, 1989).

One assumption behind behavioural/observational accuracy is that higher levels of recall or recognition will result in higher quality performance ratings. This assumption presumes a model of rater cognitive processing in which memory serves as the basis for subsequent evaluation. Much of the literature on cognitive information processing, however, suggests that evaluative judgments may be largely independent of specific information in memory (e.g., Hastie & Park, 1986; Lichtenstein & Srull, 1987; Woehr & Feldman, 1993).

Thus, even though correctly observing behaviour and providing accurate ratings may be related, the relationship is not strong (e.g., Sanchez & De La Torre, 1996; Sulsky & Day, 1992). Rating accuracy is often conceptualized as a result of correctly placing a person into a cognitive category representing how well someone has performed (DeNisi & Williams, 1988; Murphy & Cleveland, 1995). For example, if someone performs well, is that person correctly categorized by the rater into the “good” performer category? The person perception literature has shown that on-line impressions, such as those used to place someone in a performance category, are typically formed spontaneously when the information is encoded (Hastie & Park, 1986), or in other words, when behaviour is observed. The performance evaluation literature has also shown that these online
impressions are encoded separately from the memory of observed behaviours, and often raters assign ratings according to the category to which they assigned the ratee and not based on their specific memories of ratee behaviour (e.g., DeNisi & Williams, 1988; Ilgen & Feldman, 1983). As discussed earlier, FOR training focuses on improving how accurately raters can categorize ratees and consequently on improving rating accuracy.

There is some indication, however, that a by-product of FOR training is also improved observational accuracy. Woehr and Huffcut (1994) reported an effect size of $d = .37$ for observational accuracy, but it should be noted that this was based on two data points. Since then the research has been mixed as to whether FOR improves behavioural/observational accuracy. Woehr and Huffcut (1994) found that FOR trained raters used a greater number of performance dimensions and recalled significantly more behaviours than untrained raters. Roch and O’Sullivan (2003) found that FOR trained raters correctly recalled more behaviours not only directly after training but also after a 2-week delay than did untrained raters. On the other hand, Day and Sulsky (1995) found that FOR trained raters did not differ from untrained raters in the number of behavioural incidents recalled. Nonetheless, it may be that by sensitizing raters to the importance of specific behaviours when determining ratings, FOR training may have the unintended effect of also improving behavioural/observational accuracy. One of the major purposes of performance appraisal is to provide feedback (Cleveland, Murphy, & Williams, 1989), and behavioural/observational accuracy is important for feedback purposes. Thus, given the mixed research findings, another purpose of the current research is to investigate whether FOR improves behavioural/observational accuracy.

It is clear that FOR training improves rating accuracy. However, operationalizations of rating accuracy differ and to what extent FOR training influences these differing operationalizations may provide a glimpse into the cognitive processes of raters. Common to all is the requirement of a standard to which ratings are compared or, in other words, a ‘true score’ (Roach & Gupta, 1992; Sulsky & Balzer, 1988). The true scores are often based on expert ratings. These experts have the opportunity to watch/read a performance vignette several times, provide ratings, and discuss any discrepancies. Please see Sulsky and Balzer (1988) for a detailed description of the procedure most often used to obtain expert ratings, along with other means of obtaining true scores.

The operationalizations of rating accuracy most commonly seen in the literature are Cronbach’s (1955) four accuracy indices, Borman’s differential accuracy, and distance accuracy. Rating accuracy is most often assessed according to Cronbach’s (1955) four accuracy indices: elevation, differential elevation, stereotype accuracy, and differential accuracy (Murphy & Cleveland, 1995). The indices represent different strategies for decomposing the discrepancy between performance ratings and a set of ‘true scores’, most often based on expert judgments. Differential elevation is often considered the most important type of accuracy (Murphy & Cleveland, 1995; Murphy, Garcia, Kerkar, Martin, & Balzer, 1982) because it indicates how accurately one can distinguish between ratees, important for administrative decisions. However, differential accuracy is important for feedback because it indicates how well one can distinguish between dimensions for a particular ratee and thus report ratee strengths and weaknesses (Murphy & Cleveland, 1995; Murphy et al., 1982). Lastly, the other two Cronbach’s accuracy types represent whether raters are accurate over both dimensions and ratees (elevation) or distinguish between dimensions for a group of ratees (stereotype accuracy), relevant for training purposes. Roach and Gupta (1992) demonstrated that Cronbach’s accuracy indices are not significantly correlated and each provides unique information.
Distance accuracy is related to Cronbach’s accuracy indices. Distance accuracy is simply the average absolute deviation of participants’ ratings from the true scores (McIntyre, Smith, & Hassett, 1984). Conceptually, this measure is similar to Cronbach’s indices because it is also based on the distance between true scores and participant scores. See Balzer and Sulsky (1988) for a review of the formulas for all of the accuracy measures. This measure provides a global sense of how well raters are matching expert ratings.

Borman’s (1977) differential accuracy is fundamentally different than distance accuracy and Cronbach’s accuracy indices (including Cronbach’s differential accuracy index). Borman’s differential accuracy is based exclusively on correlations. Sulsky and Balzer (1988) argue that one cannot discuss accuracy unless one knows how much a ‘true’ score and a participant score differ (i.e., distance) and that a purely correlational measure does not provide this information. It is possible for ‘true’ ratings to differ from ratings provided by raters but still correlate highly. For example, if a rater consistently assigns ratings two points higher than an expert rater, the ratings of these two raters may be perfectly correlated and thus this rater may receive a high ‘accuracy’ score based on correlation measures, such as Borman’s differential accuracy. However, in this case, even though the ‘accuracy’ score may be high, the rater is not mimicking the expert rater. Thus, Sulsky and Balzer (1988) argue that Borman’s differential accuracy measure is not a true accuracy measure. Nonetheless, it is commonly used. Because it focuses on how well raters discriminate among ratees according to performance dimension, it may also be useful for feedback purposes in helping ratees identify areas of relative strengths and weaknesses.

