Is There a General Factor in Ratings of Job Performance? A Meta-Analytic Framework for Disentangling Substantive and Error Influences

Chockalingam Viswesvaran  
Florida International University

Frank L. Schmidt  
University of Iowa

Deniz S. Ones  
University of Minnesota, Twin Cities Campus

A database integrating 90 years of empirical studies reporting intercorrelations among rated job performance dimensions was used to test the hypothesis of a general factor in job performance. After controlling for halo error and 3 other sources of measurement error, there remained a general factor in job performance ratings at the construct level accounting for 60% of total variance. Construct-level correlations among rated dimensions of job performance were substantially inflated by halo for both supervisory (33%) and peer (63%) intrarater correlations. These findings have important implications for the measurement of job performance and for theories of job performance.

Because of its important implications in personnel selection and in the development of theories of work and work performance, halo error in ratings of job performance has been studied for nearly 100 years. It was first identified by Wells (1907) and was further explored by Thorndike (1920), who coined the term halo error. Guilford (1954, pp. 279–288), Stanley (1961), and Kenny and Berman (1980), as well as Zyzanski (1962), among others, have developed quantitative models of the halo and other variance components in ratings that facilitate understanding of the nature and extent of halo in ratings. Beginning in the early part of the 20th century and continuing to the present, numerous articles and studies have appeared in the research literature on the problem of halo error in ratings. W. H. Cooper (1981) provided an extensive review of this literature on different conceptualizations of halo error, methods of assessing halo error, and effects of halo error.

The question of halo in ratings is closely linked with the disagreements in the literature as to whether there is a general factor in job performance ratings beyond the effects of halo. Factor analyses of ratings data (typically intrarater ratings data) almost always reveal the existence of a large general factor. Some authors have argued that this general factor is due largely or entirely to halo error (Holzbach, 1978; Landy, Vance, & Barnes-Farrell, 1982). Others have argued that there is reason to believe the general factor reflects, at least in part, true positive manifold among job performance dimensions (e.g., Feldman, 1981; Motowidlo, Borman, & Schmit, 1997; see discussion below).

Because of problems involved in separating halo error from true interdimension correlation, this issue has been difficult to address empirically (Borman & Brush, 1993; Borman & Motowidlo, 1993, 1997; Borman, White, Pulakos, & Oppler, 1991; J. P. Campbell, 1990; J. P. Campbell, Gasser, & Oswald, 1996; J. P. Campbell, McCloy, Oppler, & Sager, 1993). One purpose of this study was to use new methods to determine whether there is a general factor in job performance ratings after controlling not only for halo error but also for rater leniency effects, random response error, and transient error (if any). A key point developed in this article is that the question of the existence of a real general factor in ratings cannot be addressed without controlling for all four of these sources of measurement error in ratings. The second purpose of this research was to estimate the size of this factor based on cumulative research findings in the literature. If a general factor does exist, it is important to know whether it is large or small. The question of whether there is a substantial general factor in ratings independent of halo error and other sources of measurement error has implications for the development of theories of job performance (which we discuss later). A third purpose of this research was to estimate the inflationary effects of halo error per se (net of other sources of measurement error) on construct-level correlations among job performance dimensions. As we discuss later, the methods developed and presented in this article allow isolation of the specific effect of halo error on construct-level correlations. To our knowledge, this has not been previously done. We also show why this can be done for construct-level correlations but not for observed correlations.

One basis for the hypothesis of a general factor is the argument that because the same abilities and traits (e.g., general cognitive ability and conscientiousness) likely contribute to performance on...
most or all dimensions of job performance, performance dimensions would be expected to be positively correlated, producing a general performance factor (Feldman, 1981; Motowidlo et al., 1997). Hulin (1982) also argued that because most human abilities are positively correlated, employees who have the ability to perform well on some job tasks are more likely to have the ability to perform well on other job tasks.

Landy, Vance, Barnes-Farrell, and Steele (1980) advocated that the problem of halo in ratings can be dealt with by partialing the general factor out of ratings. Harvey (1982), Hulin (1982), and Murphy (1982) have criticized this proposal on the theoretical grounds that doing this would partial out not only the effects of halo error but also any valid nonhalo general factor (see also Landy et al., 1982). If there is no general factor in ratings once halo error is controlled for, then this criticism is of no consequence. However, if a substantial general factor exists in ratings independent of halo error, that fact would support the position taken by Harvey (1982), Hulin (1982), and Murphy (1982) and would argue against the methodology recommended by Landy et al. (1980). Up to now, it has not been possible to subject this question to an empirical test. The data presented in this study allow for the first time such an empirical test.

Clarifying and Defining the Construct of Halo Error

Halo is often conceptualized as unrealistically large within-rater correlations between different performance dimensions; for example, an observed correlation of .85 between rated amount of output and rated personal appearance would be considered unrealistically large and therefore assumed to reflect halo. However, such apparently unrealistically large correlations are better thought of as consequences of halo error rather than halo itself. Halo is best conceptualized as a psychological process rather than an effect (Stanley, 1961; Thorndike, 1920; Wells, 1907). Conceptualizing halo as a psychological process rather than identifying it with one or more of its effects avoids controversies such as whether halo is defined as high within-rater interdimension correlations or as a Rater × Ratee interaction, or as low within-ratee interdimension variance. Instead, all three of these are viewed as outcomes or consequences of the halo psychological process. On the basis of the research literature, halo as a psychological process can be described as follows (Aiken, 1996; Guilford, 1954; Kiker & Motowidlo, 1998; Solomonson & Lance, 1997; Stanley, 1961; Viswesvaran, Ones, & Schmidt, 1996; Viswesvaran, Schmidt, & Ones, 2002). Before assigning ratings, each rater forms an overall impression of the merit or standing of each ratee. This overall impression in part overlaps the overall impressions of other raters and in part is unique to that one rater. The part of the overall impression that is in common with other raters is not halo; that part is considered true variance (Bartlett, 1983). The part that is unique to that rater—the idiosyncratic part—is halo error. This idiosyncratic part is the Rater × Ratee interaction. This interaction is directly analogous to the Subject × Item interactions in a test or other measurement scale, which is referred to as specific factor measurement error (Nunnally, 1978; Schmidt & Hunter, 1999) and is assigned to measurement error by (for example) coefficient alpha. In the context of ratings, a single rater is analogous to an item on a test. The use of several raters is analogous to use of a multiple-item test or scale. The idiosyncratic part of the overall impression influences the rater’s evaluation of the ratee on each rating dimension, increasing within-rater correlations between rating dimensions over what they would be in the absence of this halo process and, for any given rater–ratee combination, reduces the variability of ratings across the performance dimensions rated. Hence the psychological process underlying halo accounts for all these effects on observed ratings.

We need to stress that the idiosyncratic (halo) component of a rater’s overall impression is not the sum total of the rater’s overall evaluation of the ratee. If that were true, then there would be zero agreement between raters instead of the empirically obtained average value of approximately .50 (Rothstein, 1990; Viswesvaran et al., 1996); that is, in addition to the idiosyncratic component there is also a shared component of general evaluation. This shared component (shared with other raters) is not halo error but true variance, as illustrated in the structural models we present later.

Halo Error, True Halo, and Shared Halo: A Caveat

However, different arguments about the validity of this shared component can be presented. On one hand, one could argue that this shared component represents an actual general factor across job performance dimensions (Viswesvaran, 1993). On the other hand, one could argue that this shared component reflects nothing but a shared erroneous impression (shared halo error). The question is whether this shared component is construct valid as a measure of job performance (see also Murphy & DeShon, 2000; Schmidt, Viswesvaran, & Ones, 2000). This shared component may reflect a general factor (i.e., construct-valid job performance) as well as shared variance across raters that does not reflect job performance (i.e., construct-irrelevant variance; Hunter & Schmidt, 1996; Lance, Woehr, & Fiscarulo, 1991; Viswesvaran et al., 2002). However, in classical measurement theory, true score is defined as what is common across several hypothetical parallel measures (i.e., shared variance across items or raters), and all models of measurement error in ratings—including halo error—are based on classical measurement theory. Hence these models, including the one presented in this article, assume that the ratings component shared across raters is true variance.

We want to be clear here: The fact that the ratings component shared across raters is ratings–true variance does not imply that ratings–true scores correlate perfectly with “true performance.” True job performance is a hypothetical construct that could be less than perfectly correlated with ratings–true scores (Schmidt et al., 2000). Just what this correlation is is an empirical question, albeit one that is very difficult to research given the near impossibility of constructing a perfectly construct-valid measure of true job performance. Again, the focus of this article is on ratings–true scores, not on true job performance.

The concept of true halo has been used to refer to “the correlations among the actual behaviors being rated” (Solomonson & Lance, 1997, p. 665). This term is unfortunate. Because halo is by definition error, the term true halo should not be used to refer to the actual, nonhaloed relationships. A better term would be actual correlations. However, the term true halo has come into common usage in the literature. True halo does not refer to what the correlation between rated performance dimensions would be in the absence of halo error or other measurement errors in ratings, because (as noted above) even true scores on ratings might not be
perfect measures of actual performance. Rather, it refers to the unknown correlations that exist in the real world between performance dimensions, independent of ratings or other subjective judgments; for example, the real correlation between actual quantity and actual quality of output. In real world settings it is impossible to know these intercorrelations for job performance dimensions, because no perfect measures of the performance dimensions exist or can be created.

It is important to note that elimination of the effects of halo error and other measurement error on ratings as described above does not ensure that the resulting interdimension correlations are the true halo correlations. The resulting corrected correlations do estimate the rating dimension relationships unaffected by idiosyncratic halo error and other measurement errors. However, these correlations may be larger or smaller than the unknown actual correlations. For example, as noted earlier, the possibility of shared halo exists (R. J. Schneider, Hough, & Dunnette, 1996, p. 645); that is, the factor in the ratings that represents between-rater agreement (and that produces the average between-rater reliability of approximately .50) might be based in part on erroneous shared impressions. If so, even between-rater correlations corrected for measurement error as we describe later in this article could be inflated in comparison with the real world (nonperceptual) dimension intercorrelations. The literature currently contains conflicting empirical evidence about the existence of shared halo. However, even if it were shown to exist, shared halo would be, by definition, unrelated to the phenomenon of halo error in ratings (which is defined as uncorrelated across raters; Guilford, 1954; Stanley, 1961) and would have to be studied as a separate process in its own right. Conceptually, shared halo would be a form of criterion contamination (Brogden & Taylor, 1950), because it represents a process in which a variable irrelevant to actual job performance affects job performance ratings. Again, the focus of the present study is not shared halo.

