Nature and Consequences of Halo Error: A Critical Analysis

Kevin R. Murphy, Robert A. Jako, and Rebecca L. Anhalt

The definition of halo error that dominated researchers' thinking for most of this century implied that (a) halo error was common; (b) it was a rater error, with true and illusory components; (c) it led to inflated correlations among rating dimensions and was due to the influence of a general evaluation on specific judgments; and (d) it had negative consequences and should be avoided or removed.

We review research showing that all of the major elements of this conception of halo are either wrong or problematic. Because of unresolved confounds of true and illusory halo and the often unclear consequences of halo errors, we suggest a moratorium on the use of halo indices as dependent measures in applied research. We suggest specific directions for future research on halo that take into account the context in which judgments are formed and ratings are obtained and that more clearly distinguish between actual halo errors and the apparent halo effect.

When an individual is rated on multiple performance dimensions or attributes, the rater's overall impression or evaluation is thought to strongly influence ratings of specific attributes (Cooper, 1981b), a phenomenon that is referred to as halo error (Thorndike, 1920). Discussions of halo error are most frequently encountered in the context of evaluative judgment (e.g., in interviews and performance appraisals), but similar phenomena have been noted in research on illusory correlation (Chapman & Chapman, 1969), implicit personality theory (Lay & Jackson, 1969), and interpersonal judgments (Nisbett & Wilson, 1977).

Research on halo errors in rating can be traced back to the early part of this century (Thorndike, 1920; Wells, 1907). Although there are a number of different conceptual and operational definitions of halo (Balzer & Sulsky, 1992; Saal, Downey, & Lahey, 1980), throughout most of the history of research on halo error, there has been some consensus regarding the nature and consequences of halo error. First, halo error is thought to be common (Bernardin & Beatty, 1984; Blum & Naylor, 1968; Cascio, 1991; Feldman, 1986; Jacobs & Kozlowski, 1985). In this vein, Cooper (1981b) referred to the halo error as "ubiquitous." Second, halo is typically treated as a rater error. That is, halo error is assumed to reflect the rater's inability (Cooper, 1981a, 1981b; Feldman, 1986; Lance & Woehr, 1986; Murphy & Jako, 1989; Nathan & Lord, 1983; Saal et al., 1980; Vance, Winne, & Wright, 1983) or unwillingness (Banks & Murphy, 1985) to discriminate among different attributes of the individuals rated.

Third, observed halo (often operationalized in terms of the average intercorrelation among ratings) is thought to reflect two distinct influences, true halo and illusory halo; the labels valid halo and invalid halo are also sometimes used. (Bartlett, 1983; Bingham, 1939; Cooper, 1981b; Lance & Woehr, 1986; Murphy, 1982; Pulakos, Schmitt, & Ostroff, 1986). That is, observed halo reflects both the real overlap among the dimensions being rated (true halo) and a host of irrelevant factors, including deficiencies in measurement, observation and memory errors on the part of the rater, incorrect theories of covariation, and so on, that taken together constitute illusory halo. This formulation is analogous to the statement in classical measurement theory that observed scores equal true scores plus errors in measurement (Lord & Novick, 1968). As in classic measurement theory, the goal of much of the existing research on halo has been to understand and eliminate the error component, which in this case is referred to as illusory halo.

Fourth, halo error leads to inflated correlations among the attributes rated (Bernardin & Beatty, 1984; Cascio, 1991; Cooper, 1981b; Lance & Woehr, 1986; McCormick & Ilgen, 1985; Nathan & Lord, 1983; Pulakos et al., 1986); that is, the observed correlations among ratings are usually assumed to be larger than the corresponding true intercorrelations.