Whether FOR training effectiveness varies across accuracy operationalizations has important conceptual and practical implications. If Borman’s measure shows greater improvement than the measures that rely on the distance from a true score, it would provide an indication that categorization schema established by FOR training may not be based on specific ratings. For example, during FOR training, raters are shown how specific ratings, such as a 1 or a 4, correspond to specific behaviours. Accuracy measures that incorporate the distance from the true score provide an indication of the extent to which raters have adopted a categorization schema based on these ratings. However, if FOR training is more effective in regard to Borman’s measure than the distance-based ones, it would provide an indication that perhaps raters are not adopting a categorization schema based on specific ratings but one that is more general, such as the bin system suggested by Murphy and Cleveland (1995) in which ratees are placed in poor, average, and good bins. Thus, what type of accuracy operationalization is the most influenced by FOR training can give us some insight into the cognitive categorization schema used by raters. If it is Borman’s correlational based measure, FOR researchers may wish to re-evaluate how FOR training is conducted. It may be possible to conduct FOR training without relying on specific true scores, an approach discussed by Chirico et al., (2004).

**Current research**

Given that it has been over 15 years since Woehr and Huffcutt’s (1994) meta-analysis of rater training, we believe that it is time to take another look at the rater training literature to see what kinds of rater training are of interest to researchers now and what criterion measures these researchers are using. Woehr and Huffcutt found that researchers focused on rater error and accuracy measures. To what extent has this changed and are researchers using other criteria in their rater training research? During
the last 10–15 years, there has been a growing recognition that rater and ratee reactions and perceptions are important (Levy & Williams, 2004). Has this growing recognition influenced the rater training literature?

Another goal of the current research is to investigate whether the operationalization of rating accuracy is a moderator of FOR training effectiveness. Given the availability of only six rating accuracy effect sizes, Woehr and Huffcutt (1994) were unable to investigate the effect of FOR training on the various accuracy operationalizations. Thus, we investigate whether FOR training impacts some operationalizations of accuracy to a greater degree than others. Understanding whether FOR training impacts some accuracy operationalizations to a greater degree than others may provide insight into raters’ cognitive processes. Also, understanding to what extent behavioural/observational accuracy is impacted by FOR training may also provide insight into raters’ cognitive processes. As mentioned earlier, Woehr and Huffcutt (1994) only included two effect sizes regarding observational accuracy and subsequent research has been equivocal.

Lastly, we seek to provide a summary overview of the way in which FOR training is implemented, and, when possible, also conduct moderator analysis investigating whether protocol differences predict FOR training effectiveness. Understanding how basic differences in FOR training protocols, such as amount of training time and number of practice ratees, predict FOR training effectiveness will help to improve FOR training practice. Bernardin, Buckley, Tyler, and Wiese (2000) called for more research investigating FOR training protocol. Even though some primary studies (e.g., Sulsky, Uggerslev, Day, Keown, & Law, 2002; Uggerslev, Sulsky, & Day, 2003) have investigated protocol differences, there is still a need for an overview presentation of how FOR training is implemented today for both research and applied purposes.

In summary, we address a series of eight specific research questions:

(1) What types of rater training paradigms have been of interest to researchers since the Woehr and Huffcutt (1994) meta-analysis?
(2) How much empirical research has examined FOR training since 1994?
(3) To what extent are FOR training protocols similar?
(4) What are the primary criteria used for assessing the impact of FOR training?
(5) What is the estimated magnitude of the effect of FOR training on rating and behavioural accuracy based on the current body of evidence?
(6) Is there a difference between the effect size estimate reported by Woehr and Huffcutt (1994) and that based on current evidence regarding both rating accuracy and behavioural accuracy?
(7) Is the impact of FOR training on rating accuracy moderated by specific operationalizations of accuracy?
(8) Is the impact of FOR training on rating accuracy moderated by training protocol differences?

Method

Sample of studies
To find studies appropriate for inclusion, we first searched databases such as PsycINFO, ERIC, EBSCO, Business Source Premier, JSTOR, and Dissertation Abstracts for information related to rater training and accuracy of performance ratings. We limited the search to English language documents but did not limit the search by year. We used rater training, frame-of-reference training, in combination with accuracy, leniency, performance
appraisal, assessment center, interviews, and performance evaluation as search terms. Second, we examined 27 of the 29 manuscripts used in the Woehr and Huffcutt (1994) meta-analysis. Third, rater training researchers were contacted for missing statistics from rater training studies and any unpublished data. Fourth, we performed a manual search of the SIOP annual conference programmes from 1995 to 2009 and the AOM Proceedings archive from 1996 to 2008. Lastly, we performed manual searches of the reference sections of articles collected. We only focused on empirical studies.

Criteria for inclusion
We identified 90 empirical studies investigating rater training, which we evaluated for inclusion, including the articles used in the Woehr and Huffcutt (1994) meta-analysis. Fifty-seven manuscripts investigated FOR training and are included in our overview of FOR training practices. Thirty-six studies included adequate information allowing the computation of effect sizes according to Hunter and Schmidt’s (2004) formula. Of the studies included in the meta-analysis, 23 were published and 13 unpublished, with 117 effect sizes. It should be noted that in some cases, unpublished dissertations and conference presentations were later published, or the same data set was used in multiple conference presentations. We took steps to ensure that each data set was represented only once in both our list of FOR training manuscripts and in our meta-analysis. Please see Table 1 for a list of FOR training manuscripts. We would like to note that we found only one study published since the Woehr and Huffcutt (1994) meta-analysis investigating rater error training and one study investigating behaviour observation training. We found no new studies investigating performance dimension training. Thus, we only concentrated on FOR training.

Data coding
Based on standardized coding sheets, we coded for type of training and criteria, along with effect sizes. Where possible, we coded for moderators, including whether the data were published, purpose of training, type of participants (student versus employee), and details such as length of training, type of training material, training protocol, and any demographic information available regarding the participants. Each article was independently coded by two of the authors and checked for agreement. Coding discrepancies were resolved through consensus discussion. Overall interrater agreement was over 95%.

Meta-analytic procedures
We conducted the meta-analyses using procedures from Hunter and Schmidt (2004) and the Schmidt and Le (2004) software designed for conducting psychometric meta-analyses. The means and variances were corrected for sampling error. We were unable to correct for attenuation due to measurement error because reliability information is not available for either FOR training or rating accuracy measures.

