The Impact of Job Complexity and Performance Measurement on the Temporal Consistency, Stability, and Test–Retest Reliability of Employee Job Performance Ratings

Michael C. Sturman, Cornell University
Robin A. Cheramie, Winthrop University
Luke H. Cashen, Louisiana State University

Although research has shown that individual job performance changes over time, the extent of such changes is unknown. In this article, the authors define and distinguish between the concepts of temporal consistency, stability, and test–retest reliability when considering individual job performance ratings over time. Furthermore, the authors examine measurement type (i.e., subjective and objective measures) and job complexity in relation to temporal consistency, stability, and test–retest reliability. On the basis of meta-analytic results, the authors found that the test–retest reliability of these ratings ranged from .83 for subjective measures in low-complexity jobs to .50 for objective measures in high-complexity jobs. The stability of these ratings over a 1-year time lag ranged from .85 to .67. The analyses also reveal that correlations between performance measures decreased as the time interval between performance measurements increased, but the estimates approached values greater than zero.

Although it may be common to hear that past performance is the best predictor of future performance, it was noted over 40 years ago that "if the desired criterion is ultimate or total performance, there is some question whether an initial criterion measure will itself be a good predictor" (Ghiselli & Haire, 1960, pp. 230–231). Now, despite a long history of research that has recognized that individual job performance may be dynamic (e.g., Ghiselli, 1956; Ghiselli & Haire, 1960) and that predicting individual job performance to industrial– organizational psychology and organizational practice is of fundamental importance, relatively little is known about the nature of individual job performance over time (Ployhart & Hakel, 1998). Our intent in this article is to contribute to the present literature on dynamic performance by specifically investigating the extent that past performance does predict future performance and how this relationship is moderated by time, job complexity, and the method of performance measurement.

Ghiselli (1956) originally introduced the concept of dynamic criteria, stating that “it is apparent that the performance of workers does change as they learn and develop on the job” (p. 2). Soon thereafter, Humphreys (1960) demonstrated that the correlation between performance measures decreased as the amount of time between performance measures increased. Since this original work on dynamic criteria, abundant empirical evidence has verified that job performance measurements are not perfectly correlated over time (e.g., Austin, Humphreys, & Hulin, 1989; Barrett & Alexander, 1989; Barrett, Caldwell, & Alexander, 1985; Ghiselli & Haire, 1960; Ployhart & Hakel, 1998; Rambo, Chomiak, & Price, 1983; Sturman & Trevor, 2001).

The breadth of research on performance dynamism certainly has provided evidence that measures of individual job performance are not equal over time. Recent research has emerged involving performance over time, such as theoretical models on the relationship between abilities and performance (e.g., Farrell & McDaniel, 2001; Keil & Cortina, 2001), empirical models of performance trends (Deadrick, Bennett, & Russell, 1997; Ployhart & Hakel, 1998), and examinations of the implications of performance changes (Harrison, Virick, & William, 1996; Sturman & Trevor, 2001). However, there has been scant attention paid to the actual measurement of job performance in a longitudinal context. Ghiselli (1956) long ago noted that “far more attention has been devoted to the development of predictive devices than to the understanding and evaluation of criteria” (p. 1). Today, this remains true, as there has yet to be any research that has specifically estimated the extent to which job performance changes over time are attributable to unreliability of performance measures versus actual changes in job performance (Sturman & Trevor, 2001) or how the extent of performance dynamism is attributable to job or individual characteristics.

The failure to distinguish between actual changes in performance and sources of error variance has created a notable deficiency in the job performance literature; specifically, if the extent of actual performance change is unknown, then it is not clear what the research investigating performance changes over time is actually examining. In cross-sectional research, it has become standard practice to report the reliability of measures. So much so that research has emerged discussing the different types of performance measurement reliability and how these types of reliability relate to practice (e.g., Viswesvaran, Ones, & Schmidt, 1996). Not until recently has psychometric research emerged addressing the measurement of reliability of measures over time (Becker, 2000; Green, 2003). Although it has been acknowledged in the dynamic performance literature that observed performance changes may be attributable to measurement error (e.g., Barrett et al., 1985; Hanges, Schneider, & Niles, 1990), to our knowledge no research has yet specifically examined this phenomenon. Thus, the first contribution of this article is to partial out systematic changes from random fluctuations of individual performance over time, ultimately providing an estimate of the stability and test–retest reliability of performance measures over time. This is accomplished through the definition and differentiation of three concepts—temporal consistency, stability, and test–retest reliability—and how they relate to the measurement of individual job performance over time. As such, in this study we ask the following: What portion of performance dynamism is attributable to a lack of stability in individual job performance versus test–retest unreliability?

Research on dynamic performance has not examined the method of performance measurement as a potential moderator of the level of performance dynamism, although there is a notable body of literature debating the conceptual and empirical distinctness of different types and sources of performance evaluations (e.g., supervisory evaluations vs. objective performance measures; see Bommer, Johnson, Rich, Podsakoff, & MacKenzie, 1995; Lance, Teachout, & Donnelly, 1992; Sturman, 2003; Vance, MacCallum, Coovert, & Hedge, 1988; Viswesvaran, Schmidt, & Ones, 2002). As such, the second contribution of this study is that we build on existing work about performance measurement to consider how the method of performance measurement affects the temporal consistency, stability, and test–retest reliability of job performance ratings over time. Thus, in this study we ask the following: How will performance's test–retest reliability be influenced by the type of performance measure under consideration?

Similarly, studies on dynamic performance research have not examined how job characteristics might moderate the level of performance dynamism, even though job complexity has been shown to be a moderator to a number of relationships involving job performance (e.g., Farrell & McDaniel, 2001; Keil & Cortina, 2001; McDaniel, Schmidt, & Hunter, 1988; Oldham & Cummings, 1996; Sturman, 2003). As such, the third contribution of this study involves examining how job complexity influences performance longitudinally. Specifically, we ask the following: How does job complexity affect the stability and test–retest reliability of job performance measures?

Defining Temporal Consistency, Stability, and Test–Retest Reliability

Considering the issue of time in the context of a measure that is often studied cross-sectionally (e.g., job performance) can create confusion if clear steps are not taken to articulate and integrate the issues related to time into the theory and methods of a study (Mitchell & James, 2001). As such, we need to clearly delineate between three distinct terms—temporal consistency, stability, and reliability—to distinguish observed dynamism (i.e., the lack of temporal consistency) from the sources for such dynamism (i.e., the lack of stability and the lack of reliability). Furthermore, we distinguish between the types of error that cause a lack of reliability (Hunter & Schmidt, 1990; Schmidt & Hunter, 1996). Making these distinctions is important because (a) they are integral to understanding the different phenomena related to the observation of dynamic performance (Heise, 1969), and (b) prior research has used these terms inconsistently (e.g., Deadrick & Madigan, 1990; Heise, 1969; Kerlinger, 1986; Nunnally & Bernstein, 1994; Pedhazur & Schmelkin, 1991; Schwab, 1999).

Figure 1 illustrates many of the phenomena we are studying. The figure illustrates an example of a performance construct being measured by *N* items at Time 1 and Time 2. In the figure, solid lines and bold-text words represent the performance measurement model; the dotted lines and plain-text words represent the various definitions and phenomena used in our explanations below.

We define stability as the extent to which the true value of a measure remains constant over time (Carmines & Zeller, 1979; Heise, 1969). An individual's performance may change because of a change in motivation, the acquisition of job knowledge, or changes in the predictors of performance (Alvares & Hulin, 1972, 1973). The level of stability is illustrated in Figure 1 by the dotted curved arrow between the true scores at Time 1 and Time 2.

Because the construct of true performance cannot be observed, it is necessary to produce a measure of job performance. Thus, performance is estimated by a set of N items, which create the observed measures of performance (labeled as Performance Measure [Time 1] and Performance Measure [Time 2]). Frequently, research considering performance over time has examined the relationship between these measures of performance. We defined this relationship—the correlation between performance measures at different points of time—as temporal consistency (Heise, 1969). This is illustrated in Figure 1 by the dotted curved arrow between Performance Measure [Time 1] and Performance Measure [Time 2].