Fifth, the most probable cause of halo is that the rater's overall evaluation of the ratee influences his or her ratings of specific attributes (Bernardin & Beatty, 1984; Cooper, 1981b; Feldman, 1986; Fisicaro & Lance, 1990; Landy, 1989; Muchinsky, 1987; Murphy, 1982). Thorndike (1920), who coined the term halo error, defined it as "suspending ratings of special features with a halo belonging to the individual as a whole" (p. 25). Cognitively oriented research has examined a number of alternative ways in which global evaluations or evaluative schema might affect halo errors (e.g., Feldman, 1986; Nathan & Lord, 1983) but, with few exceptions (e.g., Lance & Woehr, 1986), the assumption has been that the direction of the effect is top-down, in other words, that general evaluations affect specific ratings.

Sixth, halo error is thought to have a negative impact on the quality of ratings (Cooper, 1981b; Landy, Vance, Barnes-Farrell, 1983).
correlations are often smaller than true correlations, other researchers (e.g., Vance et al., 1978) have reported similar findings. Taken together, these studies lead to the conclusion that halo errors (as traditionally defined) are likely to occur only in circumstances in which the true correlations are very low.

The evidence reviewed in this section suggests that halo errors are not ubiquitous, that inflated correlations among rating dimensions are not the norm, and that top-down theories are not always appropriate for describing the relationships between general and specific ratings.

**Comparisons Between True and Observed Correlations**

The assumption that observed correlations are typically larger than true correlations is critical to understanding common assumptions about the causes and frequency of halo errors. As Murphy and Jako (1989) noted, all top-down theories of halo (i.e., theories in which general evaluations influence ratings of specific attributes) share one characteristic, that the correlations between specific dimensions cannot be reduced and usually must be inflated as a result of the common influence of the general evaluation on the specific ratings that are being correlated (see Loehlin, 1987, for a path-analytic demonstration of this principle). That is, if halo errors are thought to be partly the result of the influence of general evaluations on ratings of specific dimensions, observed correlations should generally be higher than the true correlations among the dimensions rated. However, this is not always (or even usually) the case.

There are now a number of studies, using a variety of methods, stimuli, and rating tasks that demonstrate that observed correlations are often smaller than the corresponding true correlations (Fisicaro, 1988; Kenny & Berman, 1980; Kozlowski & Kirsch, 1987; Lance, Fisicaro, & LaPointe, 1990; McIntyre, Smith, & Hassett, 1984; Murphy & Balzer, 1986; Vance, Kunhert, & Farr, 1978). Fisicaro and Lance (1990) cautioned that the direct comparison of observed and true correlations can lead to incorrect conclusions about negative halo error (i.e., observed correlations that are smaller than true correlations). However, even in their study, which corrected for potential biases in the assessment of halo, there was evidence that some raters underestimate true halo. We know of no studies in which (a) the true correlations are known or can be reasonably estimated, and (b) there is consistent evidence that observed correlations exceed true correlations.

Studies by Murphy and Reynolds (1988) and Murphy and Jako (1989) illustrate the conditions under which observed correlations might either be smaller or larger than the true correlations among the dimensions being rated. They constructed rating tasks in which the levels of true halo could be manipulated by changing the actual behavior observed by different groups of raters. When the dimensions to be rated were, in fact, essentially orthogonal, most of the subjects in both studies committed what is traditionally referred to as halo error, that is, observed correlations greater than true correlations. However, when the level of true halo was high, most of the subjects in both studies committed what look like negative halo errors, that is, they showed observed correlations smaller than true correlations. Other researchers (e.g., Vance et al., 1978) have reported similar findings. Taken together, these studies lead to the conclusion that halo errors (as traditionally defined) are likely to occur only in circumstances in which the true correlations are very low.

**Consequences of Halo and Attempts to Remove Halo**

Although it is usually assumed that halo error detracts from the quality of rating data, there are a number of reasons to believe that halo errors are not, in fact, a serious problem and that halo errors may even contribute to the accuracy and utility of ratings.