The methods developed and used in this article make use of the fact that halo error (defined as a psychological process) affects the following in different ways: (a) within-rater reliabilities for each dimension, (b) between-rater reliabilities for each dimension, (c) within-rater interdimension correlations, and (d) between-rater interdimension correlations. These methods, and the concepts underlying them, are somewhat complex, but the presentation that follows is necessary to an understanding of how and why these methods work. We believe the result justifies the effort.

We begin below by presenting structural measurement models of ratings, based on classical measurement theory, that are appropriate for different ratings situations. We recognize that structural measurement models containing terms in addition to those contained in our models could be hypothesized (e.g., see Hoyt & Kerns, 1999). From a purely formal point of view, all such models would be compatible with classical measurement theory as a method. Classical measurement theory is capable of modeling any number of hypothesized effects on ratings (Schmidt et al., 2000). However, the effects we have included in our models are those that we know do exist from previous research on ratings. For example, we know there is a rater leniency effect, and we know there is a ratee effect, because many studies have empirically confirmed these effects. One can hypothesize additional effects beyond those contained in our models, but such effects must be confirmed empirically before being accepted as real. As we discuss later, an example of such a hypothesized effect is the rater-level effect (Mount, Judge, Scullen, Stymsma, & Hezlett, 1998; Viswesvaran et al., 2002), which hypothesizes that there is a construct difference in what is measured by supervisory ratings versus peer ratings versus subordinate ratings, based on the assumption that raters at different levels in the organization observe and emphasize different aspects of job performance, causing their ratings to assess different constructs. As we discuss later, this hypothesized rater effect has been disconfirmed by empirical research. By contrast, all the effects included in the measurement models presented below have been confirmed empirically. Furthermore, these models exclude no effects on ratings that have been empirically confirmed.

### Within-Rater Reliabilities for Performance Dimensions: Statistical Model

Within-rater reliabilities are assessed by having the same rater rate and then rereate a group of ratees with, say, a 2-week time interval intervening. As we describe later, coefficient alpha can also be used to assess within-rater reliability (based on within-rater correlations [or covariances] between different rated dimensions of job performance). Halo inflates within-rater reliabilities. This is most easily seen in the case of rate–rerate reliabilities: the rater’s idiosyncratic overall impression component affects ratings at both Time 1 and Time 2, whether the rating is of overall performance or some subdimension of performance. If the same rater rates a group of employees at Time 1 and Time 2, the structural model of the ratings is:

\[ Y_{ij}(1) = t_i + r_j + j \times i + e_{i(1)} \]

\[ Y_{ij}(2) = t_i + r_j + j \times i + e_{i(2)}, \]  

where \( Y_{ij}(1) \) is the rating given to ratee \( i \) by rater \( j \) at Time 1; likewise, \( Y_{ij}(2) \) is the same rating at Time 2; \( t_i \) = the ratee’s true score (ratee effect, conceptualized as his or her average rating over many different raters); \( r_j \) = the rater effect, representing the average degree of leniency or stringency for rater \( j \) in comparison to other raters (mean deviation score); \( j \times i = \) the Rater \( \times \) Ratee interaction (the halo effect; this value is different for each rater–ratee combination); and \( e_{i(1)} \) = random response error by rater \( j \) at Time 1; likewise, \( e_{i(2)} \) = random response error of rater \( j \) at Time 2.

In this model, the only term that differs between Time 1 and Time 2 is the random response error term. Hence, when the same rater rates a group of employees within a short intervening time interval (1 or 2 weeks), the rate–rerate correlation (actually, \( 1 - \) the rate–rerate correlation) assesses the amount of random response error (and transient error, if any; Schmidt & Hunter, 1996) in the ratings. (Transient measurement error is measurement error associated with occasions; that is, between-occasion variations in mental efficiency, emotion, mood, etc. Occasions are typically considered to be different days.) It is also clear that the idiosyncratic halo effect (the \( j \times i \) interaction) acts to increase (inflate) the rate–rerate reliability. This value differs across raters, and for any given ratee it is the same at Time 1 and Time 2. It should also be noted that, unlike the Rater \( \times \) Ratee interaction, the rater effect is a constant across rates in both Time 1 and Time 2 ratings and therefore has no effect on the rate–rerate correlation.
Between-Rater Reliabilities for Performance Dimensions: Statistical Model

Between-rater reliability is assessed as the correlation between two different raters rating the same ratees. In an ideal set of data, each rater would rate all the ratees; that is, raters and ratees would be fully crossed. However, such data sets are rare. Typically, ratees are nested within raters; that is, there will be a designated Rater 1 and a designated Rater 2, but within each such label, different raters are rated by different raters. This is illustrated in Table 1.

In Table 1, there are $N$ ratees. In a fully crossed design, Rater 1 would be the same supervisor for each ratee, as would Rater 2. However, interrater reliability data reported in the literature are different and are illustrated in Table 1, in which Ratee 1 is rated by Supervisor 1 ($S_1$) and Supervisor 4 ($S_4$), as is Ratee 2. However, for Rateee 3, Rater 1 is $S_2$ and Rater 2 is $S_6$. The only restriction here is that the same supervisor ($S_j$) cannot be both Rater 1 and Rater 2 for any ratee (because that would be a within-rater comparison). The interrater reliability estimate is computed as the correlation between Rater 1 and Rater 2. Almost all the interrater reliability estimates we have found in the literature are of this nature. In such cases, the rating received by any ratee is modeled as follows:

$$Y_{ij} = t_i + r_j + j \times i + e_{ij},$$  \hspace{1cm} (2)

where $Y_{ij}$ is the rating given to ratee $i$ by rater $j$; $t_i$ = the ratee’s true score rating (rater effect, conceptualized as his or her average rating over many different raters); $r_j$ = the rater effect, representing the average degree of leniency or stringency for rater $j$ in comparison to other raters (this is a constant for any one rater but varies across raters); $j \times i$ = the Rater $\times$ Ratee interaction caused by halo error (this is different for each $j \times i$ combination); and $e_{ij}$ = random response error on the part of rater $j$.

If there are only two raters, and each rater rates all ratees, the rater effect ($r_j$) has no effect on the correlation between the two raters, because $r_j$ is constant for any given rater across his or her ratings of ratees. However, when different raters rate different raters (as shown in Table 1), this is not the case, and differences between raters in the $r_j$ term reduce the interrater correlation. This occurs because within Rater 1 the $r_j$ term varies; it is not a constant, as it is when rates and raters are fully crossed. The same is true for Rater 2. Thus, with nested ratings, the interrater correlation is reduced by the last three terms in Equation 2. Therefore, interrater reliabilities of this sort are reduced not only by the halo effect (represented by the Rater $\times$ Ratee interaction, $j \times i$) and random response error ($e$), but also by rater effects ($r$).

Note. $S = supervisor.$

<table>
<thead>
<tr>
<th>Rates</th>
<th>Rater 1</th>
<th>Rater 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$S_1$</td>
<td>$S_4$</td>
</tr>
<tr>
<td>2</td>
<td>$S_1$</td>
<td>$S_4$</td>
</tr>
<tr>
<td>3</td>
<td>$S_2$</td>
<td>$S_6$</td>
</tr>
<tr>
<td>4</td>
<td>$S_2$</td>
<td>$S_6$</td>
</tr>
<tr>
<td>5</td>
<td>$S_3$</td>
<td>$S_4$</td>
</tr>
<tr>
<td>$N$</td>
<td>$S_k$</td>
<td>$S_1$</td>
</tr>
</tbody>
</table>

The magnitude of the variance due to random response error (and transient error, if any) can be assessed by rate–rater reliability. Viswesvaran et al. (1996, p. 565) found the average rate–rater reliability for the same rater to be .81, indicating that random response error (and transient error, if any) accounts for 19% of the variation on average. This meta-analysis found the average interrater reliability to be .52. Hence, we have $.81 - .52 = .29$, indicating that halo and rater effects (variability in leniency across raters) together account, on average, for 29% of the variation in total rating score by a single rater using a multiple-scale rating instrument. To our knowledge, there do not exist adequate data for job performance ratings to estimate how much of the 29% of total variance is due to halo effects and how much is due to rater effects.

As we show later, the results of the present study allow us to estimate these components separately—but only for estimated true score ratings, not observed ratings.

These variance percentages apply to supervisory ratings of overall performance; that is, they apply to ratings (by one supervisor) that are the sum (or average) of ratings made on multiple job performance dimensions. The figures for corresponding peer ratings are similar (Viswesvaran et al., 1996). As we discuss later, research evidence indicates that peer and supervisory ratings are equivalent at the construct level; that is, although they differ slightly in amount of measurement error, they are equivalent in the constructs assessed.

The halo component of ratings deflates the between-rater reliabilities for each dimension rated. The psychological process underlying halo error has the effect of decreasing correlations between raters; that is, it decreases between-rater reliability (Guilford, 1954; Stanley, 1961). This decrease stems from the fact that the component of overall impression that is idiosyncratic to each rater—that is, not shared by other raters; the Rater $\times$ Ratee interaction—creates halo error, thus lowering correlations between raters, both for individual performance dimensions and for overall evaluations of performance. This process explains why between-rater reliabilities are lower than within-rater reliabilities even if raters and ratees are fully crossed and rater effects do not act to lower interrater correlations.