The accuracy of job performance measurement is reduced by the introduction of various types of error variance from the measurement process. The introduction of error variance causes a lack of *reliability*, defined here as the degree of convergence between an observed score and the true score that the measure estimates (Schwab, 1999). This may be attributable to many types and sources of error, some truly random and others systematic. Thus, it is important to model and control for (or at least understand) the different sources of error variance (Murphy & De Shon, 2000). This includes sampling error, random response error, item-specific error, and transient error (Hunter & Schmidt, 1990; Schmidt & Hunter, 1996). The size and nature of this error may be a function of the measurement device, rater, ratee, or interactions of the same (Murphy & De Shon, 2000). Sampling error and random response error (discussed in depth elsewhere; cf. Hunter & Schmidt, 1990; Schmidt & Hunter, 1996; Schwab, 1999) are not represented in Figure 1 (although sampling error is discussed in the Methods section and considered in our analyses); item-specific errors can be introduced by something peculiar to the measurement situation or the different ways that individuals respond to the phrasing of an item (Hunter & Schmidt, 1990). The item-specific errors create a lack of internal consistency, which generally is captured by coefficient alpha (Cronbach, 1951) and can be mitigated somewhat by the use of multiitem measures. Transient error occurs because the items at a given point in time may be affected similarly by some influence that varies over time, such as mood (Green, 2003; Hunter & Schmidt, 1990). When performance is measured in only one time period (i.e., cross-sectionally), then the measure of internal consistency will not capture the measure's level of transient error. Consequently, when researchers use only a measure of internal consistency, they will underestimate the total amount of unreliability of a measure (Green, 2003; Hunter & Schmidt, 1990).

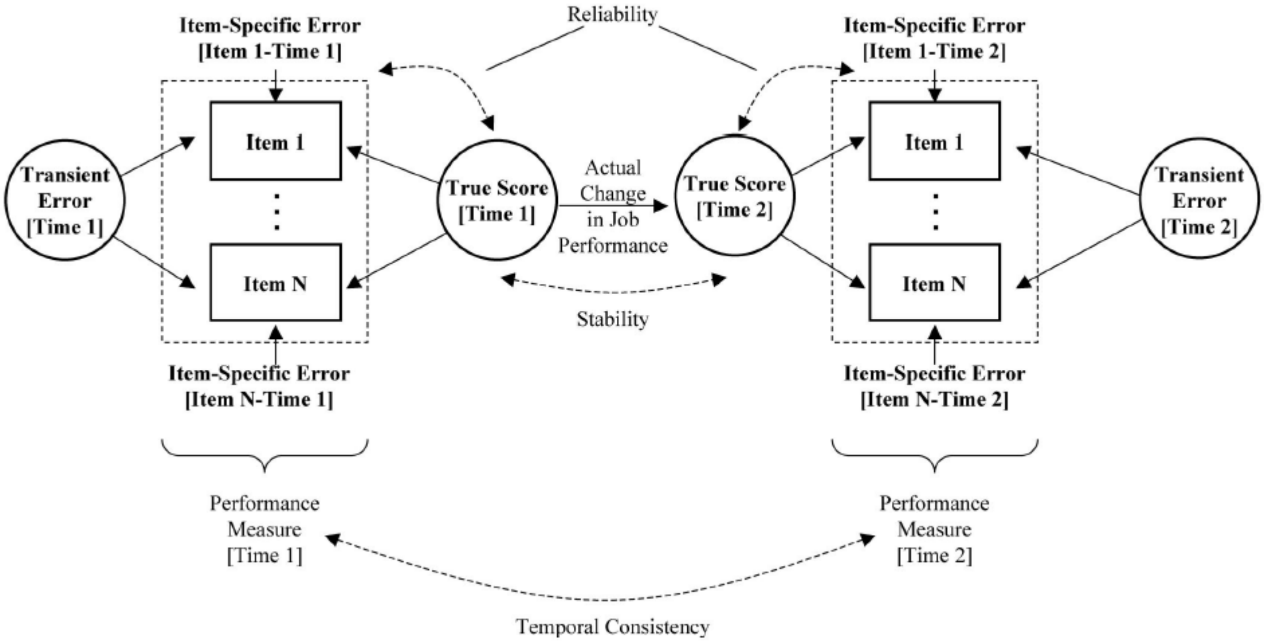
Reliability
Item-Specific Error [Item 1-Time 1]
Item-Specific Error [Item 1-Time 2]
Transient Error [Time 1]
Item 1
Item N
True Score [Time 1]
Actual Change in Job Performance
True Score [Time 2]
Item 1
Item N
Transient Error [Time 2]
Stability
Item-Specific Error [Item N-Time 1]
Item-Specific Error [Item N-Time 2]
Performance Measure [Time 1]
Performance Measure [Time 2]
Temporal Consistency

Figure 1. Stability, consistency, and error in longitudinal ratings of job performance. Solid lines and bold-text words represent the performance measurement model; dotted lines and plain-text words represent the various definitions and phenomena described in this article.

As discussed in depth elsewhere (Hunter & Schmidt, 1990; Schmidt & Hunter, 1996), the different sources of error create the need for different types of reliability estimates, each accounting for a different source of error. There is no single measure representing the reliability of a measure (Murphy & De Shon, 2000), but there are different characteristics of a measurement device that are important for generalizations to certain questions. For the purpose of contributing to the understanding of job performance over time, we are trying to assess the level of test–retest reliability of performance, defined as the relationship between performance measures over time after removing the effects of performance instability.

Note that test–retest reliability and temporal consistency are different because of the (potential) lack of stability. If the true scores of a construct were stable, then test–retest reliability would equal temporal consistency. However, if the true value of a construct changes over time, then temporal consistency captures the correlation of the measures over time, whereas test–retest reliability must partial out the lack of consistency due to actual changes in the construct.

The importance of our research is best revealed by recent work that has examined the implications of performance changes (see Harrison et al., 1996, and Sturman & Trevor, 2001). In both of these studies, the authors showed that changes in performance were predictive of employee turnover, even after controlling for the individual's most recent performance score. Although both of these studies used objective measures of performance (i.e., sales) in their models, neither study differentiated between potential test–retest unreliability and stability. Indeed, this was not directly relevant to the original focus of these studies in that they both examined the consequences of performance changes, not the causes. Consequently, neither study can truly speak to the reasons for the performance change–turnover relationship. If all performance changes are attributable to true changes in performance (i.e., no test–retest unreliability), then individuals may be more or less inclined to leave their companies because they can forecast and make decisions based on their future potential performance levels (and hence pay levels in their commission systems). However, if all performance changes are attributable to test–retest unreliability (i.e., true performance levels do not change over time), then the relationship between performance changes and turnover may be attributable to psychological factors, such as perceived fairness of the system and employees' perceptions of self-efficacy. We argue that a better understanding of the stability and test–retest reliability of performance ratings will help all research associated with longitudinal ratings of job performance.

Understanding Job Performance Over Time

Longitudinal Implications of Ackerman's Model

Theoretical development aimed at understanding job performance over time has been based on the skill acquisition model created by Ackerman (1987, 1988, 1989). This model proposed that the relative importance of various abilities for task performance changes as individuals develop skills in performing a task. Specifically, when one is new to a task, the task is performed slowly, with effort, and often with error; however, as skills in the task are developed, the task becomes more automatic, thus requiring fewer cognitive resources, and performance improves (Farrell & McDaniel, 2001). Ackerman's theory has generally been supported in both laboratory and industrial settings (Farrell & McDaniel, 2001; Keil & Cortina, 2001); however, such tests have involved cross-sectional analyses, and the model has not been extended to consider performance measured longitudinally.

Ackerman's model presented a logical explanation for the presence of dynamic performance. As individuals gain experience, their performance is expected to follow a learning curve (Farrell & McDaniel, 2001). Although learning curves follow a general pattern, individual performance changes at different rates because of individual differences in abilities, motivation levels, and opportunities to perform. Similarly, Alvares and Hulin (1972, 1973) described two general models of dynamic performance: the changing subjects model and the changing tasks model. In the former model, performance changes because of individual changes in performance- causing characteristics, such as job knowledge and motivation. In the changing tasks model, performance changes because the determinants of performance change over time.

Both Ackerman's work on skill acquisition and the dynamic performance models defined by Alvares and Hulin espoused that performance changes with the passage of time. Furthermore, individual performance changes over time in systematic (i.e., nonrandom, although not necessarily known) ways. It would be valuable for research aimed at predicting individual job performance over time to have an estimate of the stability of individual performance, as this would provide information on the nature of job performance and indicate the extent to which longitudinal performance is potentially predictable.

Evidence for a Lack of Test–Retest Reliability

Although performance may be unstable over time, error variance may make a measure appear more unstable than it truly is (Barrett et al., 1985; Epstein, 1979; Hanges et al., 1990). Yet, although researchers have examined many aspects of job performance rating reliability (e.g., Bommer et al., 1995; Viswesvaran et al., 1996), this research has focused on the reliability of performance measures at a point in time: either intrarater reliability (i.e., internal consistency of the performance measure) or interrater reliability (i.e., the consistency of performance measures across raters). The reliability of a measure over time is also a critical element of a measure's construct validity (e.g., Cronbach, 1951; De Shon, Ployhart, & Sacco, 1998; Green, 2003; Nunnally & Bernstein, 1994; Schwab, 1999), but this aspect of performance measurement is still in need of a focused research effort.