The presence of halo errors in rating has usually been taken as evidence that raters are inaccurate. A number of review articles (e.g., Becker & Cardy, 1986; Cooper, 1981b; Fisicaro, 1988) and meta-analysis (Murphy & Balzer, 1989) have shown, however, that halo error does not automatically limit rating accuracy and may, in some circumstances, lead to increased accuracy. Cooper (1981b) referred to this as the halo-accuracy paradox. Subsequent analyses (e.g., Fisicaro, 1988) suggest that this paradox may be partly a function of the halo and accuracy measures used in different studies and that the likelihood of a positive halo-accuracy correlation may vary as a function of the levels of true halo likely to be encountered in different settings. Nevertheless, it now seems clear that halo error, as traditionally defined, does not necessarily imply low levels of accuracy.

The finding that various halo measures are positively corre-
lated with a number of accuracy measures has been replicated in a number of settings (Becker & Cardy, 1986; Kozlowski & Kirsch, 1987; Lorenzo, 1984; Murphy & Balzer, 1989). Other studies have shown that what appear to be high levels of halo are not an impediment to accuracy in rating (Murphy & Balzer, 1986; Weekly & Gier, 1989). Finally, Nathan and Tippins (1990) have shown that halo in rating can lead to higher levels of validity when performance ratings (which are thought to be tainted by halo error) are validated against a variety of ability measures (which are not so tainted).

It is difficult to describe in any simple way the relationship between halo and rating accuracy, in part because of the large number of alternative indices of both halo and accuracy (Becker & Cardy, 1986; Cronbach, 1955; Saal et al., 1980; Sulsky & Balzer, 1988). However, some halo-accuracy relationships are likely to be more important than others, and if we concentrate our attention on the most critical halo-accuracy relationships, it is possible to make useful generalizations. In particular, we believe that the most important halo-accuracy relationships are between correlational measures of halo and measures of accuracy in distinguishing between rates and that the relationships between these types of measures can be easily explained in terms of classical psychometric theory.

Several authors have argued that accuracy in discriminating between rates in terms of their overall level of performance is more important than a variety of other types of accuracy (e.g., accuracy in discriminating a person's strengths from his or her weaknesses; see Murphy, Garcia, Kerkar, Martin & Balzer, 1982; Saal & Knight, 1988). Most administrative decisions (e.g., raises, promotions) depend primarily on the rank ordering of the individuals being rated and are virtually unaffected by the pattern of ratings each individual receives. Halo error makes it more difficult to distinguish an individual's strengths from his or her weaknesses but makes it easier to distinguish between rates, thus contributing to this highly critical form of rating accuracy (Murphy & Balzer, 1986). This can be shown by drawing an analogy to psychometric theory.

Consider a performance appraisal in which ratings are collected on several performance dimensions and summed to provide an index of overall performance. Formulas for the reliability of a linear combination (Nunnally, 1978) make it clear that reliability increases as the correlation among the dimensions increases. Many of the halo-accuracy correlations reported in the literature may be in part due to the increase in reliability that accompanies halo error (Jacobos & Kozlowski, 1985).

A number of strategies have been suggested for removing halo error, including the statistical control of halo (Landy et al., 1980), increased observation (Cascio, 1991), manipulation of rating scales and scale formats (Bartlett, 1983), and rater training (Bernardin & Pence, 1980), but none of these has proved to be fully effective. In some cases, they fail to control halo. In others, they control things other than halo (Murphy, 1982). In still others, they succeed in controlling halo error but do little to increase the quality of the rating data.

Strategies for statistically controlling halo have received the most attention (Becker & Cardy, 1986; Lance & Woehe, 1986; Mossholder & Giles, 1983; Murphy, 1982; Nathan & Lord, 1983). It seems clear that techniques such as partialing out the overall rating from each dimensional rating will not work. These techniques seem to remove true as well as illusory halo and do little to improve the usefulness of the resulting rating data.