---

1We could not find two papers included in the Woehr & Huffcutt (1994) meta-analysis, both presented at conferences (Pelly & Dossett, 1991; Roberson & Banks, 1986). We tried unsuccessfully to contact the authors.
Some rater training studies not only manipulated type of rater training but also manipulated another variable in a crossed design. We averaged over the ‘other’ manipulation. Also, some studies, such as Sulsky, Skarlicki, and Keown-Gerrard (2002) and Keown-Gerrard and Sulsky (2001), introduced a revised version of FOR training and compared this revised version to both a traditional FOR training and a control session. We ignored all revised versions of FOR training and only report results based on traditional FOR training. Each of the revised versions of FOR was unique, and thus we did not feel comfortable averaging across unrelated revisions to compare revised FOR training to the original FOR training protocol. Thus, for some studies, we did not use the entire sample size but just the sample that completed traditional FOR training and control training. Also, to be included in the meta-analysis, studies must have reported ratings for both a FOR training condition and a control condition. For articles investigating a time delay (e.g., Roch & O’Sullivan, 2003; Sulsky & Day, 1994), we only used the immediate ratings and not the ratings after a time delay. Unfortunately, there were not enough studies manipulating a time delay to investigate the effect of FOR training over time. Most of the studies reported multiple types of accuracy. For our overall analysis, we combined the types of rating accuracy so that each study contributed only one data point to the overall analysis. However, for the moderator analyses, we used the accuracy operationalizations presented in the study; thus a study could contribute data points to multiple moderator analyses. For example, several studies reported all of Cronbach accuracy indices, and these studies are represented in all of the moderator analyses investigating Cronbach’s accuracy indices.

Results

Tables 1 and 2 present a summary overview of the FOR training research to date. As indicated in Table 1, there appears to be a great deal of similarity across FOR training protocols with respect to the basic components of the training. Specifically, the vast majority (96%) include ‘true scores’ based on expert ratings. The majority of the studies (approximately 93%) include a discussion of the performance dimensions to be evaluated and specific behaviours that correspond to each dimension, along with evaluative standards for each behaviour. Nearly all (98%) included practice ratings with accuracy feedback, based on written example behaviours/vignettes (27%) and/or videotaped vignettes (84%) with some studies reporting that both were used. The practice material appears to vary greatly but the number of practice scenarios does not with a mean of approximately 2 and a SD of approximately 1.6.

Overall, the majority of researchers (71%) examined FOR training in a performance appraisal context but in the last 5 years, contexts have varied greatly (only five out of 14 studies in the last 5 years are in a performance appraisal context). Thus, it appears that FOR training no longer is exclusively a performance appraisal topic. The typical length for FOR training also appears to vary, with an average length of about 100 min but with a standard deviation of approximately 88 min. There is also a high level of variability with respect to the number of rating dimensions, with a mean of approximately five dimensions but with a standard deviation of 3.5. However, many researchers tend to include a relatively small number of performance dimensions. Gaugler and Thornton (1989) showed increased rating accuracy with a smaller number of dimensions within an assessment centre context and recommended that assessment centre designers limit the number of rating dimensions used. It appears that this advice has been heeded in the rater training literature. It should be noted, however, that not all studies provided
### Table 1. Empirical FOR training studies

<table>
<thead>
<tr>
<th>Authors</th>
<th>Time</th>
<th>Participants</th>
<th>No. of Dim</th>
<th>Practice</th>
<th>Description of training material</th>
<th>Discuss</th>
<th>Written practice</th>
<th>Video practice</th>
<th>In meta-analysis</th>
<th>True score</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allen, 1999</td>
<td>70</td>
<td>Students</td>
<td>7</td>
<td>1</td>
<td>Videotapes of teachers Practice job analysis items</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Aguinis et al., 2009</td>
<td></td>
<td>Employee</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>Job analysis</td>
</tr>
<tr>
<td>Angkaw et al., 2006</td>
<td>120</td>
<td>Students</td>
<td>10</td>
<td></td>
<td>Videotapes of community volunteers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Therapy</td>
</tr>
<tr>
<td>Athey and McIntyre, 1987</td>
<td>30</td>
<td>Students</td>
<td>4</td>
<td>1</td>
<td>Murphy et al., 1982 video</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Cardy and Keefe, 1994</td>
<td></td>
<td>Students</td>
<td>5</td>
<td></td>
<td>Written vignettes from Sauser, 1979</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Cellar et al., 1989</td>
<td>60</td>
<td>Students</td>
<td>15</td>
<td>1</td>
<td>Job of billing clerk</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Job evaluation</td>
</tr>
<tr>
<td>Cesare, 1989</td>
<td></td>
<td>Students</td>
<td>3</td>
<td>0, 1, or 3</td>
<td>Johnson (1987) video</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Chirico et al., 2004</td>
<td>120</td>
<td>Students</td>
<td>6</td>
<td>3</td>
<td>McCauley et al. 1990 video</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Day and Sulsky, 1995</td>
<td>90</td>
<td>Students</td>
<td>4</td>
<td>2</td>
<td>Sulsky and Day, 1992 video</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Dierdorff et al., 2010</td>
<td></td>
<td>Educators</td>
<td></td>
<td></td>
<td>Videotaped and live oral interviews</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Language proficiency</td>
</tr>
<tr>
<td>Fehrmann et al., 1991</td>
<td></td>
<td>Students</td>
<td>3</td>
<td></td>
<td>3 written psychology test questions (Angoff method)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cut-off scores</td>
</tr>
<tr>
<td>Gorman and Rentsch, 2009</td>
<td>45</td>
<td>Students</td>
<td>5</td>
<td>1</td>
<td>Gorman and Rentsch videos (assessment centre)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Gorman et al., 2009</td>
<td>45</td>
<td>Students</td>
<td>5</td>
<td>1</td>
<td>Gorman and Rentsch videos (assessment centre)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance appraisal</td>
</tr>
</tbody>
</table>

Continued
Table 1. (Continued)