A lack of test–retest reliability may be attributable to a number of factors. When completing evaluations, the rater (e.g., manager) may become distracted, misunderstand items, rush through items, or feel uncomfortable because of someone else being present. The ratee may also contribute to error, such as by trying to manage impressions, directing the conversation toward positive aspects of performance, or otherwise trying to bias the rating process. Furthermore, both the ratee's and rater's mood or disposition at any given time may affect any ratings.

It is likely that these sources of error would affect all the items simultaneously. Yet the factors adding error to the appraisal process at one point in time may be different when a second set of ratings is done at another time period. For example, the halo error in appraisal would cause the items measured at a point in time to be both inflated and highly correlated. This transient error would affect items at a particular time such that the items would be more similar at the same time than across time (Green, 2003). In other words, a multiitem performance measure may have perfect intrarater reliability (i.e., a coefficient alpha of 1.0) but still lack perfect reliability that would only be noted if the measure's test– retest reliability was assessed.

Error may also be introduced into performance measurement because of environmental factors affecting the results of job performance. For example, the opportunity to perform, or situational constraints that are beyond an individual's control, may affect individual performance (Blumberg & Pringle, 1982; Peters & O'Connor, 1980; Steel & Mento, 1986). Even if an individual's ability and motivation remained constant over time, performance ratings may appear unstable because the context within which performance occurred may have changed. Environmental constraints will also likely change over time, attenuating test–retest reliability as such constraints affect the performance rating process in different ways at different times.

For example, in a given year, a particular worker may have projects that need to be completed mostly from inside the main office. As such, that employee's supervisor may have the opportunity to observe the employee's performance accurately. In another year, the same worker may have a project that forces her to do her tasks at a different location. Although her true performance may not change, the change in this worker's observability by her supervisor may cause a change in her rated performance. Furthermore, even if she was evaluated through objective measures, then the environmental constraints that changed the type of project that she worked on each year may influence the performance evaluation received.

Some studies have tried to minimize test–retest unreliability by aggregating performance over multiple time periods (e.g., Deadrick & Madigan, 1990; Epstein, 1979; Hofmann, Jacobs, & Baratta, 1993;

Hofmann, Jacobs, & Gerras, 1992; Ployhart & Hakel, 1998). In these studies, the authors argued that by aggregating multiple performance measures, such as by using average sales over 3 months instead of monthly sales, performance would be less susceptible to random shocks. This logic is essentially an implicit argument that aggregation would help reduce the effect of test–retest unreliability. Such aggregation, however, may end up ignoring actual performance changes. Perhaps more importantly, the aggregation of data causes the loss of potentially valuable information. Indeed, when considering individual task performance over time, changes in performance from one time period to the next have been shown to be related to employee turnover (Harrison et al., 1996; Sturman & Trevor, 2001). It may therefore be inappropriate to inflate the temporal consistency of performance data with such aggregations.

Evidence for a Temporally Stable Component of Job Performance

Although our review suggests that measures of performance will be unequal over time because of a lack of performance stability and test–retest reliability, there is also evidence that some aspects of job performance are stable over time. This is not to say that performance will not change over time; rather, we argue that because of the nature of how individual characteristics affect performance, some variance in individual performance remains consistent. In other words, because of the stability of certain performance-causing attributes, we predict there will always be some positive relationship between true performance scores over any given time period.

Building on Ackerman’s skill acquisition model, Murphy (1989) hypothesized that the relationship between cognitive ability and job performance decreases with job experience, but he did not suggest that the relationship reached zero. Rather, he predicted that there are aspects of jobs that change over time, and thus, there always remains a need to learn new behaviors or adapt to new circumstances. Consequently, Murphy suggested that cognitive ability will continue to play a role in the prediction of job performance over time.

This theory is indirectly supported by the following combination of evidence: (a) Cognitive ability remains relatively stable over the period that most people are in the workforce (e.g., Bayley, 1955; Jensen, 1980; Judge, Higgins, Thoresen, & Barrick, 1999), and (b) performance is often well predicted by cognitive ability (Hunter & Hunter, 1984; Ree, Earles, & Teachout, 1994; Salgado, Anderson, Moscoso, Bertua, & Fruyt, 2003). This theory was directly supported by Farrell and McDaniel’s (2001) study. Specifically, they showed that, although the relationship between cognitive ability and performance declined with job experience, the relationship was still positive (i.e., a correlation of .20) at 10 years.

Other research provides indirect, but valuable, support of a stable component of job performance. One stream of research has examined the effects of personality—which is relatively stable over time (Costa & McCrae, 1988, 1992)—on individual job performance, with evidence that the Big Five personality characteristics are related to individual job performance (Barrick & Mount, 1991). Additionally, childhood ratings of personality characteristics were shown to relate to measures of career success measured in late adulthood (Judge et al., 1999). This personality research suggests that these traits maintain their relationship with job performance as they do over time with measures of career success; as such, a portion of individual job performance should be stable.

Summary

Job performance ratings change over time, with evidence indicating that this is attributable to a lack of stability and test–retest reliability. Because of the lack of test–retest reliability, we expect any correlation between job performance measures at different points of time to be less than one, but the nature of this relationship should be a function of time. The true stability of performance should be perfect (i.e., 1.0) when there is no time lag and then decrease as the time lag between performance measures increases. Yet, when we model temporal consistency over time, any lack of test–retest reliability decreases the magnitude of the observed relationships. Because of imperfect reliability (both internal consistency and test–retest reliability), the predicted temporal consistency levels should be less than one even when there is no time lag between measures. If we consider this relationship after correcting for the lack of internal consistency, then the remaining attenuation is attributable to the lack of test–retest reliability. In other words, when plotting the relationship between temporal consistency and time (after this relationship is corrected for attenuation due to a lack of internal consistency of the measure), the intercept of the relationship (i.e., extrapolating the relationship to the point in which there is no time lag) should represent the level of test–retest reliability. Thus, we predict that the relationship between performance scores across time will have a positive intercept that is less than one and will have a curved form that decreases as the time between performance measures increases because of a lack of performance stability, but the curve will not ultimately reach zero because of the expected stable component of performance. Translating this prediction into specific, falsifiable hypotheses, we expect the following:

Hypothesis 1a: The predicted temporal consistency between measures of individual job performance with a hypothetical time lag of zero will be less than one.

Hypothesis 1b: As the time lag between performance measures increases, the relationship between performance measures will decline.

Hypothesis 1c: As the time lag between performance measures increases, the relationship between performance measures will approach a value greater than zero.

Moderating Influences on Performance Stability

So far, our discussion has focused on the extent to which measures of overall performance exhibit dynamic and static characteristics without deference to the type of performance measure or job characteristics. However, past research has shown that sample and measurement characteristics also affect relationships with job performance (e.g., Jackson & Schuler, 1985; Sturman, 2003; Tubre & Collins, 2000). We thus turn to examining characteristics that may moderate the relationships set forth in the first hypothesis.

Performance Measurement

Studies that have examined performance ratings over time have used two different types of performance measurement: supervisory evaluations (e.g., Harris, Gilbreath, & Sunday, 1998; McEvoy & Beatty, 1989) and objective measures of employee productivity (e.g., Henry & Hulin, 1987; Ployhart & Hakel, 1998; Sturman & Trevor, 2001). Recent research has shown that supervisory evaluations capture such performance dimensions as task performance, organizational citizenship behaviors, and counterproductive behaviors (Conway, 1999; Motowidlo & Van Scotter, 1994; Rotundo & Sackett, 2002; Van Scotter, Motowidlo, & Cross, 2000). Although it is tenuous to consider supervisory ratings as equivalent to a true measure of some overarching performance construct (Scullen, Mount, & Goff, 2000), Viswesvaran, Schmidt, and Ones (2005) show that performance ratings are largely reflective of an underlying general factor. Regardless, it is undeniable that supervisory performance ratings play an important role in human resource decision making and research.

A common criticism of researchers using supervisory performance ratings is that they are subject to unreliability and bias (Bommer et al., 1995; Campbell, 1990; Feldman, 1981). Many researchers, and particularly for dynamic performance research, have therefore used objective measures of job performance. Although objective job performance measures do capture obviously important outcomes from an organization's point of view, research has demonstrated that objective (e.g., sales) and subjective (e.g., supervisory rating) measures of job performance are not interchangeable (Bommer et al., 1995; Heneman, 1986). Objective and subjective indicators of performance differ by how, versus what or why, performance is measured (Muckler & Seven, 1992).