In a widely cited study, Bernardin and Pence (1980) suggested that it was indeed possible to train raters to avoid rating errors such as halo. The problem, however, is that raters may simply substitute some new, equally invalid rating set for the old one. The evidence is somewhat equivocal regarding the effects of increased opportunities to observe ratee performance on halo errors. Cascio (1991) suggested that increased opportunities to observe ratee performance do not reduce halo. On the other hand, Kozlowski and his colleagues (e.g., Kozlowski & Kirsch, 1987; Kozlowski, Kirsch, & Chao, 1986) suggested that familiarity with the ratee and job knowledge may interact in affecting halo and that, in some circumstances, more observation might lead to less halo.

In sum, the evidence reviewed in this section leads to the conclusion that halo error is not always a problem (something that was also suggested by Wells, 1907), and that it may actually increase the accuracy and validity of ratings, especially accuracy and validity in distinguishing between rates. Efforts to control halo error in rating have not proved successful; if a method could be devised to control halo, it would quite possibly do more harm than good.

**Separating True From Illusory Halo**

In theory, it is indeed possible to separate true from illusory halo; as in true-score theories of measurement, both true and illusory components exist by definition (e.g., true halo = observed halo minus illusory halo; Lord & Novick, 1968, discuss in detail various definitions of true and error components of measurement). However, there are also several reasons to believe that the division of halo into true and illusory components is, in practice, pointless. First, we argue that it is almost impossible to measure true halo in most settings. Second, we review research showing that in settings in which true halo can be measured, its effects are minimal. Except in those rare circumstances in which the dimensions being rated are truly orthogonal, the observed correlation between dimensional ratings represents a composite of the true correlation and the net result of the cognitive distortions, errors in observation and judgment, and rating tendencies of the individual rater (i.e., illusory halo). This suggests that the distinction between true and illusory halo is fundamental; unless the true part can be removed from measures of observed intercorrelation among performance dimensions, it will never be possible to tell the extent to which observed correlations represent rater errors versus accurate depictions of the relationships among dimensions. In a number of laboratory studies and in some archival studies, it has been possible (and useful) to separate true from illusory halo (Kozlowski & Kirsch, 1987; Murphy & Jako, 1989; Murphy & Reynolds, 1988). However, the same is not true for most field settings. In most nonlaboratory settings, it is virtually impossible to measure or even estimate the level of true halo present in a set of ratings.

To understand the difficulty in measuring true halo, consider the following example. An organization has three supervisors, each of whom rates 15 subordinates on four performance
New Directions for Research on Halo Error in Rating

It is clear that the conceptualization of halo that has dominated research and application since the 1920s is no longer adequate. This does not mean that existing theory and research on halo is useless. On the contrary, there are many critical findings from this research that must be taken into account in future research on halo. For example, there is a considerable body of evidence showing that individuals often do make correlational errors in rating (see Cooper, 1981b, for a review). Furthermore, there is a good deal of evidence suggesting that the type of top-down processing that is assumed to underlie halo error (i.e., general impressions affecting more specific judgments) often does occur (Feldman, 1986; Fisicaro & Lance, 1990; Lance & Wohr, 1986; Murphy, 1982; Murphy, Martin, & Garcia, 1982; Nathan & Lord, 1983). However, the research reviewed here makes it clear that halo errors are far from ubiquitous. It is critical to understand why halo errors occur in some situations and not in others. In addition, the evidence reviewed here shows that halo error is not always a bad thing. It is critical to understand precisely what is gained or lost through halo errors in rating and to identify situations in which halo is or is not a problem.

The sections that follow suggest several directions for future research on halo errors in rating. First, problems in distinguishing halo errors from a variety of other cognitive distortions or potential shortcomings in ratings are discussed. Second, contextual influences on halo error and on the consequences of halo errors (and efforts to remove them) are examined.