Within-Rater Interdimension Correlations: Statistical Model

Halo inflates within-rater observed correlations between dimensions, because the idiosyncratic part of a rater’s overall impression (the halo error) affects ratings on all rating dimensions; that is, the Rater $\times$ Ratee interaction ($j \times i$) appears in the ratings of each separate dimension, inflating their intercorrelations. When the same rater rates a group of employees on two different performance dimensions, A and B, the structural model of the ratings is:

$$Y_{ij}(A) = t_i + r_j + j \times i + e_{ij},$$  \hspace{1cm} (3)

$$Y_{ij}(B) = t_i + r_j + j \times i + e_{ij}.$$
where \( Y_{ij}(A) \) = the rating given by rater \( j \) to ratee \( i \) on performance dimension \( A \); likewise, \( Y_{ij}(B) \) is the rating on performance dimension \( B \) by the same rater; \( t_{iA} \) = the ratee’s true score (ratee effect) on dimension \( A \); likewise, \( t_{jB} \) is the ratee effect on dimension \( B \); \( r_j \) = the rater effect, representing the average degree of leniency or stringency for rater \( j \) in comparison to other raters; \( j \times i \) = the Rater \( \times \) Ratee interaction (the halo error effect); and \( e_{ijA} \) = rater \( j \)’s random response error in rating dimension \( A \); \( e_{ijB} \) = the same quantity for dimension \( B \).

The correlation between \( Y_{ij}(A) \) and \( Y_{ij}(B) \) across raters is unaffected by the constant \( r_j \). It is reduced by the effects of the (uncorrelated) random response errors \( (e_{ijA} \text{ and } e_{ijB}) \), and the correlation is increased (inflated) by the Rater \( \times \) Ratee interaction \( (j \times i) \), because any given ratee receives the same \( j \times i \) value in each dimension rating. Use of within-rater reliabilities in the usual attenuation correction formulas (Nunnally, 1978) to correct the observed correlations for each dimension will remove (i.e., correct for) the effects of random response error, but not the effects of halo error. Thus, the resulting corrected correlations are inflated by correlated halo error.

**Between-Rater Interdimension Correlations**

Halo deflates the between-rater interdimension correlations in the same manner that it deflates between-rater reliabilities; that is, because halo error is idiosyncratic to individual raters, it is uncorrelated across raters, thus reducing interrater interdimension correlations below the value they would have in the absence of halo error. A rater’s halo error, by definition, affects his or her ratings of all performance dimensions, because halo error stems from the component of the overall impression of each ratee that is idiosyncratic or unique to that rater. Between-rater, between-dimension correlations are also biased downward by leniency effects (rater effects) in the same way that interrater reliabilities are reduced (assuming that raters and ratees are not fully crossed, which is almost always the case, as indicated earlier). Correcting between-rater, between-dimension correlations using the appropriate interrater reliabilities corrects for the biasing effects of halo error, rater leniency effects, random response error, and transient error (if any); that is, it corrects for all sources of measurement error in ratings, resulting in an unbiased estimate of the correlation of ratee effects (ratee true scores \( t_{iA} \) and \( t_{jB} \)) between performance dimensions \( A \) and \( B \).

**A Summary of Halo Effects**

The effects of halo can be summarized as follows. Between-rater reliabilities and between-rater interdimension correlations are reduced by halo error, whereas within-rater reliabilities and correlations are inflated by halo error; that is, the psychological process underlying halo error—idiosyncratic rater impressions of each ratee—is symmetrical: It inflates all within-rater correlations and deflates all between-rater correlations in comparison to values that would be observed in the absence of halo error.

Directly paralleling traditional measurement scales, the halo component in ratings is measurement error variance stemming from specific factor variance in the raters (Nunnally, 1978; Schmidt & Hunter, 1996, 1999); that is, the idiosyncratic bias in each rater’s ratings (Rater \( \times \) Ratee interactions) is analogous to specific factors in items (Examinee \( \times \) Item interactions) on a test. Such specific factors are considered measurement error. For example, coefficient alpha for personality or ability measures assigns such specific factors (Examinee \( \times \) Item interactions) to measurement error (Nunnally, 1978).

**The Proposed Framework**

A key question is whether there is a way to determine the correlations between rated dimensions unaffected by the inflating or deflating effect from halo error. Although observed between-rater interdimension correlations are deflated by halo error (as well as by rater effects and random response error), correlations that are corrected for between-rater reliability of each dimension are not. The deflationary effect of halo error is encoded in the between-rater reliabilities, which are reduced by halo error (as well as by rater effects and random response error), and correction for measurement error eliminates the deflationary effect of halo error on the dimension intercorrelations. This correction process corrects for rater leniency effects, random response error (and transient error, if any; Schmidt & Hunter, 1996, 1999), as well as halo error; that is, the corrected correlations estimate the between-rater interdimension correlations in the absence of any type of measurement error (halo error, rater effects, random response error, or transient error; Schmidt & Hunter, 1996). These, then, are the true score correlations in the absence of halo error and other measurement errors. From the point of view of theory development and understanding of the impact of halo error, these are the correlations of interest.

**Purposes of the Present Study**

As stated earlier, one purpose of this study was to estimate the inflationary effects of halo error on construct-level (true score) correlations. The inflationary effect of halo on interdimension correlations at the true score (i.e., ratee-effect) level is the difference between within-rater correlations corrected using within-rater reliabilities and between-rater correlations corrected using between-rater reliabilities (see Figure 1). The former are the correlations corrected for random response error (and transient error) but inflated by halo error; the latter are the correlations corrected for random response error (and transient error), halo error, and rater leniency effects. Hence, the difference is the inflationary effect of halo error at the true score (construct) level. Again, the true score level is the ratee effect level as defined in Equations 1, 2, and 3. It is also important to note that the inflationary effects of halo error on true score correlations is not a measure of the amount of halo in observed ratings. In fact, it is not possible to empirically assess this value in observed ratings.

In addition to comparing the average value of two types of correlation, one can examine the inflationary effect of halo error by comparing the size of the general factor extracted from the two types of correlation matrix. The general factor derived from corrected within-rater correlations among performance dimensions is inflated by halo error, whereas the general factor derived from corrected between-rater correlations is unaffected by halo error. The difference in the percentage variance accounted for by the general factor is the index of the effect of halo error.
A GENERAL FACTOR

1. Within-rater observed correlations; increased by halo error; reduced by random response and transient errors. (Upper triangular matrices, Tables 5 & 6)

2. Between-rater observed correlations; reduced by halo error, rater leniency effects, random response and transient errors. (Upper triangular matrix, Table 7)

3. Corrected using within-rater reliabilities; random and transient errors corrected, but corrected correlations still inflated by halo. (Lower triangular matrices, Tables 5 & 6)

4. Corrected using between-rater reliabilities; halo, rater leniency effects, random and transient errors are all corrected for. (Lower triangular matrix, Table 7)

5. Subtract matrix in 4 from matrix in 3 to estimate inflationary effects of halo error on within-rater true score correlations. Apply confirmatory factor analysis to matrices in 3 and 4 and compare the size of the general factor. The difference in size indicates the inflationary effect of halo error. The general factor found in the confirmatory factor analysis of the matrix in 4 estimates the size of the general factor in ratings controlling for halo error and other forms of measurement error.

Figure 1. Steps in the research process used to estimate the inflationary effects of halo error and the size of the general factor in job performance ratings, controlling for halo and other forms of measurement error.

Again, as stated earlier, another purpose of this study was to determine whether there is a general factor in job performance ratings after controlling for the effects of halo error, rater leniency effects, random response error (and transient error, if any) and, if there is, to estimate its size on the basis of cumulative research findings in the literature.

Method

Database

We used meta-analysis (Hunter & Schmidt, 1990b, 2004) to estimate the effects of halo on job performance ratings. We first describe the specification of the domain of the job performance construct and the dimensions that make up that construct. Given a definition of the dimensional components of the construct of job performance, within- and between-rater correlations between dimensions as well as between and within-rater reliabilities for each dimension can be tabulated from the research literature.

To specify the domain of job performance dimensions as comprehensively as possible, we draw a parallel to the lexical hypothesis in the personality domain (Goldberg, 1995). According to the lexical hypothesis, a comprehensive specification of personality traits (dimensions) can be obtained through an examination of all the interpersonal adjectives found in the language. Similarly, we argue that a comprehensive specification of the construct domain of rated job performance can be achieved by collating all rated job performance dimensions found in the research literature.

The various dimensions of job performance analyzed in this article were identified by grouping the measures obtained from published studies into conceptually similar dimensions (Viswesvaran et al., 1996). Viswesvaran et al. (1996) compiled a comprehensive list of job performance measures used in the extant literature, and two of the authors independently grouped the dimensions into conceptually similar groups. The initial overall agreement rate was 92.4%, and remaining differences were resolved through discussion until agreement was reached. The nine dimensions that resulted are shown in Table 2.

We note here that there were few studies that reported correlations between peer and supervisory ratings of communication competence. In fact, there were no studies reporting correlations between peer or supervisory ratings of communication competence and the other eight dimensions. As a result, we dropped that dimension from the analysis.

Ratings can be obtained from a number of sources (supervisors, peers, subordinates, customers). Of these, supervisors are the most commonly used source (H. J. Bernardin & Beatty, 1984; Cascio, 1991). However, we found that too few studies reported correlations between ratings by different supervisors for individual performance dimensions. However, correlations between ratings of performance dimensions by supervisors and ratings of the same dimensions by peers were fairly common, allowing use of these between-rater interdimension correlations. To the extent that there are theoretically relevant differences in the perspectives of supervisors and peers, interdimension correlations would be underestimated in our analyses. However, the evidence indicates that the same dimensions underlie supervisory and peer ratings, and agreement between supervisor and peer ratings is nearly as high as agreement within each group of raters (Mount et al., 1998; Viswesvaran et al., 2002). Viswesvaran et al. (2002) found that, corrected for measurement error, peer and supervisory ratings of different dimensions of job performance correlate essentially 1.00, indicating that the same constructs are being rated by the two types of raters. In a large data set, Mount et al. (1998) examined job performance ratings from four levels of raters: (a) supervisory ratings, (b) peer ratings, (c) subordinate ratings, and (d) self ratings. The model that best fit their data (see p. 567) was one in which there were halo factors for each individual rater (which they called “methods factors”) and three “skills factors” (job dimension factors) but no level factor; that is, the model that best fit their data is one that assumed the construct-level equivalence of supervisory, peer, subordinate, and self-ratings. Hence their findings, like those of Viswesvaran et al. (2002), support the use of peer ratings as equivalent to supervisory ratings at the construct level. Perhaps the strongest evidence that ratings by supervisors and peers represent the same constructs comes from Facteau and Craig’s (2001) study. Using a large database from a corporate 360 feedback program, they examined the construct equivalence of ratings by supervisors, peers, subordinates, and the self for each of seven rating scales. Their analysis was based on both confirmatory factor analysis (CFA) and analysis of differential item function and differential scale functioning derived from item response theory (IRT). The CFA analysis indicated that all the scales were invariant across rating groups. For each rater group, the same seven factors were found to underlie the raters’ responses to the items on the seven scales. In the IRT-based analysis, evidence of differential item and scale functioning was limited to effects of 3 of the 36 items and was described as trivial in magnitude. Facteau and Craig (2001) concluded that their seven-scale rating instrument showed measurement invariance across all four rater groups. A study conducted by Maurer, Raju, and Collins (1998) also is relevant. Using the same CFA and IRT methods used by Facteau and Craig (2001), they found that peer and subordinate ratings of managers’ team-building skills assessed the same underlying construct. Although this study did not include ratings by supervisors or managers, the findings provide support for the same sort of general equivalence of ratings across all rater groups found by Facteau and Craig (2001).