Subjective measures of performance are affected by the process of performance measurement (Bommer et al., 1995). Furthermore, subjective measures of overall performance ratings are influenced by the different performance dimensions evaluated by the supervisor (Rotundo & Sackett, 2002). The construct validity of a subjective performance measure is therefore threatened by the decision processes of the rater making the evaluation—such as from bias, contamination, scale unit bias, different weighting schemes or perceived relative importance of various performance dimensions, and criterion distortion (Heneman, 1986)—and perhaps the behaviors of the individual being evaluated, such as from impression management or attempts to influence the evaluator. Consequently, objective measures of performance are believed to have higher reliability at a given point of time.

Yet, the reliability advantage associated with objective performance scores at a given point in time does not necessarily translate to similarly higher test–retest reliability. There are a number of potential reasons for us to hypothesize this effect. First, as illustrated by Figure 1, a lack of item-specific error does not mean there will be a lack of transient error. Even though temporal consistency will be attenuated by a lack of internal consistency, there is no added benefit for researchers to use objective performance measures in reference to minimizing transient error.

Second, objective scores, by their very nature, do not account for circumstances outside of the individual’s control that may affect performance ratings. Objective measures of performance focus on outcomes or results of behavior, not behaviors themselves (Cascio, 1998). Although there is expected overlap between outcomes and behaviors, objective measures of performance also capture the effects of factors outside of employees’ control that have an impact on performance results (Cascio, 1998). For example, economic conditions vary over time, and sales performance has been shown to vary extensively (Barrett et al., 1985). Objective measures of task performance ignore the opportunity factors that may influence the temporal consistency of performance over time, whereas subjective measures of performance provide a means for a rater to consider factors outside of the employee’s control when evaluating performance. Thus, supervisory ratings of performance may actually yield less transient error than objective ratings.

Third, some of the individual decision-making cognitive biases that lead to lower inter- and intrarater reliability may actually lead to greater test–retest reliability. For example, a confirmatory bias may increase the apparent stability of performance ratings across time. The reason for this is that individuals often seek confirmatory evidence and exclude disconfirming information in the decision-making process (Kahneman, Slovic, & Tversky, 1982). Having given a certain evaluation at Time 1, managers may seek information that confirms that rating and thus give similar ratings at Time 2. It is also possible that managers may suffer from the anchoring and adjustment bias (Bazerman, 1998). That is, managers may use past evaluations as a starting point when making future evaluations and may make future ratings by adjusting from this initial value. The problem with this decision process is that adjustments from the starting value often prove to be inadequate, resulting in final assessments that are inappropriately close to the anchor (Tversky & Kahneman, 1973).

To summarize, many of the factors that lead to lower intrarater reliability when considering performance measurement cross-sectionally may actually enhance the temporal consistency of performance ratings. This includes factors that arguably increase the accuracy of performance evaluations (i.e., by considering contextual factors when evaluating performance) and those that decrease the accuracy (e.g., heuristics, biases, and prejudices). Both the benefits of supervisory ratings (i.e., the greater validity due to considering more dimensions of performance) and the negatives associated with supervisory ratings (i.e., the potential biases and heuristics) would cause subjective measures to have less transient error. This would cause subjective ratings of performance to be more consistent over time and, more specifically, to have greater test–retest reliability. Thus, we hypothesize the following:

Hypothesis 2: The test–retest reliability of individual performance measures will be greater with subjective job performance measures than with objective job performance measures.

Performance evaluations are the assessment of either the behaviors or results. We do not expect the method of performance evaluations to affect the way employees change over time. Thus, we do not expect any relationship between measurement type and performance stability.

Job Complexity

Performance research has shown that occupational group characteristics, and more specifically job characteristics, influence relationships with job performance (e.g., Schmitt, Gooding, Noe, & Kirsch, 1984; Sturman, 2003; Tubre & Collins, 2000). Ackerman's (1987, 1988) skill acquisition model and Murphy's (1989) model of performance both suggested that job complexity plays a role in the nature of job performance over time. Furthermore, research has shown that job complexity moderates a number of relationships with job performance (e.g., Farrell & McDaniel, 2001; Gutenberg, Arvey, Osburn, & Jeanneret, 1983; Keil & Cortina, 2001; McDaniel et al., 1988; Sturman, 2003).

As predicted by Ackerman's and Murphy's models, greater complexity necessitates the use of cognitive ability over time to adjust to changing tasks and to learn new skills. One could argue that more complex jobs will have greater stability because cognitive ability remains relatively stable over time, the component of performance that is predicted by cognitive ability should be larger in more complex jobs, and cognitive ability is a good predictor of job performance (e.g., Hunter & Hunter, 1984; Salgado et al., 2003). However, by their very nature, more complex jobs lead to changing job requirements and changing predictors of job performance (i.e., more like the changing tasks model; see Alvares & Hulin, 1972, 1973). This greater complexity will reduce the stability of job performance over time. Furthermore, more complex jobs by their very nature will make the assessment of performance more difficult. The greater complexity of the job should actually increase the amount of transient error at any point in time. In Hypotheses 3 and 4, we predict that the curve posited in Hypothesis 1 will be lower in more complex jobs. More specifically, we predict the following:

Hypothesis 3: The test–retest reliability of individual job performance will be lower in more complex jobs.

Hypothesis 4: The stability of individual job performance will be lower in more complex jobs.

Method

The intent of this article is to estimate and partial out the sources causing a lack of temporal consistency in job performance ratings: a lack of stability and imperfect test–retest reliability. To make this assessment, we have a number of specific methodological requirements. Ultimately, this led us to use a random-effects meta-analysis to test our hypotheses. The reason for this choice is based on the following three requirements.

First, we wanted to differentiate between stability and test–retest reliability; thus, we needed to have data from at least three data-collection periods (Heise, 1969). By assuming that the variance of contextual error was constant across time periods (Green, 2003; Heise, 1969), the use of multiple observations was necessary to estimate the two unknown variables in which we were interested. Second, we wanted to test the moderating effects of performance measurement and job complexity on the stability and test–retest reliability of job performance. To gather sufficient data that (a) had correlations from at least three data periods and (b) represented samples of varying levels of job complexity that used both subjective and objective measures of performance, we chose to quantitatively review the existing literature. Third, given these two requirements, the meta-analytic technique needed to account for error that came from both within and across studies. That is, because there were several correlations for a given sample, the potential existed for correlated sampling errors. Given our needs (a) to account for the correlated errors that may exist because of multiple correlations being obtained from the same sample, (b) to model covariates, and (c) to select a meta-analytic technique that fits with the methodological requirements of the research question being investigated (Hedges & Vevea, 1998; Overton, 1998), we used a random-effects metaanalysis (e.g., Bryk & Raudenbush, 1992; Erez, Bloom, & Wells, 1996; Snijders & Bosker, 1999).

Defining Our Analytic Model

The intent of our methodology was to model the relationship between performance measures over time. Consistent with random-effects metaanalysis, the dependent variable at the first (most micro) level of analysis was based on the correlation of individual performance scores from the original studies. This Level-1 model was the same for all the models described below:

$$\mathrm{r_{ij}} = \mathrm{p_{ij}} + \mathrm{e_{ij}}$$

The observed correlation (r_{ij}: correlation i of study j) is equal to the true correlation (p_{ij}) plus sampling error (e_{ij}, which has a mean of zero and whose variance equals σ^2). A requirement of random-effects meta-analysis is that the variance of the error at this first level of analysis be known (Bryk & Raudenbush, 1992). The variance of the correlation can be estimated with formulae designed to estimate the variation of raw correlations, correlations corrected for statistical artifacts (i.e., unreliability, range restriction; see Raju & Brand, 2003; Raju, Burke, Normand, & Langlois, 1991), or transformed correlations (i.e., Fisher, 1932; Mendoza, 1993).

The second level of the model allowed us to predict the true relationship estimated at Level 1 (p_{ij}). We began our analyses with a base case, both to serve as a null hypothesis and to provide a source of comparison to evaluate the amount of variance explained in our subsequent models. This null model, labeled Model 1, included no covariates to the relationship between performance scores over time. In this model, performance was assumed not to change, and thus, the lack of consistency in performance measures was a function of test–retest reliability:

$$p_{ij} = \beta_{0j} + n_{ij}$$

The error at this second level of analysis is represented by n, which has a mean of 0 and a variance of τ^2. Because of potentially correlated error terms (with multiple correlations from the same sample), we modeled B_{0j} as a random effect:

$$\beta_{0j} = \delta_0 + \varepsilon_j$$

The test–retest reliability is represented by; $_j$has a mean of 0 and a variance of τ^2.