Inferring and Identifying Halo Errors

The problem of determining whether or when halo errors occur has received little attention, for two reasons. First, most researchers have assumed that these errors are ubiquitous and that the only real question is how much halo error is present in any given set of ratings. Second, research on halo has not consistently distinguished between the rating error that occurs when correlations are inflated and the rater error that is thought to be responsible for this effect (Fisicaro & Lance, 1990). That is, it is often difficult, when reading research on halo, to distinguish between discussions of halo as a cause and halo as an effect. The distinction between the halo effect (i.e., inflated correlations) and halo errors (i.e., spillover of general impressions onto ratings of specific dimensions) is critical, for two reasons.

First, because the level of true halo may vary across raters, situations, time, and so on, there is no easy way to determine whether the correlations among rating dimensions are, in fact, inflated. Therefore, a simple examination of the correlations among ratings on separate dimensions does not provide sufficient evidence for recognizing either the halo effect or halo errors. Second, even if it could be shown that observed correlations were larger than their true counterparts, inflated correlations do not necessarily imply that halo errors have occurred; any number of processes might lead to inflated correlations among ratings. For example, a rater who is asked to fill out a performance appraisal form with nine dimensions might mentally group them into several related sets, none of which represents his or her global evaluation of the ratee. Both schema theories and associationist theories (e.g., Anderson, 1983) could be used to explain inflated correlation without recourse to any global evaluative dimension.

The inferential strategy typically used to infer halo may need to be completely reversed. Typically, when the halo effect is thought to be present in a set of ratings (e.g., on the basis of high intercorrelations), the halo error is inferred. A better analytic strategy might be to first identify situations in which the type of top-down processing that characterize halo error is most likely to occur and in which there are no serious situational constraints to inflated correlations (e.g., ceiling effects in ratings) and then infer that, in those situations, the halo effect is most likely to occur. In the section that follows we examine research
on situational or contextual determinants of halo that might provide a starting point for this type of analysis.

Several years ago, Landy and Farr (1980) called for a moratorium on research comparing different rating scale formats. We think that it is time for a similar moratorium on the use of indices of observed halo as dependent measures in applied psychological research. This is not to say that halo measures are always useless; in situations in which the level of true halo is relatively homogeneous across raters and situations, these measures might be quite useful. The reason for the proposed moratorium is that, at present, we have no way of knowing whether the halo measures used in a particular study or context are good measures of halo error or uninterpretable jumbles of true and illusory halo. Until we develop a sufficient understanding of the factors that might make observed halo measures either appropriate or uninterpretable, the continued use of these measures simply makes no sense.

Effects of Context on Halo Errors and the Halo Effect

The context in which judgments about performance are formed and in which formal performance appraisals are obtained is likely to influence both halo errors and the halo effect. Several models of the performance appraisal process (e.g., Balzer & Sulsky, 1990; Cleveland & Murphy, 1992; Landy & Farr, 1980) suggest that it is useful to examine separately the ways in which contextual variables influence judgments, ratings, and the evaluation of those ratings.

Contextual effects on judgments. The halo error represents a distortion in judgment in which overall impressions have an undue effect on specific judgments. Several contextual factors are thought to contribute to this sort of error. First, substantial delays between observation and judgment (e.g., appraisals usually cover periods of several months to a year) increase the likelihood that raters will rely on global impressions in making specific judgments (Murphy & Balzer, 1986; Murphy et al., 1982; Murphy & Reynolds, 1988; Nathan & Lord, 1983). Second, the rater's familiarity with the ratee may influence halo; ratings of unfamiliar ratees seem to show more halo (Kozlowski & Kirsch, 1987; Kozlowski et al., 1986). This may be due to the relatively undifferentiated impressions of unfamiliar ratees; with experience, raters may develop more fine-grained impressions.