<table>
<thead>
<tr>
<th>Authors</th>
<th>Time</th>
<th>Participants</th>
<th>No. of Dim</th>
<th>Practice</th>
<th>Description of training material</th>
<th>Discuss</th>
<th>Written practice</th>
<th>Video practice</th>
<th>In meta-analysis</th>
<th>True score</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gorman et al., 2010</td>
<td></td>
<td>Students</td>
<td>4</td>
<td>2</td>
<td>Murphy et al., 1982 video</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Unpublished data</td>
</tr>
<tr>
<td>Hartog, 1991</td>
<td></td>
<td>Students</td>
<td>4</td>
<td>2</td>
<td>Murphy et al., 1982 video</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Hauenstein et al., 1999</td>
<td></td>
<td>Students</td>
<td>6</td>
<td>3</td>
<td>Stamoulis and Hauenstein, 1993 video (interview)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Interview</td>
</tr>
<tr>
<td>Hessler, 1992</td>
<td>10</td>
<td>Employees</td>
<td>4</td>
<td>1</td>
<td>Videotaped interviews</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Hoffman and Sulsky, 1998</td>
<td>40</td>
<td>Students</td>
<td>2</td>
<td>2</td>
<td>Sulsky and Day, 1992 video</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Jackson et al., 2005</td>
<td>360</td>
<td>Assessors</td>
<td>7</td>
<td>2</td>
<td>Written vignettes</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Jen, 1998</td>
<td>240</td>
<td>Supervisors</td>
<td>21</td>
<td>1</td>
<td>Video of administrative employees</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Johnson, 1987</td>
<td>210</td>
<td>Students</td>
<td>3</td>
<td>1</td>
<td>Self-developed video</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Keefe, 1988</td>
<td></td>
<td>Students</td>
<td>5</td>
<td>1</td>
<td>Vignette of prof. classroom behaviour based on Sauser, 1979</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Keown-Gerrard and Sulsky, 2001</td>
<td></td>
<td>Students</td>
<td>3</td>
<td>3</td>
<td>Written vignettes based on McCauley’ et al., 1990</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Lee, 1994</td>
<td></td>
<td>Students</td>
<td>8–12</td>
<td>1</td>
<td>Murphy et al., 1982 video</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Lewis, 2002</td>
<td>100</td>
<td>Students</td>
<td>5</td>
<td>1</td>
<td>Schleicher and Day, 1998 video</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Lievens, 2001</td>
<td></td>
<td>Students</td>
<td>3</td>
<td>1</td>
<td>Videotaped vignette</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Lievens and Sanchez, 2007</td>
<td>480</td>
<td>Consultants</td>
<td>5</td>
<td>2</td>
<td>Assigned behaviours to competencies</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Competency modelling</td>
</tr>
<tr>
<td>Lundstrom, 2008</td>
<td></td>
<td>Students</td>
<td>7</td>
<td>2</td>
<td>Evaluated two written biodata profiles</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Evaluating biodata</td>
</tr>
</tbody>
</table>

Continued
<table>
<thead>
<tr>
<th>Authors</th>
<th>Time</th>
<th>Participants</th>
<th>No. of Dim</th>
<th>Practice</th>
<th>Description of training material</th>
<th>Discuss</th>
<th>Written practice</th>
<th>Video practice</th>
<th>In meta-analysis</th>
<th>True score</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>McIntyre, et al., 1984</td>
<td>30</td>
<td>Students</td>
<td>12</td>
<td>1</td>
<td>Murphy et al., 1982 video</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Melchers et al., 2011</td>
<td>70</td>
<td>Students</td>
<td>3</td>
<td>2</td>
<td>Klehe et al. video</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Interview</td>
</tr>
<tr>
<td>Mills, 1999</td>
<td>75</td>
<td>Students</td>
<td>4</td>
<td>2</td>
<td>Sulsly and Day, 1992 video</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Noonan and Sulsly, 2001</td>
<td>180</td>
<td>Employees</td>
<td>9</td>
<td>2</td>
<td>Infantry corporals video</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Powell and Goffin, 2009</td>
<td>120</td>
<td>Students</td>
<td>6</td>
<td>1</td>
<td>Behavioural interview of a bank manager trainee</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Personality ratings</td>
</tr>
<tr>
<td>Pulakos, 1984a</td>
<td>90</td>
<td>Students</td>
<td>5</td>
<td>2</td>
<td>Video of managers talking to subordinates</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Pulakos, 1984b</td>
<td>90</td>
<td>Students</td>
<td>5</td>
<td>2</td>
<td>Based on Borman, 1977 video</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Pulakos, 1986</td>
<td>90</td>
<td>Students</td>
<td>5</td>
<td>2</td>
<td>Video of managers talking to subordinates</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Roch and O’Sullivan, 2003</td>
<td>45</td>
<td>Students</td>
<td>6</td>
<td>1</td>
<td>McCauley et al., 1990 videos</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Roth, 1987</td>
<td></td>
<td>Students</td>
<td>5</td>
<td>2</td>
<td>Based on Borman, 1977 video</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Schleicher, 1998</td>
<td>90</td>
<td>Students</td>
<td>3</td>
<td>4</td>
<td>Videotaped assessment centre exercises</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Assessment centre/Performance appraisal</td>
</tr>
<tr>
<td>Schleicher and Day, 1998</td>
<td>90</td>
<td>Students</td>
<td>4</td>
<td>2</td>
<td>Video of managers talking with subordinate</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Schleicher et al., 2002</td>
<td>90</td>
<td>Students</td>
<td>3</td>
<td>4</td>
<td>Videotaped assessment centre exercises</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Assessment centre/Performance appraisal</td>
</tr>
<tr>
<td>Schlientz et al., 2009</td>
<td>23</td>
<td>Students</td>
<td>2</td>
<td>6</td>
<td>Video of school children engaged in activity</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Ratings of behaviour</td>
</tr>
</tbody>
</table>
Table 1. (Continued)