Thus, the first step in this research was to cumulate the literature using psychometric meta-analyses to establish both between- and within-rater intercorrelations across dimensions of job performance. Within-rater correlations were cumulated separately for supervisory ratings and peer rat-
Table 2

<table>
<thead>
<tr>
<th>Dimension rated</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpersonal competence</td>
<td>Ratings of ability to work well with others. Ratings of individuals on cooperation with others, customer relations, working with coworkers, acceptance by others, as well as nominations for being easy to get along with.</td>
</tr>
<tr>
<td>Administrative competence</td>
<td>Ratings of proficiency in handling the coordination among different roles in an organization; proficiency in organizing and scheduling work periods, administrative maintenance of records (note, though, that clarity will go to communication competence), ability to place and assign subordinates, and having a clear idea about the job duties and responsibilities of others.</td>
</tr>
<tr>
<td>Quality</td>
<td>Ratings of statements referring to the quality of tasks completed, lack of errors, accuracy of specifications, thoroughness, and amount of wastage.</td>
</tr>
<tr>
<td>Productivity</td>
<td>Ratings of the quantity of work produced. Ratings are based on productivity or sales. Examples include ratings of the number of accounts opened by bank tellers and ratings of number of transactions completed by salesclerks.</td>
</tr>
<tr>
<td>Effort</td>
<td>Ratings of amount of work expended in striving to do a good job. Initiative, attention to duty, alertness, resourcefulness, enthusiasm about work, industriousness, earnestness at work, persistence in seeking goals, dedication, personal involvement in the job, effort and energy expended on the job.</td>
</tr>
<tr>
<td>Job knowledge</td>
<td>Ratings of the individual’s on-job knowledge, keeping up to date, as well as nominations of who knows the job best and nominations of who keeps up to date, and so on.</td>
</tr>
<tr>
<td>Leadership</td>
<td>Ratings of ability to inspire, to elicit high performance from others, to motivate others. Illustrative examples include performance appraisal statements such as “gets subordinates to work efficiently,” “stimulates subordinates effectively,” and “maintains authority easily and comfortably.”</td>
</tr>
<tr>
<td>Compliance/acceptance of authority</td>
<td>Ratings on general perspective about rules and regulations. Obeying rules, conforming to regulations in the workplace, a positive attitude toward supervision, conforming to organizational norms and culture, not complaining of organizational policies, and following instructions.</td>
</tr>
<tr>
<td>Communication competence</td>
<td>Ratings of skill in gathering and transmitting information (both in oral and written format). The proficiency to articulate, either in written or oral format, information/views/opinions/positions. Illustrative examples include performance appraisal statements such as “very good in making reports,” “reports are clear,” “reports are unambiguous,” and “reports need no further clarification.”</td>
</tr>
</tbody>
</table>

Table 2: Definition of Job Performance Rating

In searching the literature, we used both electronic and manual search strategies. We searched the PsycLIT database for the period 1974 through 1999, and the same search was carried out in Wils, Infotrac, Oasis, and Central Libraries Research Council databases. However, the most fruitful search strategy was the manual search strategy. We searched the major journals for articles that reported the correlation between any two dimensions of job performance. We searched all issues from journal inception up to February 1999 of the following journals: Journal of Applied Psychology, Personnel Psychology, Academy of Management Journal, Human Relations, Journal of Business and Psychology, Journal of Management, Accident Analysis and Prevention, International Journal of Intercultural Relations, Organizational Behavior and Human Decision Processes, Journal of Vocational Behavior, Journal of Applied Behavioral Analysis, Human Resources Management Research, Journal of Occupational Psychology, Psychological Reports, and Journal of Organizational Behavior.

Again, all between-rater correlations were correlations between ratings by supervisors and ratings by peers.

Studies that reported correlations between the different rated dimensions of job performance were located. To be included in the database, a study had to be from a published source; conference presentations and unpublished data were not included. We excluded unpublished studies only for feasibility of data management, in light of the vastness of the relevant research literature. Hence, we did not attempt to include dissertations, books, and unpublished reports in the database. However, in the case of correlations among performance measures, there would appear to be no reason to expect reporting or publication bias, because these correlations were almost never the focus of the research. Rather, they were secondary by-products of the research. Hence, we would expect published correlations to be an unbiased sample of all correlations.

In searching the literature, we used both electronic and manual search strategies. We searched the PsycLIT database for the period 1974 through 1999, and the same search was carried out in Wils, Infotrac, Oasis, and Central Libraries Research Council databases. However, the most fruitful search strategy was the manual search strategy. We searched the major journals for articles that reported the correlation between any two dimensions of job performance. We searched all issues from journal inception up to February 1999 of the following journals: Journal of Applied Psychology, Personnel Psychology, Academy of Management Journal, Human Relations, Journal of Business and Psychology, Journal of Management, Accident Analysis and Prevention, International Journal of Intercultural Relations, Organizational Behavior and Human Decision Processes, Journal of Vocational Behavior, Journal of Applied Behavioral Analysis, Human Resources Management Research, Journal of Occupational Psychology, Psychological Reports, and Journal of Organizational Behavior.

We included in the database studies that report data on individuals regarding their rated performance on actual jobs. Studies reporting group-level data on departments, productions units, or teams were excluded. Experimental simulations reporting data on student samples engaged in laboratory work (or ratings of videotaped and manipulated work samples) were also excluded. We also omitted assessment center ratings as well as interviewer and recruiter ratings, because they represent ratings on job performance predictors rather than on job performance itself. Finally, the database also includes studies that report only between- and within-rater
reliability estimates for the different dimensions of job performance. We compiled the reliability values separately for supervisor and peer ratings to construct reliability distributions (described below) that were used in the psychometric meta-analyses. A total of 303 articles were located and coded.

If a study reported data on a total sample as well as on subsamples (e.g., Blacks and Whites, or males and females), we included the data only from the total sample, because subgroup correlations contain more sampling error. Also, averaging subgroup correlations can introduce a downward bias in the correlation (due to range restriction) as compared to the correlation computed on the total sample (Hunter & Schmidt, 1990b, p. 465). When studies reported correlations between measures of the same dimensions with different labels or wordings of the same dimension along with correlations with other dimensions of job performance, we formed a composite measure by summing over the measures of different wordings of the same dimension. We used unit (equal) weights for the different measures of the same dimension in computing the composite correlations (Harman, 1976; Nunnally, 1978). The correlations of this sum with other dimensions of job performance were computed, and these computed correlations were used in the meta-analyses. We also computed the Mosier reliability1 of the composite (Hunter & Schmidt, 1990b, pp. 461–463), separately for intra- and interrater reliability.

Meta-Analyses

We used the meta-analysis methods of Hunter and Schmidt (1990b, 2004). In all, we cumulated three types of interdimension correlations: (a) within-rater correlations obtained from the same supervisor, (b) within-rater correlations obtained from the same peer rater, and (c) between-rater correlations in which one rater was a supervisor and the other was a peer rater. Correlations computed on artificially dichotomized continuous scales or variables were corrected for downward bias induced by dichotomization and their sample sizes were appropriately adjusted downward to allow for the increased sampling error introduced by the correction (Hunter & Schmidt, 1990a; Ones, Viswesvaran, & Schmidt, 1993; Schmidt et al., 1993).

We computed the mean observed within-rater correlations among the eight performance dimensions separately for supervisors and peers. These are the observed within-rater interdimension correlations. Thus, we constructed two 8 \times 8 matrices (one for supervisors and one for peers). As described earlier, the correlations in these two matrices are inflated by halo and deflated by random response error. For each pair of job performance Dimensions A and B, we corrected the within-rater correlations between Dimension A and Dimension B rated by supervisors using average within-rater reliability estimates for supervisory ratings of Dimension A and average within-rater reliability estimates for supervisory ratings of Dimension B. Few within-rater reliabilities in the form of rate–rate correlations for the same rater were available in the literature for specific performance dimensions. (They are available for composite overall job performance ratings. Viswesvaran et al., 1996, reported an unweighted average value of .81 for composite ratings; the sample size weighted average was .84.) However, many estimates of within-rater coefficient alpha were available for individual performance dimensions. Within-rater coefficient alpha estimates rate–rate reliability (Cronbach, 1947, 1951; Schmidt & Hunter, 1996), and so we used it here.2 We conducted similar analyses for within-rater interdimension correlations for peers (using within-rater reliabilities for peers). These corrections resulted in two 8 \times 8 matrices (one for supervisors and one for peers) in which the cells are the within-rater interdimension correlations corrected for random response error but not for halo error.

Next, we computed the between-rater interdimension correlations. To estimate each between-rater interdimension correlation, we conducted two separate meta-analyses. The first meta-analysis cumulated the correlations between peer ratings of Dimension A with supervisor ratings of Dimension B. The second meta-analysis cumulated the correlation between peer ratings of Dimension B with supervisor ratings of Dimension A. The two mean observed correlations were averaged (sample size weighted) across the two meta-analyses to provide the final estimate of the observed correlation between Dimension A and Dimension B.