Our subsequent models challenged the null hypothesis and provided tests of our hypotheses. In both of our hypothesized models (below), the Level-1 and Level-3 formulae were the same, and test–retest reliability was captured by δ_0; the second-level formulae, though, is different.

As predicted by Hypothesis 1, there should be some level of test–retest unreliability (Hypothesis 1a), p_{ij} should decrease with time (Hypothesis 1b), but the decrease should be nonlinear so that it does not directly approach zero (Hypothesis 1c). We thus provided an initial test of Hypothesis 1 through the following Level-2 model:

$$\mathrm{p_{ij}} = \beta_{0j} + \left(\beta_1 \times \mathrm{Time_{ij}}\right) + \left(\beta_2 \times \mathrm{Time_{ij}}^2\right) + \mathrm{n_{ij}}(\mathrm{Model\ 2})$$

In our subsequent hypotheses, we predicted that job complexity and measurement type would affect the level of temporal consistency. In Hypotheses 2 and 3, we predicted that measurement type and complexity would affect the level of test–retest reliability (_$_{0j}$). These hypotheses thus suggest the inclusion of main effects for these two variables. In Hypothesis 4, we predicted that job complexity would moderate stability and thus the effect of time in Model 2. We tested this by examining the interaction of complexity and time. Model 3 is as follows:

$$\begin{aligned}\mathrm{p_{ij}} = {} & \beta_{0j} + \left(\beta_1 \times \mathrm{Time_{ij}}\right) + \left(\beta_2 \times \mathrm{Time_{ij}}^2\right) + \left(\beta_3 \times \mathrm{Measurement\ Type_{ij}}\right) + \left(\beta_4 \times \mathrm{Complexity_{ij}}\right) \\ & + \left[\beta_5 \times \left(\mathrm{Time_{ij}} \times \mathrm{Complexity_{ij}}\right)\right] + \left[\beta_6 \times \left(\mathrm{Time_{ij}}^2 \times \mathrm{Complexity_{ij}}\right)\right] + \mathrm{n_j}(\mathrm{Model\ 3})\end{aligned}$$

Model 3 represents our full model for the consistency of job performance over time. From this model, test–retest reliability was approximated by the sum of the Level-3 intercept and the Level-2 main effects ($\delta_0 + \beta_3 \times \mathrm{Measurement\ Type} + \beta_4 \times \mathrm{Complexity}$). The (lack of) stability of job performance is revealed by the Level-2 parameters associated with time.

Literature Search and Study Characteristics

A search was conducted for articles in which researchers studied individual performance over three or more time periods. We did not include studies that used athletes (i.e., Henry & Hulin, 1987; Hofmann et al., 1992) or student grades, as the type of performance from these samples are less generalizable to more common forms of employment. The search involved four different strategies. First, articles that reviewed literature on dynamic performance were used as a source of potential studies (Barrett et al., 1985; Hulin, Henry, & Noon, 1990; Rambo et al., 1983). Second, a manual search was conducted of the following management and marketing journals: *Journal of Applied Psychology*, *Academy of Management Journal*, *Administrative Science Quarterly*, *Personnel Psychology*, *Organizational Behavior and Human Decision Processes*, *Journal of Management*, *Human Resource Management*, *Human Relations*, *Journal of Marketing*, and *Journal of the Academy of Marketing Science.*

This search was from 1980 to 2003. Third, a computer search was conducted with ABI/INFORM, which contains abstracts–articles for business and psychological research. Fourth, unpublished articles (or other articles not revealed through our prior search) were requested through Human Resource Division Net and Research Methods Net, both electronic mailing lists associated with their respective divisions of the Academy of Management Association.

From all the above sources, a total of 22 articles were obtained. A summary of the sample and characteristics of each of the studies is provided in Table 1. From each study, the following variables were included in the meta-analysis: sample size for each time interval, correlations for performance at each time interval, time interval between performance measures (in years), type of performance measure used within each study (i.e., objective or subjective), and job complexity. Measurement type was coded as 1 if the study used objective measures of performance; otherwise, it was set equal to 0 (for supervisory ratings).

For each sample, the job being investigated was recorded, and a measure of job complexity was estimated. Similar to previous studies (Farrell & McDaniel, 2001; Sturman, 2003), we used Hunter's (1983) complexity scale, which is based on the Data and Things dimension of the *Dictionary of Occupational Titles* (DOT; U.S. Department of Labor, 1991). Gandy (1986) questioned the reliability of the Things dimension reported in the DOT; therefore, we replicated the measurement method and the practice of using only the Data dimension used in previous studies (e.g., Farrell & McDaniel, 2001; Sturman, 2003). The Data scale ranges from 0 (high complexity) to 6 (low complexity); we reverse coded it (and added 1) so that 7 represented high complexity and 1 represented low complexity. We did this to facilitate the interpretation of our results so that higher values would represent higher complexity. To further enhance the interpretability of our analyses, we centered complexity on the basis of its grand mean (raw score M = 4.19; SD = 2.11). Therefore, the interpretation of our intercept was based on the idea of a job with average complexity, rather than a job with theoretically no complexity (which does not exist, as our minimum score is 1).

We also examined the situations described in each of the 22 studies. For most studies (k = 17), there were no particularly noteworthy events that occurred during the span of the study. That is, the studies were focused on researching individuals over time, and no mention was made of any intervention (or event) that would suggest a major disruption to performance. However, there were some exceptions (k = 5). For example, Griffin (1991) examined the effects of work redesign. In this study, data were purposely collected before and after the redesign occurred. Harrison et al. (1996) examined employee performance when employees' pay system changed from a flat amount and commission (during their first 2 months of service) to a purely commission-based system. In Tziner, Ronen, and Hacohen's (1993) study, the researchers conducted the study after an assessment center was set up to evaluate, develop, and promote employees to upper management levels. Finally, Warr and Bunce (1995) collected data prior to and after a 4-month, self-paced, open-training program. We point these studies out to provide a complete picture of our metaanalysis. Our investigations, though, suggest that the relationships across performance scores in these studies were not significantly different from the relationships found in the rest of our sample. There were no significant differences for this set of studies compared with the others with regard to the measurement type, job complexity, length of time examined in the longitudinal study, sample size, or mean correlation (corrected or uncorrected). We also examined residual scores of these five studies and compared them with the distribution of residuals from the 17 other studies but found no significant differences. We argue that the changes examined in these five studies (i.e., training, work redesign, pay system changes) are all types of changes that frequently occur over time as the natural course of work, and thus we did not expect these five studies to be notably different. Nonetheless, it is important to point out these situations so that readers can draw their own conclusion on the validity and generalizability of our results.