Third, several characteristics of the rating task can influence halo. For example, research in social cognition suggests that making judgments about a particular dimension affects memory for and judgments concerning other related performance dimensions (Judd, Drake, Dowing, & Krosnick, 1991; Torangeau & Rasinski, 1988). Whereas research on halo has, for the most part, concentrated on top-down models, this research has drawn from associationism models that suggest that making judgments about one dimension primes or activates information about other related dimensions, which may affect subsequent judgments without necessarily involving the rater's global evaluation of the ratee. This spreading activation model suggests that the order of dimensions may affect the degree of halo in judgments. If early judgments involve dimensions that are strongly related to one's global evaluation, spreading activation effects could inflate halo. If early judgments involve dimensions largely unrelated to one's global evaluation, spreading activation effects could diminish halo. Finally, halo errors seem most likely when there are only a few dimensions, each of which is highly relevant to one's overall evaluation, and least likely when there are many dimensions, several of which are apparently unrelated to overall performance.

Contextual effects on ratings. It is important to distinguish between the rater's judgments about each subordinate and the ratings he or she records on performance appraisal forms (Banks & Murphy, 1985; Murphy & Cleveland, 1991). Distortions in ratings that lead to inflated intercorrelations (i.e., the halo effect) may reflect errors in judgment, but they might also reflect a variety of contextual effects that have relatively little to do with the rater's private judgments about the strengths and weaknesses of his or her subordinates. For example, research on the effects of the purpose of rating suggests that when ratings are used to make administrative decisions (e.g., salary or promotion), (a) ratings tend to be inflated, and (b) ratings tend to minimize within-ratee distinctions and maximize distinctions between ratees (Cleveland, Murphy, & Williams, 1989; Landy & Farr, 1983; Murphy & Cleveland, 1991). Research on rater goals (Cleveland & Murphy, 1992) and on political behavior in appraisal (Bjerke, Cleveland, Morrison, & Wilson, 1987; Longenecker, Sims, & Gioia, 1987) suggests that these effects represent distortion in ratings rather than distortions in raters' judgments. This research suggests that raters often consciously distort ratings to achieve a variety of goals (e.g., win promotions for deserving subordinates or preserve harmony in the work group), particularly when ratings are strongly tied to administrative decisions. These distortions in ratings can either inflate dimensional intercorrelations (i.e., when raters exaggerate differences between ratees) or diminish them (i.e., when ratings are so inflated that it is not possible to distinguish between ratees).

As an example of the process described earlier, Bjerke et al. (1987) described how the context in which evaluations of officers in the U.S. Navy are obtained contributes to rating inflation and range restriction (see Hackworth, 1989, for a less technical description of the same dynamics in evaluating Army officers). Because of the norms of the organization and "up-or-out" retention policies (i.e., officers who are passed over for promotion within specified intervals must retire), there is a strong tendency to give uniformly high ratings to most officers. In fact, one of the key problems a rater faces in this context is communicating the fact that a particular subordinate really is a star performer when almost everyone receives near-perfect ratings. Although raters may in fact commit halo errors in forming their private evaluations of their subordinates, the demands of the rating system lead to extremely low correlations among rating dimensions.

As noted earlier, likelihood of the halo effect depends largely on whether the true correlations among dimensions are large or small. If true correlations are already large, it may be difficult to further inflate them no matter what rating strategy is used. Note that the level of true correlation may not have a marked effect on the likelihood of halo errors. However, when the true correlations are already quite high, it may be hard to discriminate (on the basis of observed intercorrelations) between raters who either rely too heavily on or completely ignore global im-
pressions in rating subordinates on specific performance dimensions.

Contextual influences on rating may also provide an explanation for negative halo (i.e., observed correlations smaller than true correlations; see Lance et al., 1990, for discussions of statistical issues related to negative halo). Rating scales typically require the rater to provide judgments about 5–9 performance dimensions, although some appraisal systems might require judgments on many more dimensions. Furthermore, from our experience, it seems that the distinctions between dimensions are sometimes quite subtle. The fact that separate judgments are required should create a demand characteristic, that is, the assumption that raters should receive somewhat different ratings on different dimensions. We hypothesize that the likelihood of negative halo increases as the number of dimensions increases, as their distinguishability decreases, and as the true correlations among dimensions increase. There is empirical support for the hypothesized role of true intercorrelations (Murphy & Jako, 1989; Murphy & Reynolds, 1988); the other two portions of this hypothesis have not been directly tested.