We carried out a similar averaging process for the estimated true score correlations to provide the final estimate of the true score correlation between Dimensions A and B. The observed correlation between peer ratings of Dimension A with supervisor ratings of Dimension B was corrected with the intersupervisor reliability for Dimension B and interpeer reliability for Dimension A. A parallel correction was made for supervisory ratings of Dimension A and peer ratings of Dimension B. The two corrected correlations were averaged. Note that although the original correlation matrix here is nonsymmetrical, this averaging process yields a symmetric correlation matrix. This process produced a single between-rater intercorrelation matrix of observed correlations between the different dimensions of job performance. Likewise, it produced (after corrections) a single between-rater matrix of estimated true score intercorrelations among the different job performance dimensions.

We estimated the inflationary effect of halo on dimension intercorrelations at the true score (construct) level separately for supervisory and peer ratings. To estimate this effect for supervisory ratings, we subtracted the

---

1 An alternative to the Mosier reliability estimate for the composite is the standardized coefficient alpha, which is the same as use of the Spearman–Brown formula based on the average intercorrelation between the different wordings of the same dimension. Both the standardized coefficient alpha and the Spearman–Brown formula assign the specific factor variance (specific but reliable to each measure forming the composite) to measurement error, whereas in the Mosier reliability estimate the specific factor variance is treated as true variance. Whether the specific factor variance should be treated as error or true variance is a theoretical question. The use of the Mosier formula (instead of Spearman–Brown or the standardized coefficient alpha) results in higher estimates of the reliability (because the specific factor variance is treated as true variance), which lowers the corrected correlation and provides a more conservative estimate of the size of any general factor at the construct level.

2 A reviewer raised the question of how coefficient alpha is computed in these cases. The individual (single) rater rates a number of rates (say, 25 rates) on each of a number of statements that describe that dimension of job performance (e.g., 15 statements that relate to the employee’s output/productivity). The resulting Rates \times Statements (items) data matrix is the basis for computing alpha. If computed on the intercorrelations among the statements (items), the result is the standardized coefficient alpha, which weights all statements (items) equally. If computed on the variance–covariance matrix of the items, the result is the raw score alpha, which weights each item by its standard deviation. Raw scores and standardized alphas are typically almost identical, because item standard deviations typically do not differ much and because there are usually more than a few rating items. (The larger the number of variables, the smaller is the effect of differential weighting.) Note that for any given performance dimension it is not necessary that different studies use exactly the same items (or subdimension elements) or item wordings. The reviewer also questioned whether the data matrix can be expected to be unidimensional, contending that if it were not, then alpha would underestimate reliability. In this case the data matrices are almost certainly unidimensional (or nearly so), because all items refer to the same dimension of job performance. However, unidimensionality is not a requirement for use of alpha in this manner. If the data are not fully unidimensional, alpha underestimates only the strict parallel forms reliability of a scale (and typically by only small amounts, rarely exceeding .02; Nunnally & Bernstein, 1994). It does not underestimate random parallel forms reliability, which is often the more appropriate measurement model (Nunnally & Bernstein, 1994).
corrected between-rater correlation matrix from the corrected within-rater correlation matrix for supervisors (element by corresponding matrix element). These two matrices differ only in that correlations in the within-rater matrix are inflated by halo error, whereas those in the between-rater matrix are unaffected by halo error in either direction. The differences between those two matrices reveal the general inflationary effect of halo error on construct-level correlations, independent of random response error and transient error. The average value in this difference matrix is the average inflationary effect of halo error. We conducted this same analysis for peer ratings. To our knowledge, no such analysis has previously been conducted.

Detecting and Calibrating the Performance General Factor

We subjected the intercorrelation matrices obtained from the psychometric meta-analyses to CFA (Viswesvaran & Ones, 1995). The CFA hypothesized a single general factor. We used analyses based on the corrected between-rater matrices to assess the actual size of the general factor in ratings independent of halo error. CFA results for the other matrices were used for comparison purposes. We conducted a CFA on the three observed (uncorrected) correlation matrices using LISREL 8.5.2 (student edition) to assess the size of the general factor as well as the loadings of the eight dimensions on this general factor. Maximum likelihood (ML) estimation procedures were used. However, for the three reliability-corrected correlation matrices, because of multicollinearity, the ML estimates were distorted as the correlation matrix was positive definite. Rather than using the ridge option to obtain the ML estimate of the general factor as well as factor loadings, we used the unweighted least squares analyses with these three reliability-corrected correlation matrices. As noted earlier, the major focus of these analyses was on the size of any general factor, expressed as percentage of total variance accounted for (and the square root of that figure; see Discussion section).

The extent to which the estimated general factor loadings reproduce the intercorrelation matrix between the various dimensions of job performance reflects the fit of a model postulating only a single general factor (Hunter & Gerbing, 1982). The root-mean-square residual (RMSR) and the normed fit index (NFI) were used as indices of the extent to which such a single general factor model fit the empirical data. Following convention, RMSR values less than .10 and NFI of .90 or larger were taken as indicating good general factor model fit the empirical data. Following convention, RMSR and the normed fit index (NFI) were used as indices of the extent to which such a single general factor model fit the empirical data. Following convention, RMSR values less than .10 and NFI of .90 or larger were taken as indicating good general factor model fit (Hunter & Gerbing, 1982).

Note that the focus in these CFAs is on estimating the inflation in the size of the general factor caused by halo as well as the underestimation caused by unreliability. Therefore, we do not place great emphasis on the RMSR or NFI per se. The exact factor structure of job performance is a different question from the question of how the magnitude of the general factor is affected by halo and other measurement error. Smaller group factors may remain after the extraction of a general factor. Thus, it is not crucial for this research that the model postulating only a single general factor show excellent fit to the data. Excellent fit might require identification of smaller group factors in addition to the general factor.) The critical question is whether there is a substantial general factor and, if so, how large it is. Even given a general factor, there may still be group factors in addition to the general factor, and hence the general factor alone may not fully account for the correlations among the performance dimensions. Such smaller group factors (if any) are not the focus of the present study.

Results

Analysis of Halo Error

Mean between-rater reliabilities for ratings are summarized in Table 3. Also provided are the observed (uncorrected) standard deviations of the reliability estimates. The estimates reported in Table 3 are the weighted means. (In this data set, the sample size weighted mean reliabilities [not shown] are slightly lower than the weighted values used in the meta-analytic cumulations here; cf. Viswesvaran et al., 1996. This may be due to a publication bias favoring the publication of small-sample studies that report higher reliabilities due to positive sampling errors; that is, it is possible that, among small-sample studies, reliability estimates with positive sampling error were more likely to be published than reliability values with negative sampling error.) Except for relatively low reliability for peer ratings on three job performance dimensions, the values reported in Table 3 are consistent with previous meta-analytic results for supervisory ratings (e.g., Rothstein, 1990).

Table 4 shows, for each of the eight performance dimensions, the weighted average within-rater reliability and its observed standard deviation. As expected, these values are larger than the between-rater reliabilities in Table 3. In addition, 12 of the 16 values are lower than the mean rate–rerate same rater reliability reported by Viswesvaran et al. (1996) for composite (summed) ratings of job performance (.81). This is as expected given the fact that individual performance dimensions are expected to be less reliable than an overall composite of multiple rating dimensions.

The within-rater correlations for supervisors are shown in Table 5. Observed rs (M = .55) are shown above the diagonal, and estimated true score rs (M = .72) are shown below the diagonal. Reliabilities used for these corrections are the supervisory within-rater reliabilities shown in Table 4. Comparable within-rater correlations for peer ratings are shown in Table 6. For peer ratings, the average observed correlation is .58, and the average true score correlation is .88. Again, the reliabilities used to make the corrections are the corresponding within-rater reliabilities for peer ratings shown in Table 4. Values averaged across supervisory and peer ratings are .57 for observed correlations and .80 for corrected correlations.

Table 7 shows the correlations between peer and supervisor ratings for the eight performance dimensions. As described earlier, this matrix was produced by averaging corresponding elements in the original nonsymmetric matrix. For example, the correlation between supervisory ratings of quality and peer ratings of effort was averaged with the correlation between peer ratings of quality and supervisory ratings of effort. As expected, these between-rater
correlations are considerably smaller than the within-rater correlations in Tables 5 and 6. Observed correlations (M = .27) are shown above the diagonal, and true score correlations (M = .54) are below the diagonal. The reliabilities used to make these corrections are the between-rater reliabilities presented in Table 3. We used the appropriate between-rater reliability for both supervisor and peer ratings to correct each correlation, because the ratings going into each correlation came from both sources.

CFA Results

The CFA findings related to the question of a possible general factor in ratings independent of halo error are presented in Tables 8, 9, and 10. Table 11 contains a summary of these analyses.

The factor loadings for supervisory ratings for the within-rater correlations from Table 5 are presented in Table 8. These corresponding results for the peer ratings within-rater correlations from Table 6 are shown in Table 9. For both supervisory ratings (Table 8) and peer ratings (Table 9), all job performance dimensions load substantially on the general factor. Of the 16 loadings, the smallest is .66 (for ratings of productivity by supervisors). As indicated earlier, RMSRs of .10 or smaller indicate excellent fit of the one-factor model. As can be seen in Table 11, all RMSRs except one are somewhat smaller than this figure, indicating acceptable fit of the general factor model. Likewise, all NFI values except one are .90 or larger (and the single exception is .891). The general factor is quite large in the data in Tables 8 and 9. As can be seen in Table 11, at the level of observed correlations the general factor accounts for 55.8% of the variance of supervisory ratings and 58.4% of the variance of peer ratings. At the true score levels, these figures rise to 75.8% and 88.8%, respectively.