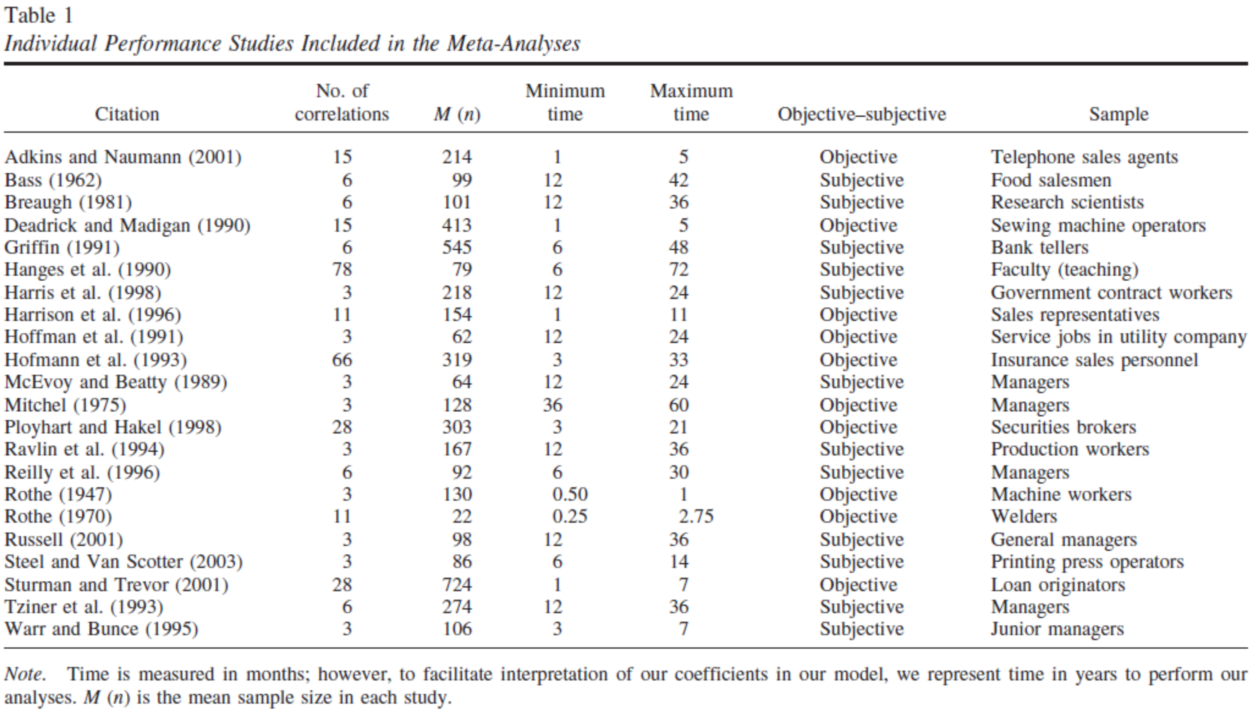

Table 1
Individual Performance Studies Included in the Meta-Analyses

Citation	No. of correlations	*M* (*n*)	Minimum time	Maximum time	Objective–subjective	Sample
Adkins and Naumann (2001)	15	214	1	5	Objective	Telephone sales agents
Bass (1962)	6	99	12	42	Subjective	Food salesmen
Breaugh (1981)	6	101	12	36	Subjective	Research scientists
Deadrick and Madigan (1990)	15	413	1	5	Objective	Sewing machine operators
Griffin (1991)	6	545	6	48	Subjective	Bank tellers
Hanges et al. (1990)	78	79	6	72	Subjective	Faculty (teaching)
Harris et al. (1998)	3	218	12	24	Subjective	Government contract workers
Harrison et al. (1996)	11	154	1	11	Objective	Sales representatives
Hoffman et al. (1991)	3	62	12	24	Objective	Service jobs in utility company
Hofmann et al. (1993)	66	319	3	33	Objective	Insurance sales personnel
McEvoy and Beatty (1989)	3	64	12	24	Subjective	Managers
Mitchel (1975)	3	128	36	60	Objective	Managers
Ployhart and Hakel (1998)	28	303	3	21	Objective	Securities brokers
Ravlin et al. (1994)	3	167	12	36	Subjective	Production workers
Reilly et al. (1996)	6	92	6	30	Subjective	Managers
Rothe (1947)	3	130	0.50	1	Objective	Machine workers
Rothe (1970)	11	22	0.25	2.75	Objective	Welders
Russell (2001)	3	98	12	36	Subjective	General managers
Steel and Van Scotter (2003)	3	86	6	14	Subjective	Printing press operators
Sturman and Trevor (2001)	28	724	1	7	Objective	Loan originators
Tziner et al. (1993)	6	274	12	36	Subjective	Managers
Warr and Bunce (1995)	3	106	3	7	Subjective	Junior managers

Note. Time is measured in months; however, to facilitate interpretation of our coefficients in our model, we represent time in years to perform our analyses. *M* (*n*) is the mean sample size in each study.

Implementing the Meta-Analysis

The first step of implementing the meta-analysis was to decide whether or not to correct the observed correlations for statistical artifacts (i.e., intrarater reliability and range restriction). Although some strongly argue for such corrections (e.g., Hunter & Schmidt, 1990; Schmidt & Hunter, 1996), there are important methodological considerations for our analyses. Most methods of random-effects meta-analysis described elsewhere (cf. Bryk & Raudenbush, 1992; Erez et al., 1996) transform the correlations with Fisher's r-to-Z transformation. This has the advantage of putting the correlations into a form that is more normally distributed and with a stable variance. Unfortunately, there is limited research on the use of Fisher's r-to-Z transformation after coefficients have been corrected for statistical artifacts. The only example we found was by Mendoza (1993), who showed that the computation for variance after a correlation has been corrected for range restriction is far more complex than the simple formula associated with the regular transformation. There is no research that has examined the effects of Fisher's r-to-Z transformation for correlations corrected for unreliability, or unreliability and range restriction.

Because random-effects meta-analysis requires that Level-1 variances be known, and with the lack of research on the correct method of estimating this variance after Fisher's r-to-Z transformation, we did not feel it would be appropriate to use the transformation on corrected coefficients. However, given that one of the main goals of our research was to provide an accurate estimate of the amount of unreliability and stability in job performance over time, we also felt it was critical that we yield as accurate an estimate of the true correlation as possible, and thus felt we needed to correct correlations for statistical artifacts whenever we could. Therefore, we decided to estimate Level-1 variances on the basis of the formulae that exist for estimating the variance of corrected correlation coefficients (see Raju & Brand, 2003; Raju et al., 1991).

Internal consistency ratings for supervisory evaluations were obtained from our set of studies whenever possible (internal consistencies ranged from .54 to .95; M = .89, SD = .06; 90% of values were greater than .86); however, two studies (Bass, 1962; Harris et al., 1998) did not include reliabilities for their subjective performance measures. We, therefore, used the estimate of average intrarater reliability (α= .86) for the missing data from Viswesvaran et al.'s (1996) study. Note that we used the measure of intrarater reliability and not interrater reliability because the measures were of a single individual's (supervisor's) evaluation and not from a set of different judges' ratings. No correction for unreliability was made for objective performance measures.

Additionally, a decision needed to be made whether to correct the correlations for range restriction. There is evidence that performance is related to turnover (e.g., Trevor, Gerhart, & Boudreau, 1997; Williams & Livingstone, 1994) and that turnover creates range restriction when considering performance longitudinally (Sturman & Trevor, 2001). Given our desire to estimate the true test–retest reliability and stability of job performance, it would be desirable to correct the correlations for this effect. To make this correction, we would need information on (a) the relationship between performance and turnover and (b) the rate of turnover. Unfortunately, with the exception of Sturman and Trevor's (2001) study, none of the studies provided any information on this performance–turnover relationship; furthermore, only a few of the studies provided data on the number of employees in each time period (most of the studies simply used list-wise deletion and reported only the final sample size). Thus, it would be very difficult to provide any accurate estimate of range restriction. Furthermore, as shown by Trevor et al. (1997), the relationship between performance and turnover is nonlinear and moderated by the strength of the link between pay and performance. Again, very few of our studies provided any information on the nature of the studies' compensation systems. To further compound our difficulties, we were unaware of any methodological research that describes how to correct range restriction (a) when there is double range restriction, (b) when the restriction is based on an unmeasured third variable, and (c) when the relationship between our predictor and the third variable is nonlinear. Thus, we did not correct for range restriction because we did not have confidence that any estimate we produced would necessarily be accurate. We discuss this limitation in the conclusion of our study.

With the correlation coefficients corrected for unreliability and their variances computed, we proceeded with the performance of our metaanalysis. As noted above, because our analytic strategy required several correlations from each sample, our meta-analytic method needed to account for the possibility that there existed correlated sampling errors within each study. Using Model 2 to illustrate, we fit the following model:

$$(\text{Level } 1\ )\ \mathrm{r_{ij}} = \mathrm{p_{ij}} + \mathrm{e_{ij}} \qquad \mathrm{e} \rightarrow \mathrm{N}(0, \sigma^2)$$

$$(\text{Level } 2)\ \mathrm{p_{ij}} = \beta_{0j} + \left(\beta_1 \times \mathrm{Time_{ij}}\right) + \left(\beta_2 \times \mathrm{Time_{ij}}^2\right) + \mathrm{n_j} \qquad \mathrm{n} \rightarrow \mathrm{N}(0, \tau^2)$$

$$(\text{Level } 3)\ \beta_{0j} = \delta_0 + \varepsilon_j \qquad \varepsilon \rightarrow \mathrm{N}(0, \mathrm{v}^2).$$

The subscripts are used to represent correlation *i* of study *j*. The error terms—$\mathrm{e_{ij}}$, $\mathrm{n_j}$, and ε_j—are assumed to follow a normal distribution with a mean of 0 and with variances equal to σ^2,τ^2, and v^2, respectively. The formulae for Level 1 and Level 3 are the same for all three models; the Level-2 model is given by Models 1–3.

The specific methods for performing the meta-analysis are described in detail elsewhere (e.g., Bryk & Raudenbush, 1992; Erez et al., 1996). In brief, they entail researchers using maximum likelihood estimation to approximate the βs, τ^2, and v^2. Because of the nature of this data set, the statistical algorithm for implementing the meta-analysis was created for the purposes of this study. We used Visual Basic tied into an interface from Microsoft Excel. The program was verified by comparing meta-analyses without covariates with the results of HLM (4.0) output and by replicating the results reported in Erez et al.'s (1996) study. Although the user interface is primitive, the program is available from Michael C. Sturman on request.