<table>
<thead>
<tr>
<th>Authors</th>
<th>Time</th>
<th>Participants</th>
<th>No. of Dim</th>
<th>Practice</th>
<th>Description of training material</th>
<th>Discuss practice</th>
<th>Video practice</th>
<th>In meta-analysis</th>
<th>True score</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Senderak, 1983</td>
<td></td>
<td>Students</td>
<td>7</td>
<td>4</td>
<td>Borman tapes (1983)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Sinclair and Hauenstein, 2001</td>
<td>150</td>
<td>Students</td>
<td>3</td>
<td>3</td>
<td>Video of lectures of teaching assistants</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Stamoulis and Hauenstein, 1993</td>
<td>90</td>
<td>Students</td>
<td>7</td>
<td>3</td>
<td>Video of an interviewer and interviewee</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Stevens, 1995</td>
<td>150</td>
<td>Students</td>
<td>6</td>
<td>3</td>
<td>Stamoulis and Hauenstein, 1993 videos</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Interview</td>
</tr>
<tr>
<td>Sulsky, 1988</td>
<td>75</td>
<td>Students</td>
<td>3</td>
<td>2</td>
<td>Video of managers talking with subordinates-Borman, 1977</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Sulsky et al., 2002</td>
<td>75</td>
<td>Students</td>
<td>3</td>
<td>2</td>
<td>Sulsky et al., 2002 video</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Sulsky and Day, 1992</td>
<td>75</td>
<td>Students</td>
<td>3</td>
<td>2</td>
<td>Video of manager talking with subordinates-Roberson and Banks, 1986</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Sulsky and Day, 1994</td>
<td>75</td>
<td>Students</td>
<td>3</td>
<td>2</td>
<td>Same as Sulsky and Day, 1992</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Sulsky and Kline, 2007</td>
<td>85</td>
<td>Students</td>
<td>3</td>
<td>2</td>
<td>Same as Sulsky and Day, 1992</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Sulsky et al., 2002</td>
<td>60</td>
<td>Managers</td>
<td>3</td>
<td>2</td>
<td>Video of salesperson at work</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Sulsky and Uggerslev, 2010</td>
<td>75</td>
<td>Students</td>
<td>3</td>
<td>2</td>
<td>Sulsky et al., 2002 video</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Svoboda, 1986</td>
<td></td>
<td>Students</td>
<td>3</td>
<td>2</td>
<td>Video of managers, Borman, 1979</td>
<td>x</td>
<td>X</td>
<td>x</td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Uggerslev, 2004</td>
<td>90</td>
<td>Students</td>
<td>5</td>
<td>2</td>
<td>Video lecturing instructors</td>
<td>x</td>
<td>X</td>
<td>x</td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Uggerslev et al., 2003</td>
<td>75</td>
<td>Students</td>
<td>3</td>
<td>2</td>
<td>Sulsky et al., 2002 video</td>
<td>x</td>
<td>X</td>
<td>x</td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
<tr>
<td>Woehr, 1994</td>
<td></td>
<td>Students</td>
<td>5</td>
<td>1</td>
<td>McCauley et al., 1990 video</td>
<td>x</td>
<td>X</td>
<td>x</td>
<td>x</td>
<td>Performance appraisal</td>
</tr>
</tbody>
</table>

Note. PA, performance appraisal. Time represents the reported FOR training time in minutes. Participants represent the type of participants used in the study. No. of Dim represents the number of dimensions used in training and rating. Practice represents how many practice scenarios/videos/items were used in training. Discuss represents whether the trainers discussed the rating dimensions and behavioural examples for the rating dimensions. Written practice represents whether trainees practiced rating either written behaviours and/or written scenarios and received feedback on their ratings. Video practice represents whether trainees practiced rating videotaped behaviour and received feedback on their ratings. In meta-analysis represents whether the study was included in our meta-analysis. True score represents whether true scores/expert ratings were used in the study. Purpose represents the topic of the study. In all cases, we are basing this information on our reading of the article; it may be that the authors had indeed included one of the steps but not reported it.
Table 2. Criteria used by FOR studies over time

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of FOR articles</th>
<th>E</th>
<th>DE</th>
<th>SA</th>
<th>DA</th>
<th>Borman</th>
<th>Distance</th>
<th>Recall &amp; behavioural accuracy</th>
<th>Halo</th>
<th>Leniency</th>
<th>Raw means</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up to 1993</td>
<td>18</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>8</td>
<td>10</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1994–2000</td>
<td>13</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>8</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2001–2010</td>
<td>26</td>
<td>14</td>
<td>15</td>
<td>13</td>
<td>16</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>9</td>
</tr>
</tbody>
</table>

Note. Woehr and Huffcutt (1994) reported four effect sizes for halo, three effect sizes for leniency, six effect sizes for rating accuracy, and two effect sizes for observational accuracy (recall and behavioural accuracy). Please see footnote 1. It was not possible to include all of the studies represented in this table in the meta-analysis.

adequate information, and these statistics are based only on the information reported in the studies. The lack of information in some of the columns, especially in regard to FOR training procedures, does not necessarily indicate that a researcher did not include this procedure but only that the procedure was not discussed in sufficient detail to determine whether these aspects were included.

However, all studies provided information regarding criteria. As indicated in Table 2, the majority of the FOR training studies use rating accuracy as their criteria, with Cronbach’s differential accuracy as the most frequently used criterion. It should be noted that the ‘other’ category represents varied criteria, such as rater agreement, rater reliability, attention time, and performance theory agreement. In general, Cronbach’s accuracy indices have been a popular choice in the last 10 years. Very few FOR training studies in the last 10 years report rater error measures. In fact, of the 26 studies dated since 2000, none reported halo or leniency.

Effectiveness of FOR training

Given the experimental nature of our data, we report our effect sizes in terms of the effect size statistic ‘\(d\)’, 95% confidence intervals, and 80% credibility intervals. All effect sizes represent the difference between a FOR trained group and a control group. The confidence interval provides ‘the range of values that has a 95% chance of containing the estimated mean effect size’ and the credibility interval is interpreted as ‘80% of the values in the \(\rho\) (i.e., true score) distribution lie in this interval’ (Hunter & Schmidt, 2004, p. 205). Table 3 presents the results of the overall meta-analysis and the moderator analyses. Five of the studies also reported pre- and post-training accuracy scores in terms of Cronbach’s accuracy indices (see Table 4). Even though based on a small sample, the results of the pre- and post-accuracy measures present an even more positive view of the effect of FOR rater training.