However, these figures serve for comparison purposes only, because the size of the general factor in within-rater data is inflated by halo error. The within-rater general factor does not reflect the actual (nonartifactual) size of the general factor. The size of the general factor is smaller for between-rater data. The factor loadings for the between-rater data (Table 7 correlations) are presented in Table 10. Although the general level of loadings is lower than in the case of the within-rater correlations (as expected), all loadings remain substantial. The RMSR is .050 for the observed and .228 for the corrected between-rater correlations, respectively. The NFI is .921 for the uncorrected correlations and .869 for the corrected values. The fit index values for the corrected correlations appear to be large enough to indicate the likely existence of additional, smaller factors beyond the general factor (which future research might examine). However, this possibility does not argue against the existence of a general factor, as noted earlier. For the uncorrected between-rater correlations, the general factor accounts for 27.4% of the variance. However, the size of this general factor is substantially reduced by measurement error (i.e., by rater effects; halo error; random response error; and possibly transient error, if any exists here). Correcting between-rater correlations using the appropriate between-rater reliabilities corrects for all four of these sources of measurement error. For the resulting

Table 4
Within-Rater Reliabilities: Frequency Weighted Means and Standard Deviations

<table>
<thead>
<tr>
<th>Job performance dimension</th>
<th>Supervisor ratings</th>
<th>Peer ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>M</td>
</tr>
<tr>
<td>Interpersonal competence</td>
<td>10,955</td>
<td>.75</td>
</tr>
<tr>
<td>Administrative competence</td>
<td>4,754</td>
<td>.79</td>
</tr>
<tr>
<td>Quality</td>
<td>739</td>
<td>.81</td>
</tr>
<tr>
<td>Productivity</td>
<td>2,697</td>
<td>.85</td>
</tr>
<tr>
<td>Effort</td>
<td>3,112</td>
<td>.75</td>
</tr>
<tr>
<td>Job knowledge</td>
<td>959</td>
<td>.77</td>
</tr>
<tr>
<td>Leadership</td>
<td>3,821</td>
<td>.77</td>
</tr>
<tr>
<td>Compliance/acceptance of authority</td>
<td>3,438</td>
<td>.76</td>
</tr>
</tbody>
</table>

Note. N = total sample size across all estimates that were averaged. All values are within-rater coefficient alpha values computed based on within-rater, between-dimension correlations.

Table 5
Within-Rater Correlations for Supervisory Ratings

<table>
<thead>
<tr>
<th>Dimension</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Interpersonal competence</td>
<td>—</td>
<td>.47</td>
<td>.48</td>
<td>.49</td>
<td>.59</td>
<td>.56</td>
<td>.62</td>
<td>.55</td>
</tr>
<tr>
<td>2. Administrative competence</td>
<td>.62</td>
<td>—</td>
<td>.55</td>
<td>.39</td>
<td>.57</td>
<td>.64</td>
<td>.70</td>
<td>.39</td>
</tr>
<tr>
<td></td>
<td>(7,680)</td>
<td></td>
<td>(3,456)</td>
<td></td>
<td>(8,530)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Quality</td>
<td>.63</td>
<td>.69</td>
<td>—</td>
<td>.56</td>
<td>.62</td>
<td>.54</td>
<td>.50</td>
<td>.47</td>
</tr>
<tr>
<td></td>
<td>(11,153)</td>
<td></td>
<td>(4,121)</td>
<td></td>
<td>(4,791)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Productivity</td>
<td>.63</td>
<td>.47</td>
<td>.68</td>
<td>—</td>
<td>.61</td>
<td>.37</td>
<td>.48</td>
<td>.61</td>
</tr>
<tr>
<td></td>
<td>(10,753)</td>
<td></td>
<td>(8,602)</td>
<td></td>
<td>(10,963)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Effort</td>
<td>.81</td>
<td>.74</td>
<td>.81</td>
<td>.77</td>
<td>—</td>
<td>.59</td>
<td>.67</td>
<td>.59</td>
</tr>
<tr>
<td></td>
<td>(7,002)</td>
<td></td>
<td>(4,725)</td>
<td></td>
<td>(1,905)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Job knowledge</td>
<td>.76</td>
<td>.82</td>
<td>.69</td>
<td>.46</td>
<td>.79</td>
<td>—</td>
<td>.63</td>
<td>.54</td>
</tr>
<tr>
<td></td>
<td>(11,142)</td>
<td></td>
<td>(6,492)</td>
<td></td>
<td>(1,785)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Leadership</td>
<td>.84</td>
<td>.91</td>
<td>.64</td>
<td>.60</td>
<td>.90</td>
<td>.83</td>
<td>—</td>
<td>.67</td>
</tr>
<tr>
<td></td>
<td>(11,590)</td>
<td></td>
<td>(4,129)</td>
<td></td>
<td>(4,740)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Compliance/acceptance of authority</td>
<td>.75</td>
<td>.51</td>
<td>.60</td>
<td>.72</td>
<td>.80</td>
<td>.72</td>
<td>.90</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>(11,590)</td>
<td></td>
<td>(4,129)</td>
<td></td>
<td>(4,740)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Sample size weighted mean observed correlations corrected for within-rater unreliability are below the diagonal; uncorrected correlations are above the diagonal. Reliabilities used in corrections are described in Table 4. Numbers in parentheses are the total sample size across the correlations averaged.
construct-level correlations, the general factor accounts for 60.3% of the variance, indicating that at the level of the performance constructs underlying job performance ratings, and after the effects of halo and other measurement errors have been removed, a general factor accounts for more than half the variation in job performance ratings. As shown in the lower part of Table 11, the three forms of measurement error considered together reduce the general factor from 60.3% to 27.4% of the total variance, a 55% reduction.

**Discussion**

**Role of Halo Error**

One purpose of this study was to estimate the extent to which halo error inflates true score correlations between job performance dimensions. As noted above, the average true score correlation between performance dimensions controlling for the inflationary effects of halo error is .54. When the effects of halo error are not controlled, this value is .72 for supervisory ratings, yielding an

---

**Table 6**

**Within-Rater Correlations for Peer Ratings**

<table>
<thead>
<tr>
<th>Dimension</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Interpersonal competence</td>
<td>—</td>
<td>.52</td>
<td>.41</td>
<td>.62</td>
<td>.56</td>
<td>.53</td>
<td>.47</td>
<td>.63</td>
</tr>
<tr>
<td>2. Administrative competence</td>
<td>.77</td>
<td>—</td>
<td>.69</td>
<td>.57</td>
<td>.54</td>
<td>.75</td>
<td>.48</td>
<td>.58</td>
</tr>
<tr>
<td></td>
<td>(1,867)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Quality</td>
<td>.57</td>
<td>.85</td>
<td>—</td>
<td>.65</td>
<td>.64</td>
<td>.59</td>
<td>.47</td>
<td>.64</td>
</tr>
<tr>
<td></td>
<td>(320)</td>
<td>(248)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Productivity</td>
<td>.99</td>
<td>.91</td>
<td>.95</td>
<td>—</td>
<td>.60</td>
<td>.49</td>
<td>.65</td>
<td>.63</td>
</tr>
<tr>
<td></td>
<td>(1,708)</td>
<td>(437)</td>
<td>(341)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Effort</td>
<td>.89</td>
<td>.76</td>
<td>.82</td>
<td>.99</td>
<td>—</td>
<td>.52</td>
<td>.73</td>
<td>.57</td>
</tr>
<tr>
<td></td>
<td>(1,432)</td>
<td>(592)</td>
<td>(274)</td>
<td>(248)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1,942)</td>
<td>(782)</td>
<td>(134)</td>
<td>(108)</td>
<td>(527)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1,170)</td>
<td>(924)</td>
<td>(159)</td>
<td>(434)</td>
<td>(431)</td>
<td>(833)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Compliance/acceptance of</td>
<td>.91</td>
<td>.74</td>
<td>.75</td>
<td>.95</td>
<td>.75</td>
<td>corr.</td>
<td>corr.</td>
<td>—</td>
</tr>
<tr>
<td>authority</td>
<td>(939)</td>
<td>(161)</td>
<td>(161)</td>
<td>(161)</td>
<td>(167)</td>
<td>avg.</td>
<td>avg.</td>
<td>avg.</td>
</tr>
</tbody>
</table>

**Note.** Sample size weighted mean observed correlations corrected for within-rater unreliability are below the diagonal; uncorrected correlations are above the diagonal. Reliabilities used in corrections are described in Table 4. Numbers in parentheses are the total sample size across the correlations averaged. Avg. = The average of all the observed correlations was imputed; corr. avg. = The average observed correlation corrected for the unreliability of the measures correlated was imputed.

---

**Table 7**

**Between-Rater Correlations Between Supervisory and Peer Ratings**

<table>
<thead>
<tr>
<th>Dimension</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Interpersonal competence</td>
<td>—</td>
<td>.22</td>
<td>.10</td>
<td>.29</td>
<td>.24</td>
<td>.17</td>
<td>.21</td>
<td>.33</td>
</tr>
<tr>
<td></td>
<td>(1,158)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Administrative competence</td>
<td>.38</td>
<td>—</td>
<td>.22</td>
<td>.34</td>
<td>.31</td>
<td>.23</td>
<td>.20</td>
<td>.29</td>
</tr>
<tr>
<td></td>
<td>(322)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Quality</td>
<td>.18</td>
<td>.34</td>
<td>—</td>
<td>.27</td>
<td>.30</td>
<td>.38</td>
<td>.29</td>
<td>.18</td>
</tr>
<tr>
<td></td>
<td>(322)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Productivity</td>
<td>.60</td>
<td>.68</td>
<td>.51</td>
<td>—</td>
<td>.28</td>
<td>.46</td>
<td>.26</td>
<td>.31</td>
</tr>
<tr>
<td></td>
<td>(2,849)</td>
<td>(580)</td>
<td>(322)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Effort</td>
<td>.49</td>
<td>.54</td>
<td>.54</td>
<td>.65</td>
<td>—</td>
<td>.22</td>
<td>.29</td>
<td>.34</td>
</tr>
<tr>
<td></td>
<td>(1,358)</td>
<td>(376)</td>
<td>(548)</td>
<td>(376)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Job knowledge</td>
<td>.35</td>
<td>.46</td>
<td>.66</td>
<td>.99</td>
<td>.49</td>
<td>—</td>
<td>.34</td>
<td>.29</td>
</tr>
<tr>
<td></td>
<td>(1,399)</td>
<td>(956)</td>
<td>(113)</td>
<td>(174)</td>
<td>(578)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Leadership</td>
<td>.44</td>
<td>.44</td>
<td>.70</td>
<td>.60</td>
<td>.52</td>
<td>.79</td>
<td>—</td>
<td>.22</td>
</tr>
<tr>
<td></td>
<td>(802)</td>
<td>(174)</td>
<td>(352)</td>
<td>(430)</td>
<td>(174)</td>
<td>(319)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Compliance/acceptance of</td>
<td>.57</td>
<td>.43</td>
<td>.27</td>
<td>.61</td>
<td>.60</td>
<td>.70</td>
<td>.48</td>
<td>—</td>
</tr>
<tr>
<td>authority</td>
<td>(1,536)</td>
<td>(322)</td>
<td>(322)</td>
<td>(1,006)</td>
<td>(361)</td>
<td>(435)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note.** Sample size weighted mean observed correlations corrected for between-rater reliability are below the diagonal; uncorrected correlations are above the diagonal. Between-rater reliability values used in corrections are described in Table 3. Numbers in parentheses are the total sample size across the correlations averaged. The corrected correlations in this table are true score correlations free of halo error. As explained in the text, a symmetric correlation matrix was produced by averaging corresponding cells; for example, $r_{s3p5}$ and $r_{p3s5}$ were averaged, where $s =$ quality, $p =$ effort, $s =$ supervisor rating, and $p =$ peer rating.
The inflationary effect of halo is therefore .888 – .603 = .285, or a 47% inflation in percentage variance accounted for. The corresponding figures for supervisory ratings are: .758 – .603 = .155, or a 25% inflation. Hence we reach the same conclusion: Halo error substantially inflates both types of within-rater true score correlations but considerably more so for peer ratings than for supervisory ratings.