Results

Table 2 presents the results of the tests from our series of models. As noted above, Model 1 is our null hypothesis and a useful starting point with which to make comparisons. This model is analogous to estimating a single correlation in more traditional meta-analyses. We also used this case to perform a test for homogeneity to determine whether our search for moderators was merited (Hedges & Olkin, 1985; Hunter & Schmidt, 1990). Indeed, this was strongly supported ($p < .0001$). Thus, we continued on to test Models 2 and 3.

Model 2 presents a test of our first hypothesis and supports our predictions. As expected, the estimate of test–retest reliability (δ_0) was positive, fairly large, but less than 1.0. Its 95% confidence interval ranged from .53 to .61. The results also show that as the time lag between measures increased, the correlation between those measures decreased. This decrease, though, was nonlinear. The quadratic term for time was significant ($p < .05$) and its magnitude and direction indicated that the curve did not approach zero. Thus, Hypothesis 1 (i.e., Hypotheses 1a–1c) was supported. Additionally, Model 2 explained 18% of the Level-2 variance and, with a log-likelihood test, was significantly more predictive than Model 1 ($p < .0001$).

Model 3 presents the test of the full model. Here, we were interested in the estimate of test–retest reliability ($\delta_0 + \beta_3 \times \text{Measurement Type} + \beta_4 \times \text{Complexity}$) and the effects associated with time $[\beta_1 \times \text{Time} + \beta_2 \times \text{Time}^2 + \beta_5 \times (\text{Time} \times \text{Complexity}) + \beta_6 \times (\text{Time}^2 \times \text{Complexity})]$. This model's results, shown in the right-most column of Table 2, support Hypotheses 2–4. Specifically, objective measures of performance were associated with lower test–retest reliability ($\beta_3 = -.22, \mathrm{p} < .0001$, greater complexity was associated with lower test–retest reliability ($\beta_4 = -.03, \mathrm{p} < .001$), and complexity moderated the relationships with time (both at $\mathrm{p} < .01$). Again, the third model was significantly more predictive than the second model (log-likelihood test: $p < .0001$); similarly, the model explained 40% of the Level-2 variance in the correlation coefficients, compared with 18% in Model 2. Additionally, when we replicated our analyses without correcting for unreliability (available from

Michael C. Sturman on request) all the coefficients remained statistically significant; thus, our finding for lower test–retest reliability for objective performance measures is not attributable to our correction for a lack of internal consistency in the subjective measures.

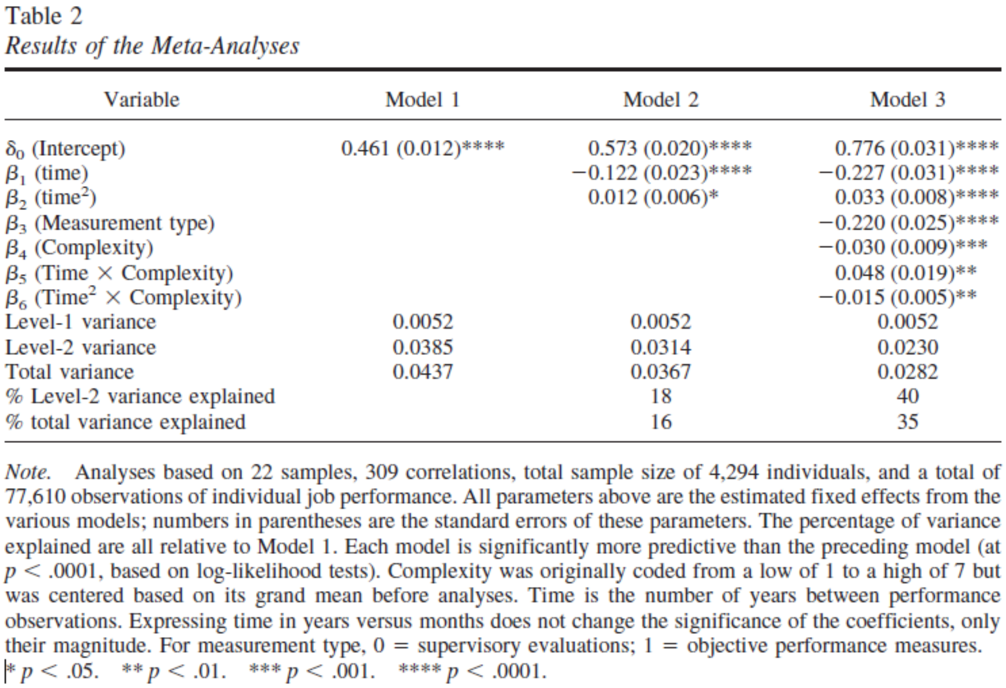

Table 2
Results of the Meta-Analyses

Variable	Model 1	Model 2	Model 3
δ_0 (Intercept)	0.461 (0.012)****	0.573 (0.020)****	0.776 (0.031)****
β_1 (time)		−0.122 (0.023)****	−0.227 (0.031)****
β_2 (time^2)		0.012 (0.006)*	0.033 (0.008)****
β_3 (Measurement type)			−0.220 (0.025)****
β_4 (Complexity)			−0.030 (0.009)***
β_5 (Time × Complexity)			0.048 (0.019)**
β_6 (Time^2 × Complexity)			−0.015 (0.005)**
Level-1 variance	0.0052	0.0052	0.0052
Level-2 variance	0.0385	0.0314	0.0230
Total variance	0.0437	0.0367	0.0282
% Level-2 variance explained		18	40
% total variance explained		16	35

Note. Analyses based on 22 samples, 309 correlations, total sample size of 4,294 individuals, and a total of 77,610 observations of individual job performance. All parameters above are the estimated fixed effects from the various models; numbers in parentheses are the standard errors of these parameters. The percentage of variance explained are all relative to Model 1. Each model is significantly more predictive than the preceding model (at $p < .0001$, based on log-likelihood tests). Complexity was originally coded from a low of 1 to a high of 7 but was centered based on its grand mean before analyses. Time is the number of years between performance observations. Expressing time in years versus months does not change the significance of the coefficients, only their magnitude. For measurement type, 0 = supervisory evaluations; 1 = objective performance measures.
* $p < .05$. ** $p < .01$. *** $p < .001$. **** $p < .0001$.

To facilitate the interpretation of our results, we computed the expected test–retest reliability for different hypothetical groups, in addition to the temporal consistency and stability of performance scores for these groups at a number of different time lags. These results are reported in Table 3. The four groups were for the following conditions: (a) an objective measure, high-complexity (measured at one standard deviation above the mean of complexity ratings) group, (b) a subjective measure, high-complexity group, (c) an objective measure, low-complexity (measured at one standard deviation below the mean of complexity ratings) group, and (d) a subjective measure, low-complexity group. These estimates were derived from Model 3 of Table 2. The time lags were chosen as values of potential practical interest (0.50 years, 1 year, 2 years, and 3 years). To have stable estimates, we do not provide predicted values for any time lag exceeding the 90th percentile of time spans from the original set of studies (i.e., 36 months). The 95% confidence intervals are also reported for each of these estimates.

To compute the estimated performance stability, we used the test–retest reliability values reported at the top of Table 3 and corrected the original correlations for this amount of error (and recomputed the resultant variance). We then replicated our analyses (available from Michael C. Sturman on request) using Model 3. As we expected, the level of test–retest reliability for each group was near

1.0 (none differed significantly from 1.0), and the effect of measurement type was not significant. The other coefficients of Model 3 remained significant at $p < .01$. We used this model to estimate the level of stability at the same time lags as in the middle of Table 3, and we report the expected values and their 95% confidence intervals.

Table 3 illustrates the support of our hypotheses. Clearly, temporal consistency and stability decrease with greater time lags but do not reach values whose confidence intervals include zero. Table 3 also shows the magnitude of the effects associated with measurement type and complexity.

Discussion

Studying any phenomenon longitudinally is difficult because statistical analyses have “difficulty determining true changes from error” (Mitchell, 1997, p. 126). Our study facilitates the study of job performance over time by examining the causes of performance dynamism and thus allowing the differentiation of true changes (a lack of stability) from error (a lack of test–retest reliability). Our results show that it is impossible to estimate a single true stability of job performance, as this value is contingent on the time interval being conceptualized and job complexity. Nonetheless, the present study yields a model that can be used to estimate the level of temporal consistency, stability, and test–retest reliability for a wide variety of circumstances.

Ultimately, this study makes a number of contributions for a variety of literature streams. Most obviously, our intent was to contribute to the research on dynamic performance. Research on dynamic performance has evolved from studies simply showing the existence of dynamic criteria to research investigating the causes and consequences of job performance over time. An understanding of the nature of job performance over time is critical to allow such investigations to continue. Our study’s findings on the amount of job performance’s stability and test–retest reliability should help future researchers interpret and understand findings associated with performance changes over time.