As seen in Table 3, the overall effect of FOR training was .50, smaller than the effect size of .83 reported in Woehr and Huffcutt (1994). Zero is not in the confidence interval. We would also like to note that only 20.55% of the variance in the observed \(d\)-values can be attributed to sampling error variance. Thus, most likely important moderators influence to what extent FOR training is effective.
Table 3. Meta-analysis results for FOR training

<table>
<thead>
<tr>
<th></th>
<th>K</th>
<th>N</th>
<th>Avg d</th>
<th>SD</th>
<th>95% confidence interval</th>
<th>Lower &amp; upper 80% credibility interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>36</td>
<td>3,754</td>
<td>.50</td>
<td>.39</td>
<td>.37  .63</td>
<td>−.00   1.00</td>
</tr>
<tr>
<td>Woehr and Huffcutt (1994)</td>
<td>6</td>
<td>365</td>
<td>.83</td>
<td>.41</td>
<td>.17  1.33</td>
<td></td>
</tr>
</tbody>
</table>

Moderator analyses

DV operationalization
Cronbach’s accuracy
Elevation
Differential elevation
Stereotype accuracy
Differential accuracy
Borman differential accuracy
Distance accuracy
Recall/Behavioural accuracy

Protocol differences
Source of data
Journal
Dissertation
Other
Type of participants
Students
Employees
Purpose of training
Performance appraisal
Other
Type of training material
Video only
Written scenario only
Both
Published
Yes
No
Number of practice ratees
1
2
3 or more

Note. From left to right, number of effect sizes (k), total sample size (N), the sample size weighted average effect size (Avg d), the standard deviation corrected for sampling error (SD). It should be noted that that we calculated the confidence intervals based on Hunter and Schmidt’s (2004) formulas. Cronbach’s accuracy indices and distance accuracy were reversed scored to be consistent with the remainder of the measures.

Moderator analysis – accuracy operationalization

It appears that FOR training is most successful in improving Borman’s differential accuracy, representing a moderate to large effect (.77). All of Cronbach’s accuracy indices appear to be moderately improved by FOR training, with effect sizes in the low to high .40s. These effect sizes for differential elevation, stereotype accuracy, and differential accuracy are close to a medium effect size (.50 according to Cohen’s, 1988, guidelines). Also, it should be noted that Cronbach’s accuracy indices have overlapping
Table 4. Meta-analysis results for FOR training pre versus post only

<table>
<thead>
<tr>
<th>Accuracy operationalization</th>
<th>K</th>
<th>N</th>
<th>Avg d</th>
<th>SD</th>
<th>Lower and upper 95% confidence interval</th>
<th>Lower and upper 80% credibility interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cronbach’s Elevation</td>
<td>5</td>
<td>369</td>
<td>.22</td>
<td>.09</td>
<td>.09 .35</td>
<td>.10 .34</td>
</tr>
<tr>
<td>Differential elevation</td>
<td>5</td>
<td>369</td>
<td>.51</td>
<td>0</td>
<td>.44 .59</td>
<td>.51 .51</td>
</tr>
<tr>
<td>Stereotype accuracy</td>
<td>5</td>
<td>369</td>
<td>.49</td>
<td>0</td>
<td>.41 .57</td>
<td>.49 .49</td>
</tr>
<tr>
<td>Differential accuracy</td>
<td>5</td>
<td>369</td>
<td>.61</td>
<td>.25</td>
<td>.38 .83</td>
<td>.29 .92</td>
</tr>
</tbody>
</table>

Note. From left to right, number of effect sizes (k), total sample size (N), the sample size weighted average effect size (Avg d), the standard deviation corrected for sampling error (SD), and the upper and lower bounds of the 95% confidence interval (CI). It should be noted that we calculated the confidence intervals based on Hunter and Schmidt’s (2004) formulas. It should also be noted Cronbach’s accuracy indices are reverse coded so that for all effect sizes a positive number indicates more accuracy in the post-test relative to the pretest.

Table 5. Correlations among the accuracy operationalizations

<table>
<thead>
<tr>
<th></th>
<th>CE</th>
<th>CDE</th>
<th>CSA</th>
<th>CDA</th>
<th>DA</th>
<th>BDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cronbach’s Elevation (CE)</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cronbach’s differential elevation (CDE)</td>
<td>.39</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cronbach’s stereotype accuracy (CSA)</td>
<td>.34</td>
<td>.40</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cronbach’s differential accuracy (CDA)</td>
<td>.02</td>
<td>.45</td>
<td>.32</td>
<td>–</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance accuracy (DA)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>.28</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Borman’s differential accuracy (BDA)</td>
<td>.38</td>
<td>.27</td>
<td>.30</td>
<td>.19</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Recall/Recognition</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>.62</td>
<td>.67</td>
<td>–</td>
</tr>
</tbody>
</table>

Note. None of the correlations reached significance (p < .05). The number of effect sizes represented in each correlation ranged from 20 (Cronbach’s elevation and differential elevation) to 2 (Recall/recognition and Cronbach’s distance accuracy). Not enough data points were available to correlate recall/recognition or distance accuracy with many of accuracy operationalizations. The correlations were weighted by sample size.

confidence intervals but the confidence interval for Borman’s accuracy measure does not overlap with Cronbach’s elevation and differential elevation measures and has only minimal overlap with Cronbach’s stereotype accuracy and differential accuracy measures. Table 5 presents the correlations among the accuracy indices. It should be noted that many of these correlations are sizable, larger than those reported by Roach and Gupta (1992), which indicates that the accuracy indices may not be as empirically distinct as believed. However, these correlations are based on relatively few data points ranging from two to 20 depending on the correlation.