As has been noted in the literature (e.g., Hunter & Schmidt, 1990b, pp. 199–202; Rosenthal & Rubin, 1979, 1982), interpretations of data in terms of percentage variance accounted for are often deceptive. Such interpretations often underestimate the importance of the relationship in question, making it desirable to interpret statistical relationships in other ways. Consider our finding that the general factor accounts for 60% of the variance of ratings at the true score level. This finding means that the average correlation between true scores on individual dimensions and the general factor is .77 (i.e., \( \sqrt{.60} = .77 \)). This can also be seen in the second column of Table 10, which gives the loadings of dimensions on the general factor; the loadings range from .63 to .92, with an average of .77. This means that at the level of actual (true score) job performance constructs, individual performance dimension constructs are highly correlated with the general factor in job performance.

The correlation of .77 places constraints on the prospects for discriminant validity—at least for high discriminant validity—for ratings but considerably more so for peer ratings than for supervisory ratings.

At the construct level, both inflationary effects of halo error (on the mean intercorrelation and on the size of the general factor) were large, as expected. The within-rater general factor accounted for 76% and 89% of the total variance of supervisory and peer ratings, respectively (see Table 11), for an average of 82%. However, when all forms of measurement error, including halo error, were removed from ratings, the general factor at the true score level accounted for only 60% of total variance, for a difference of 22 percentage points. One conclusion from these findings is that factor analyses of within-rater dimension intercorrelations are deceptive and should be avoided in substantive research. Another conclusion is that at the construct level, more than half the variance of the ratings stems from a general factor; that is, when the effects of all four forms of measurement error—(a) rater effects, (b) halo error (specific factor error), (c) random response error, and (d) transient error—are controlled, the general factor accounts for 60% of the variance. Hence, a general factor not only exists but also is fairly large. As we discuss later, this has implications for theories of job performance.
observed ratings of performance dimensions. An argument can be made that it is not the average correlation of .77 that is important but the pattern of correlations. The implication is that some dimensions of performance might have low correlations with the general factor, even though the average is .77. However, examination of Table 10 (second column) shows that the range of correlations of performance dimensions with the general factor is .63 to .92. (The average of these loadings is the .77; this average is the square root of the proportion of variance accounted for by the general factor.) So, none of the performance dimensions had a low correlation with the general factor; in light of this, prospects for discriminant validity for even the dimension with the loading of .63 are constrained.

In approaching the issue of discriminant validity, we must first view the prospects for discriminant validity from the viewpoint of dimension intercorrelations at the true score level. Then, as the second step, one must view discriminant validity prospects from the viewpoint of observed scores. If dimension intercorrelations are substantial at the true score level, this indicates that at the construct level there are impediments to the achievement of discriminant validity for measures of those constructs (dimensions). The concept of discriminant validity applies to measures (i.e., observed scores), but the likelihood of achieving discriminant validity in measures depends in part on the size of construct intercorrelations (estimated by true score correlations). At the extreme, if these correlations are 1.00, then discriminant validity of measures is impossible a priori. If dimension intercorrelations are substantial at the true score level, then they can be substantially reduced at the observed score level simply by using unreliable measures of the dimensions. Such reduced intercorrelations are of course a methodological artifact and do not imply improved discriminant validity. In fact, such “discriminant validity” can be demonstrated even if all measures measure exactly the same construct (i.e., all correlate 1.00 at the true score level).

Consideration of the properties of observed scores (actual ratings) suggests that the prospects for discriminant validity are even more constrained at the observed score level than is indicated by the true score correlation findings. The term discriminant validity refers to relations among measures, not to true score relationships. As we have shown, relationships at the observed score level are usually based on intrarater correlations—and these are greatly inflated by halo error. Beyond the data presented in our research here showing this, Mount et al.’s (1998) findings provide support for the conclusion that discriminant validity prospects are poor at the true score level:

As [Table 3] indicates, the method [halo] loadings are very strong. In fact, the average rater factor loading is .84. Although the trait [dimension] loadings are nonzero (confidence intervals around the loadings exclude zero), they are considerably weaker in magnitude than the rater level loadings. The average trait [dimension] loading is .30. That the method [halo] loadings are much higher than the trait [dimension] loadings is consistent with the multitrait–multimethod results presented earlier. (p. 567)

These findings, based not on meta-analysis but on a single very large data set, mirror our findings that the factor structure of ratings (and therefore the discriminant validity of dimensional ratings) is greatly obscured by halo error (which Mount et al., 1998, referred to as method variance, where the “method” is the individual rater). In the present study, by focusing both on interrater correlations of dimensions (rarely available in study data sets) and on interrater true score correlations, we were able to eliminate the effects of halo error and other forms of measurement error. Neither of these is possible in regular operational measurement; that is, interrater correlations between dimensions are rarely available, and at the same time, because the term discriminant validity refers to measures, and therefore to observed scores, corrections for measurement error are not part of demonstrating discriminant validity for scales. Stated another way, the fact that at the true score level there is some limited room for discriminant validity does not necessarily mean that this same opportunity exists at the level of observed score measurements, and it is at the observed score level that discriminant validity must be assessed.

Considering between-rater correlations, the general factor accounted for 60% of the variance at the construct level but only 27% of the variance at the observed level (see Table 11), indicating that the combined effects of all four types of measurement error reduce the size of the general factor by more than half. This finding reveals the power of measurement error to obscure the underlying structure in observed data (even when the appropriate type of data—inter-rater correlations—is used). Corrections for measurement error can reveal the obscured structure in the data.

The finding of a substantial general factor in ratings that is not due to halo error supports the arguments of Harvey (1982), Hulin (1982), and Murphy (1982) that partialing out the general factor in ratings is not a viable method for removing the effects of halo error.

<table>
<thead>
<tr>
<th>Correlation type</th>
<th>Supervisor ratings</th>
<th>Peer ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSR</td>
<td>NFI</td>
</tr>
<tr>
<td>Intrarater</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uncorrected</td>
<td>.058</td>
<td>.928</td>
</tr>
<tr>
<td>Corrected</td>
<td>.149</td>
<td>.965</td>
</tr>
<tr>
<td>Interrater</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uncorrected</td>
<td>.050</td>
<td>.921</td>
</tr>
<tr>
<td>Corrected for between-rater reliability</td>
<td>.228</td>
<td>.869</td>
</tr>
</tbody>
</table>

Note. RMSR = root-mean-square residual; NFI = normed fit index.

* Percentage variance accounted for by the general factor.  
  b Corrected for within-rater reliability.  
  c Correlations between supervisor and peer raters.
from ratings. It is clear from the present findings that such a procedure would partial out not only the effects of halo error but also the substantial nonhalo general factor in rated job performance. At the time that Landy et al. (1980) recommended this procedure, there were no systematic analyses of empirical data available to prove that it would be flawed in this manner; neither were there any methods for conducting such an analysis. The findings of the present study provide such an analysis, based on a meta-analysis of the published research literature from over 90 years, and would seem to settle this issue. The simple expedient of partialing out the general factor in ratings will not provide a solution to the problem of halo error.

**Theoretical Basis for the General Factor**

Why is there a general factor independent of halo in the ratings data? One possible reason is that citizenship behaviors (called *contextual performance* by Borman & Motowidlo, 1993) affect ratings of all performance dimensions (including noncitizenship dimensions) for both supervisor and peer raters (see also Brief & Motowidlo, 1986). If this is the case, the effect would be to yield (or increase the size of) a general factor. Orr, Sackett, and Mercer (1989) presented evidence that citizenship behaviors do affect supervisory ratings of job performance, even on noncitizenship dimensions of job performance. Note that for many of the performance dimensions (e.g., effort, leadership, quality, administrative competence, interpersonal competence, compliance with rules), it appears that citizenship behaviors, broadly defined, should be considered in assigning performance ratings. Citizenship behaviors would presumably contribute to enhanced performance on these dimensions. Other dimensions, such as job knowledge and productivity (output), appear not to involve a citizenship aspect. If ratings on all dimensions, including those not affected by citizenship behavior, are affected by shared rater perceptions of citizenship performance, as suggested by research, this would, at a minimum, increase the size of any general factor.

Basic individual-differences variables may also affect performance on all rated job performance dimensions. For example, research findings suggest that both general mental ability and the personality trait of conscientiousness contribute to performance on all dimensions of job performance (Feldman, 1981; Motowidlo et al., 1997; Nathan & Alexander, 1988; Ones & Viswesvaran, 1996; Pearlman, Schmidt, & Hunter, 1980; Schmidt, Ones, & Hunter, 1992). Indeed, this finding is a major reason for hypothesizing a general factor in job performance measures (Viswesvaran, 1993). From a theoretical point of view, it would appear likely that the combined effect of these two processes would be sufficient to produce a valid, nonartifactual general factor in job performance and therefore in job performance ratings.