Contextual influences on the evaluation of halo errors. Funder (1987) distinguished between errors and mistakes in interpersonal perception. An error is a deviation from a normative standard that may not have any real-world implications. A mistake is more serious and causes a person to make an incorrect decision or choice. It is clear from the research reviewed here that halo is sometimes an error but rarely a mistake. Indeed, it is likely that halo “error” is often highly adaptive in the sense that it is the result of the rater focusing on what is important and paying less attention to less critical features of the ratees’ behavior (Cleveland & Murphy, 1992; Funder, 1987; Murphy & Balzer, 1986). The question of whether halo errors should be suppressed or avoided depends largely on the context in which judgments are elicited or ratings are obtained.

As noted earlier, halo error makes it easier to distinguish between rates and harder to distinguish within rates (Cleveland et al., 1989). That is, halo errors magnify differences in the mean ratings received by different individuals and flatten the overall profile of ratings, compressing differences between ratings of different areas of performance. It follows that halo is an error (i.e., incorrect but not harmful) when the primary purpose of rating is to distinguish good from poor performers, and that it is a mistake (i.e., incorrect and harmful) when the primary purpose of rating is to distinguish strengths from weaknesses. Surveys of the uses of performance appraisal (e.g., Cleveland et al., 1989) suggest that ratings are most often used to distinguish between individuals in terms of their overall performance, which implies that halo errors are more likely to contribute to the quality of measurement (by increasing the reliability of summated ratings) than to detract.

Assuming that halo errors are sometimes beneficial, it might be useful to know how we could encourage halo errors. It might be possible to design a number of interventions that would increase the reliability of discriminations between rates, at the perhaps negligible cost of reducing the reliability of within-person distinctions. Examples might include adjustments to rating scales (e.g., include dimensions that are strongly related to the global evaluations, or “g-loaded”), rater training programs (e.g., encourage raters to explicitly consider the ratee’s overall performance in rating specific dimensions), or rating procedures (e.g., ask raters to record their overall evaluation before rating specific dimensions).

The assertion that halo errors are rarely harmful and sometimes beneficial adds further force to our concerns about the continued use of halo indices as dependent variables in applied research. In situations in which the mix of true and illusory halo is unknown and in which the consequences (if any) of making halo errors are unclear, we do not see a clear justification for using halo measures as dependent variables. Halo measures collected under these circumstances (which we believe are representative of most rating situations) will be virtually uninterpretable, and any conclusions drawn on the basis of such measures (e.g., Training Program A is better than Program B because it leads to less halo) have a high likelihood of being wrong.

Summary

The research reviewed here leads to three critical conclusions, as follows: (a) Halo is not ubiquitous, (b) the presence of halo does not necessarily detract from the psychometric quality of ratings, and (c) it is impossible to separate true from illusory halo in most field settings. Contrary to assumptions that have guided halo research since the 1920s, it is often difficult to determine whether halo errors have occurred, why they have occurred (e.g., they might result from errors on the part of the rater or from contextual factors unrelated to the rater’s judgment of the individuals being rated), or what to do about them.

Because of the ambiguity inherent in all operational measures of halo (i.e., the difficulty of distinguishing true from illusory halo and errors from mistakes; see Balzer and Sulsky, 1992, for additional critiques of halo measures), we have called for a moratorium on the use of halo indices as criteria for evaluating training programs, rating scales, and so on. However, this does not mean that research on halo error should be abandoned. Rather, we believe that this research should be redirected to solving the following three critical problems: (a) determining whether or when a halo error has occurred, (b) determining how contextual factors affect measures of observed halo, and (c) determining when halo errors are harmful or beneficial. We believe that substantial progress must be made in all three areas before it will be possible to develop measures of halo that are useful in the field.

References

Bartlett, C. J. (1983). What's the difference between valid and invalid