We also investigated distance accuracy and recall/behavioural accuracy. Please see Table 3 again. The effect size for distance accuracy also represents approximately a medium effect but the confidence interval included zero and the standard deviation is
Table 6. Regression analyses for continuous moderators

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>SE B</th>
<th>β</th>
</tr>
</thead>
<tbody>
<tr>
<td># of dimensions</td>
<td>-.06</td>
<td>.03</td>
<td>-.31*</td>
</tr>
<tr>
<td>Training time</td>
<td>-.00</td>
<td>.00</td>
<td>-.32</td>
</tr>
</tbody>
</table>

Note. The table above represents two separate regression equations, all data points weighted by sample size. The number of studies for number of dimensions is 31, and 26 studies for training time. For training time, we also analysed the data according to a median split and, mirroring the regression results, there appears to be a greater effect size (.57) associated with training time less than the median than greater than the median (.38). For the number of practice dimensions, we created four categories (two and three dimensions, four and five dimensions, six and seven dimensions, and greater than seven dimensions). Similar to the regression analysis, fewer dimensions are associated with greater effect sizes (.61, .43, .33, .15, respectively). * = p < .10.

large. It should be noted that of the moderator analyses based on accuracy operationalizations, this is the only one with zero in the confidence interval. Thus, more research is needed to determine whether FOR training improves distance accuracy.

The results for recall/behavioural accuracy are more promising. Even though how many behaviours recalled is not the same as behavioural accuracy, usually computed based on signal detection theory (Lord, 1985; Snodgrass & Corwin, 1988), we combined these dependent variables because there were not adequate number of studies to separate them. Both recall and behavioural accuracy are based on specific behaviours. Thus, even though the results regarding our combined measure are tentative, it appears that FOR training may have a large effect on recall/behavioural accuracy (d = .88), the largest effect found in our meta-analysis. Also, the confidence interval for the recall/behavioural accuracy measure does not overlap with Cronbach’s elevation and differential elevation measures and has only minimal overlap with Cronbach’s stereotype accuracy and differential accuracy measures.

Moderator analysis – training protocol differences
Also see Table 3 for analyses of the categorical moderators related to training protocol differences. It appears that studies published in journals, using employees, and for non-performance appraisal purposes have larger effect sizes. However, these results are, at best, tentative. As seen in Table 3, for all moderators, except whether the study was published, the vast majority of the studies fall into one of the categories and only a handful fall in the alternative categories. It should be noted, however, that there does appear to be a curvilinear relationship between the number of practice ratees and the training effect size, with two practice ratees representing the ideal.

Table 6 provides the regression results for the continuous moderators. It does appear that increasing number of dimensions and increasing training time are related to reduced accuracy, but the regression coefficients are not significant and only suggest trends. Thus, we can draw no firm conclusions regarding the effects of protocol differences and only highlight trends.

Discussion
FOR training is alive and well and not just in the performance appraisal literature. We found 39 manuscripts investigating FOR training dated 1994 or more recently, with 26
manuscripts dated since 2000. Of the 14 studies we found investigating FOR training in the last 5 years, only five were in a traditional performance appraisal context: nine were in varying contexts from assessing language proficiency (Dierdorff, Surface, & Brown, 2010), competency modelling (Lievens & Sanchez, 2007) to evaluating biodata items (Lundstrom, 2008). Thus, FOR training no longer is just a performance appraisal topic. Some of the criteria of interest have also changed since the Woehr and Huffcutt (1994) meta-analysis. The manuscripts written since the Woehr and Huffcutt meta-analysis use a number of criteria to evaluate FOR training. Unfortunately, the criteria not related to accuracy in these recent studies vary too greatly to be examined meta-analytically. Given that FOR training was designed to increase rating accuracy, it is not surprising that accuracy is still the criterion of choice. We found that differential accuracy is the most commonly used criterion with the remainder of the Cronbach’s accuracy indices also widely used. It is also not surprising that rater error indices have fallen out of favour, given Murphy and colleagues’ (Murphy & Balzer, 1989; Murphy et al., 1993) suggested moratorium on the use of rater error measures. We believe that FOR training has become the rater training method of choice partly because rater error training has declined in popularity given its focus on rater errors and partly because the Woehr and Huffcutt (1994) meta-analysis concluded that FOR training has the largest increase in regard to rating accuracy, in comparison to the other rater training paradigms discussed in their meta-analysis.

It is worth noting that even though accuracy indices have been criticized as insufficient indicators of rating quality (e.g., Murphy & Cleveland, 1995; Sulsky & Balzer, 1988) and other criteria representing rating quality have gained popularity (e.g., fairness perceptions, utility perceptions, etc.; Hedge & Teachout, 2000; Murphy & Cleveland, 1995), accuracy indices are still commonly used to investigate performance ratings (London, Mone, & Scott, 2004). For example, in the last 3 years, Cronbach’s accuracy indices were used as indicators of rating quality in several empirical studies (e.g., Bernardin, Tyler, & Villanova, 2009; Gorman & Rentsch, 2008; Melchers, Kleinmann, & Printz, 2010; Melchers, Lienhardt, von Aarburg, & Kleinmann, 2011; Uggerslev & Sulsky, 2008). These are only a few of the more recent studies investigating rating accuracy. Many more such studies have been conducted in the last 10 years. Thus, despite the common criticism that accuracy indices are of limited use, given the need of a standardized performance viewed by all raters (both experts and naïve), for some research questions rating accuracy still represents the most relevant criterion.

It is clear that FOR training does improve rating accuracy. Across all operationalizations of accuracy, FOR training had an average effect size of $d = .50$, somewhat smaller than the effect size ($d = .83$) reported by Woehr and Huffcutt (1994). Moreover, as would be expected with the larger number of studies and corresponding increase in sample size, the current estimate of the impact of FOR training is far more precise than that provided by Woehr and Huffcutt (1994). As mentioned earlier, Woehr and Huffcutt based their conclusions regarding rating accuracy on only six effect sizes and 365 participants. Our conclusion is based on 117 effects sizes from 36 studies and 3,754 participants.