**Implications for Theory Development**

The finding of a general factor bears on the development of theories of job performance. This general factor refers to a psychological construct and not to the economic construct that underlies all job performance measures (e.g., Krug, 1961; Schmidt & Kaplan, 1971; Stone & Kendall, 1956; Tiffin & McCormick, 1958; see also the following section), hence it is relevant to psychological theories of job performance. Job performance has been the dependent variable in much research and has been used as a construct in theories of work behavior and human motivation (cf. Schmidt & Hunter, 1992). Some theories (e.g., J. P. Campbell, 1990; J. P. Campbell et al., 1996; J. P. Campbell et al., 1993) may superficially appear to downplay a role for a general factor because they focus on conceptually distinct substantive dimensions of job performance. However, there is no contradiction between an emphasis on substantive performance dimensions and the existence of a general factor if performance dimensions are positively correlated, as they almost certainly are. Thus, these theories do not deny the existence of a general factor; rather, they simply do not address the issue. On the other hand, some causal theories (e.g., Schmidt & Hunter, 1992), because of their emphasis on overall job performance, appear to emphasize the importance of the general factor. The results reported in this article indicate that different rated job performance dimensions share substantial variance that is independent of halo error. This suggests that theories of job performance need to be developed that address and accommodate this common variance (i.e., the general factor). Theories that assume either the absence of a general factor or that assume that the general factor is small and unimportant are unlikely to be scientifically productive. However, group factors represented in the 40% of true score variance not accounted for by the general factor may come to play an important role in future theories of job performance, in addition to the general factor.

**Other Implications**

The finding of a large general factor in job performance also has implications for criterion measurement in validation studies. Specifically, it suggests that the practice of combining component measures of job performance into a measure of overall job performance, as has been done in most primary validity studies and validity generalization studies, rests on a justifiable (and, as we discuss below, a psychological) foundation. Schmidt and Kaplan (1971) pointed out that if different dimensions of job performance are uncorrelated (i.e., there is no general factor), then a measure of overall job performance produced by summing across the dimensions does not represent a psychologically meaningful construct of performance, and hence such a measure does not contribute to the psychological understanding of the nature of job performance. This is not to say that such a measure cannot represent the economic construct of value to the organization. For example, if salespeople all sell two products, and the correlation between their sales of the two products is zero, the sum of these two sales indices may still be a construct-valid measure of the economic value of the employee to the organization. Although such a measure may have practical value in applied work, it is unlikely to be able to advance psychological understanding of the construct of job performance. On the other hand, when performance dimensions are substantially correlated (i.e., there is a general factor), a measure of overall job performance is psychologically meaningful (i.e., represents a psychologically meaningful construct of overall job performance) and hence has the potential, when used in research, to contribute to psychological understanding of the nature of job performance (e.g., by examination of the kinds of predictors that correlate with it). This does not imply that group factors in job performance are not important. Group factors represented in the 40% of true score variance not accounted for by the general factor may have impor-
tant applied implications, in addition to possibly playing a role in future theories of job performance. In summary, if there were no general factor, then any such measure of overall job performance would merely be a composite of uncorrelated variables and would lack psychological meaning (Schmidt & Kaplan, 1971). Psychologically meaningful validation research would require separate validation for each performance dimension. However, the presence of a large general factor in job performance indicates that measures of overall job performance can have construct validity—and are not merely heterogeneous indices of the economic value of the individual to the organization (Schmidt & Kaplan, 1971).

A reviewer inquired as to whether this analysis implies that different causes of job performance are necessarily highly correlated. It does not. For example, on a particular job the two main causes of job performance might be general mental ability (GMA) and conscientiousness. These two causes may be uncorrelated, while at the same time both of them contribute to correlation among the performance dimensions. Hence, high correlations among performance dimensions would be due in part to the fact that GMA is a determinant of standing on all dimensions and in part to the fact that conscientiousness is a determinant of standing on all dimensions.

Limitations

A limitation of this study is the fact that some of the correlations and reliability estimates were based on small sample sizes. However, even the smallest sample sizes were in the hundreds, thus mitigating the effects of sampling error as compared to individual primary studies. A related concern is the small number of samples we meta-analyzed for some interdimension correlations and reliability estimates. The small number of samples we meta-analyzed raises concerns of the generalizability of the results to the entire population of jobs in the U.S. economy. This limitation can be addressed only by more primary studies that report intercorrelation and reliability estimates for the different job performance dimensions assessed in those respective samples.

A reviewer suggested that the use of lexically derived job performance dimensions was a limitation in this research, hypothesizing that such use leads to inflation of interdimension correlations and hence to an overestimation of the size of the general factor. The inflation of correlations was hypothesized to stem from the fact that dimensions such as quality and productivity (see Table 2) are quite broad and contain aspects of different performance dimensions. However, it must be remembered that no rater ever used the lexical dimensions in our study to rate job performance. Instead, within each study we coded the job performance dimensions rated in that study into the lexical classification categories. The average correlation between any two dimensions is the average across all studies that contained ratings classified into those two dimensions. Thus, this average is based on the actual, narrower (more specific) rating dimensions used in the individual studies and is not affected by the somewhat greater breadth (or abstractness) of the lexical dimensions. It is these averages that are the basis of our study. If raters in the individual studies had rated employees on the actual broad lexical categories we used to organize the literature, then possibly the broad nature of these categories might have caused inflation of interdimension correlations. In fact, however, this never happened, and so it cannot be the basis for a hypothesis that interdimension correlations are inflated. Another, more direct way to address this hypothesis is to examine the actual uncorrected interdimension correlations in Tables 5, 6, and 7. In Table 5, the average uncorrected within-rater correlation between dimensions for supervisory ratings is .55, a value that is not particularly large relative to those typically observed in individual studies. The comparable average value for peer ratings (see Table 6) is .58, again a value not larger than those we have typically seen in individual studies. Finally, in Table 7, the average interrater–interdimension correlation is .27, once again not a value that seems inflated. It is important to note that the data in Table 7 are the basis for estimation of the size of the general factor.

Another potential limitation mentioned by a reviewer was the possibility that the lexical dimensions might not correspond to the factors or dimensions resulting from factor analyses in the literature. As we noted earlier, factor analyses of ratings data in the literature have been based on intraindividual correlations between rated performance dimensions and have therefore been dominated by a large artifactual halo factor. An important conclusion from the present research is that such factor analyses cannot provide accurate information about the actual nonartifactual factor structure of ratings. Hence there would be no point in comparing the lexical dimensions with “factors” derived from such factor analyses.

One could argue that a limitation of this study is the fact that the estimate of the size of the general factor is based on averages across jobs and that the size of the general factor may vary across jobs; that is, that we have estimated the size of the general factor for the average job, but it may be larger or smaller than this for different jobs. A key question is whether there is any basis for hypothesizing a larger general factor for some jobs than others. One possibility concerns the role of GMA. The validity of GMA for predicting job performance is larger for jobs of higher complexity. However, this statement refers to job applicants and does not necessarily apply to job incumbents, who are affected by range restriction on GMA. In a world without range restriction, if GMA is a more important determinant of performance on all performance dimensions in jobs of higher complexity, this fact would cause higher interdimension correlations on jobs of higher complexity and hence a larger general factor in job performance. In reality, however, range restriction on GMA exists and is typically more severe in jobs of higher complexity (Schmidt & Hunter, 2004). Range restriction lowers the dimension intercorrelations among incumbents and therefore lowers the size of the general factor in performance. Hence, when the focus is on incumbents (as it necessarily and appropriately is in the present research), the general factor in job performance may be no larger in high- than in low-complexity jobs. Furthermore, this refers to actual performance, not rated performance. Even if there were differences across jobs in the size of the general factor in actual performance, there would still remain the question of whether such a difference would be reflected in ratings of job performance; that is, there would still be the question of whether raters could pick up this difference. These are questions for future research.

Future Directions

Conceptualization of the construct of job performance has been evolving. Our conceptualization of the construct was based on grouping the different measures reported in the literature over the
last 90 years into conceptually similar categories. Newer dimensions (cf. Howard, 1995) that partly overlap with the dimensions studied here are being emphasized in recent years and may become more central to the construct of overall job performance as jobs change. For example, customer service can be construed as part of interpersonal competence and quality of job performance (two dimensions we studied in the present research). Future research should include these newer, often compound dimensions (e.g., customer service) in examining the general factor across ratings. There is also a need to study narrower, more specific performance dimensions. The construct of job performance can be sliced in different ways, depending on the level of specificity desired in its measurement. Future research should explore whether the size of the general factor in ratings and the amount of halo differ depending on the specific set of performance dimensions studied.

The general impression driving job performance ratings by an individual rater is composed of a rater-idiosyncratic component and a component shared across raters. With respect to the shared component, our estimates of the magnitude of the general factor cannot distinguish between real world (objectively existing) covariance across dimensions and covariance due to misperceptions that might be shared across raters (i.e., shared halo); that is, although we were able to isolate the effects of rater-idiosyncratic halo in ratings, we were not able to further estimate effects due to shared halo versus real covariance. Future research should attempt to disentangle these two sources of variance.

We examined both peer and supervisor ratings and found very few studies that reported correlations between peer or supervisor ratings and subordinate or customer ratings. In one of the rare studies of this kind, Mount et al. (1998) reported correlations based on subordinate ratings. More such research is needed. Another avenue of fruitful future research involves an investigation of moderator effects of the interdimensional correlations as well as reliability estimates. However, prior cumulative research (cf. Peterson, 1994; Viswesvaran et al., 1996) has found little support for hypothesized moderators. Nevertheless, future research along those lines may refine our understanding of the construct of job performance.

On a related note, this research focused exclusively on the magnitude of the general factor in ratings and did not explore the magnitude of potential group factors. The extant literature discusses in detail the differences among task performance, contextual performance, and counterproductive behaviors (Borman & Motowidlo, 1993; Ones et al., 1993; Viswesvaran & Ones, 2000). Researchers (e.g., Guys & Sackett, 2003) have called for the examination of the intercorrelation between different subdimensions that make up these three domains—task performance, contextual performance, and counterproductive behaviors. Future research should extend the results reported here to assess the magnitude of group factors net of rater idiosyncratic halo error.

**References**

References marked with an asterisk indicate studies included in the meta-analysis.


A GENERAL FACTOR

of physical ability tests for predicting performance in craft jobs. Journal of Applied Psychology, 64, 262–274.


Received December 6, 2002
Revision received December 15, 2003
Accepted December 30, 2003