However, it appears that FOR training does not impact all accuracy operationalizations to the same degree. The largest effect size was associated with Borman’s correlation-based differential accuracy measure ($d = .77$). Borman’s measure indicates how closely participants’ relative rank ordering of ratees corresponds to established standards (i.e., ‘true’ scores). This suggests that FOR training may be most effective with respect to training raters to recognize relative patterns of performance as opposed to absolute levels.
Thus, it appears that FOR trained raters more accurately detect patterns of strengths and weaknesses and more accurately rank order ratees than untrained raters. However, it is also possible that the larger effect size for Borman’s measure may also be partly due to the statistical properties of correlations versus distance indices. Borman’s measure only requires raters to rank order ratees but the other accuracy operationalizations require raters to both rank order ratees correctly and to match ‘true scores’.

Perhaps the most surprising finding of the meta-analysis is the large effect size for recognition/recall ($d = .88$), much larger than the one ($d = .37$) found in Woehr and Huffcutt (1994). However, even though our sample is four and a half times as large ($n = 9$ compared to $n = 2$) as the sample used in Woehr and Huffcutt, it is still small. We made the decision to combine recall and recognition, even though they depend on different cognitive processes, to maximize our number of data points. Both recall and recognition focus on the accuracy of observation versus the accuracy of rating. As discussed earlier, FOR training is designed to improve rater categorization and, therefore, rating accuracy and not observational accuracy. However, FOR training most often does include a discussion of specific behaviours that belong to performance dimensions. Perhaps this behavioural emphasis in FOR training sensitizes the rater to the importance of observing behaviours and thus improves behavioural accuracy.

These results provide an interesting snapshot into rater cognitive processes. The results regarding behavioural accuracy suggest that FOR training not only helps to create shared performance categories but perhaps it also encourages raters to create stronger behavioural memories by emphasizing the importance of behaviour. An alternative explanation is that increased behavioural accuracy may be due to the creation of performance categories that are also associated with more relevant behaviours than the untrained rater’s idiosyncratic categories. Future research is needed to explore which of these explanations is the case. Regardless, the results show that FOR training influences cognitive processing of behavioural information.

The differences among accuracy operationalizations also have some interesting implications regarding rater cognitive processing. Given the stronger results for Borman’s differential accuracy measure than for the accuracy operationalizations based on the distance between raters and ‘true scores’, it does seem that FOR training may help raters categorize ratees but these categories may not be based on specific numerical labels but more general category labels such as poor, average, or good bins as suggested by Murphy and Cleveland (1995). Whether this is an issue depends on the use of the ratings. If the ratings are used in a cut-score situation in which those above a certain score are treated differently than those below a certain score, it is important that raters learn how to mimic true scores. In this case, perhaps a more enriched FOR training protocol, such as that suggested by Sulsky and Kline (2007) involving modelling and role-playing, would be beneficial. However, if the purpose of the ratings is simply to decide whom to hire, terminate, or promote based on the rank ordering of candidates, the current FOR protocols involving an introduction to the measure, rating practice, and feedback may be sufficient.

Furthermore, if FOR training is indeed most successful in improving correlational types of accuracy, it may be that Chirico et al.’s (2004) alternative FOR training in which ‘true’ ratings are not used may be worth an increased consideration. In the Chirico et al. (2004) version of FOR without ‘true’ ratings, raters still practice assigning ratings and receive feedback but the feedback is in more qualitative terms that does not emphasize specific numbers. Developing the true scores needed to conduct FOR training
is time intensive and may be perceived as one of the barriers to implementing FOR training in organizations.

Another potential barrier may be that little is known regarding to what extent FOR training persists over time. Very few studies have investigated FOR training’s effects over time, and to the best of our knowledge, the longest time frame investigated was 2 weeks (Roch & O’Sullivan, 2003). Researchers should investigate time frames that represent organizational performance appraisal cycles more closely. If it does appear that FOR training accuracy declines over more realistic time frames, researchers should investigate the benefits of short FOR refresher training sessions. It would also be beneficial if future research investigated the perceived barriers to implementing FOR training in organizations. Once researchers better understand the perceived barriers to the implementation of FOR training in organizations, they will be able to better focus their research to eliminate these barriers.

Future FOR research should also transition from focusing on rating accuracy to focusing on both rater motivation and rater and ratee reactions, in addition to investigating relevant evaluation criteria suggested by the training literature. Our results show that FOR training increases accuracy; more research establishing the effectiveness of FOR training within the training session is not needed. However, research investigating to what extent FOR training transfers out of the training session is needed. The issue is that in most applied settings the accuracy of performance ratings is impossible to establish outside of the rater training session. Thus, researchers may wish to focus on rater and ratee perceptions and motivation instead when investigating the effectiveness of FOR training outside of the training session. As mentioned earlier, rater motivation and rater and ratee reactions have become a primary focus of performance appraisal researchers (e.g., Levy & Williams, 2004). However, both rater and ratee reactions have been largely ignored by rater training researchers, with a few exceptions (e.g., Uggerslev & Sulsky, 2007; Sulsky & Kline, 2007).

Furthermore, the FOR training literature has virtually ignored the greater training evaluation literature. Depending on the context and the research question, FOR training researchers should perhaps consider Kirkpatrick’s (1994) training criteria (reactions, learning, behaviours, and results) or Kraiger’s (2002) more sophisticated training evaluation taxonomy, concentrating on cognitive, affective, and behavioural training outcomes. The most relevant criterion depends on the research question, which in turn is partly a function of the training context. For example, different research questions are relevant for FOR training in the context of job analysis than for performance appraisal; the question asked should determine the criteria chosen.

Regardless of the criteria chosen, FOR training researchers should move beyond only comparing the post-training ratings or reactions of a trained group to an untrained group, especially in applied settings where random assignment to training condition often is not possible. Cook, Campbell, and Peracchio (1990) provide a wealth of quasi-experimental designs useful in evaluating training, including FOR training. Thus, when evaluating FOR training, researchers and practitioners should not just consider the performance appraisal literature when choosing appropriate criteria to evaluate the success of the training but should consider the training evaluation literature. Choosing training evaluation criteria from the training literature, such as evaluating not only reactions to training but also the amount learned in training and changes in rating behaviour from before to after FOR training, may help to convince practitioners of the benefits of FOR training. Overall, the results of our meta-analysis show that FOR training is not only alive and well but also is an effective rater training method.
References

*References included in the meta-analysis.


*Received 16 February 2011; revised version received 25 August 2011*