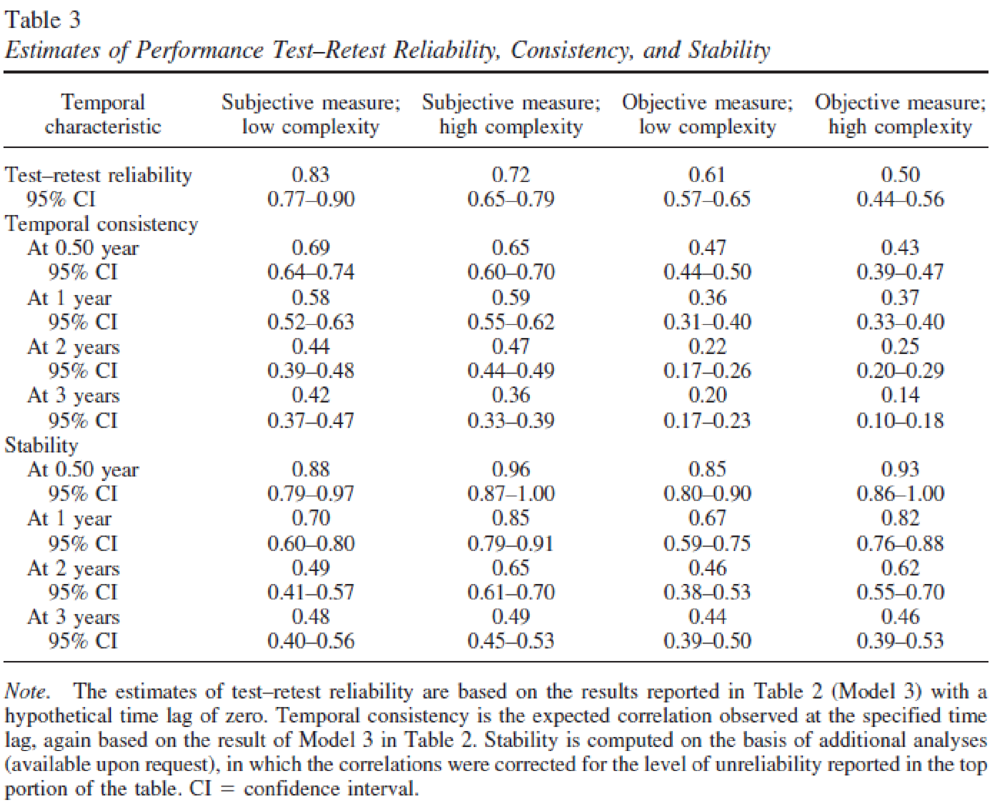

Table 3

Estimates of Performance Test–Retest Reliability, Consistency, and Stability

Temporal characteristic	Subjective measure; low complexity	Subjective measure; high complexity	Objective measure; low complexity	Objective measure; high complexity
Test–retest reliability	0.83	0.72	0.61	0.50
95% CI	0.77–0.90	0.65–0.79	0.57–0.65	0.44–0.56
Temporal consistency				
At 0.50 year	0.69	0.65	0.47	0.43
95% CI	0.64–0.74	0.60–0.70	0.44–0.50	0.39–0.47
At 1 year	0.58	0.59	0.36	0.37
95% CI	0.52–0.63	0.55–0.62	0.31–0.40	0.33–0.40
At 2 years	0.44	0.47	0.22	0.25
95% CI	0.39–0.48	0.44–0.49	0.17–0.26	0.20–0.29
At 3 years	0.42	0.36	0.20	0.14
95% CI	0.37–0.47	0.33–0.39	0.17–0.23	0.10–0.18
Stability				
At 0.50 year	0.88	0.96	0.85	0.93
95% CI	0.79–0.97	0.87–1.00	0.80–0.90	0.86–1.00
At 1 year	0.70	0.85	0.67	0.82
95% CI	0.60–0.80	0.79–0.91	0.59–0.75	0.76–0.88
At 2 years	0.49	0.65	0.46	0.62
95% CI	0.41–0.57	0.61–0.70	0.38–0.53	0.55–0.70
At 3 years	0.48	0.49	0.44	0.46
95% CI	0.40–0.56	0.45–0.53	0.39–0.50	0.39–0.53

Note. The estimates of test–retest reliability are based on the results reported in Table 2 (Model 3) with a hypothetical time lag of zero. Temporal consistency is the expected correlation observed at the specified time lag, again based on the result of Model 3 in Table 2. Stability is computed on the basis of additional analyses (available upon request), in which the correlations were corrected for the level of unreliability reported in the top portion of the table. CI = confidence interval.

Furthermore, our results provide support for Ackerman's model, as applied to longitudinal investigations of performance. Although this model has been receiving empirical support lately (e.g., Farrell & McDaniel, 2001; Keil & Cortina, 2001), our study provides support for applying the model to employee job performance ratings over time. Additionally, our results provide support for Murphy's application of Ackerman's model to field settings. An important addition of our study, though, is that we provide evidence on the functional form of job performance over time. Our finding that there is a stable component of job performance is an important confirmation of Murphy's hypothesis that certain characteristics (e.g., cognitive ability) will continue to play an important role when predicting job performance. However, the predictive ability of prior job performance ratings is still higher than the predictive power of cognitive ability ratings, even over time. This suggests that there are other characteristics that explain some of this stable performance component. Perhaps, in addition to cognitive ability, this includes the personality characteristics studied by Judge et al. (1999). There may certainly be other job or individual characteristics that may prove useful for predicting performance over time. More research is needed to help understand the causes of performance and consideration needs to be given to the implications of these factors in a longitudinal context.

Beyond our contribution to the domain of dynamic performance, this study also yields findings that can inform a number of other content domains. By providing an estimate of test–retest reliability, this study contributes to the growing literature aimed at understanding the construct of performance. With information now available on the magnitude of different types of job performance reliability (i.e., intrarater reliability, interrater reliability, and test– retest reliability), the use of such estimates (particularly if the intent is to correct an estimate for attenuation) should depend on the needs of the researcher or practitioner, the nature of the question being asked, and the characteristics of the data. Viswesvaran et al. (1996) noted that if one needs an answer to the question "Would the same ratings be obtained if a different but equally knowledgeable judge rated the same employees?" (p. 565), then interrater reliability is the appropriate estimate for making corrections for criterion unreliability in validation research. This is consistent with generalizability theory, which argues that there is not one value representing the reliability of a construct but different sources of error variance whose importance depends on the nature of the generalization being made (Murphy & De Shon, 2000). For longitudinal research involving performance, and particularly for research examining the extent to which performance over time can be predicted, being able to partial out attenuation due to a lack of test–retest reliability may be most relevant.

Ultimately, this stream of research will contribute to the wide range of studies that consider any number of the relationships associated with job performance. Support for our hypothesis that objective measures of performance are less reliable over time is counter to the notion that objective measures have some inherent advantage in research. Our rationale for the greater test–retest reliability of subjective measures, however, draws attention to the idea that high reliability, when assessed only through a single measure, does not necessarily connote a complete lack of error variance.

For practice, these results provide a means to estimate the predictability of performance. Although it is widely accepted that past performance is the best predictor of future performance, actual validity data, let alone meta-analytic results, have been lacking. It is no surprise that such a correlation, say over a 1-year time lag, would be statistically significant; however, it is often valuable to have a precise estimate of the validity of such a predictive process, such as to help inform the evaluation of internal selection systems. Using our results and correcting for test–retest unreliability, researchers can make estimates of the correlation between performance scores over various time lags. The results in Table 3 show these values at a variety of hypothetical time lags. The difference between the uncorrected correlations and the predicted values demonstrates the significance of our findings and their importance for those interested in making longitudinal predictions of performance, be it for theoretical or practical reasons.

Our results can also help inform practitioners making decisions on the type of performance data to collect. For example, we feel our findings suggest that objective measures of performance may not be very useful in highly complex jobs because of the lack of test–retest reliability in this context. Although it is not our intent to recommend a minimum cutoff for test–retest reliability, our study does provide estimates on the level of test–retest reliability that can help describe the quality of performance measures. This information will allow practitioners and researchers to interpret and use performance data more appropriately. More holistically, our results suggest that performance measures should be chosen carefully, as there are notable consequences—in terms of what exactly is measured and how much error is associated with that measurement— that accompany such choices.

The conclusions from our results, though, must be tempered by a number of limitations. Most notably, our estimate of test–retest reliability is based on synthesizing a number of studies and methodologically partialing out variance attributable to different covariates. The difficulty with this approach is that we have no direct control over the circumstances. A more structured approach to assessing test–retest reliability would be to perform a well-controlled study in which performance stability could be assessed more directly. Of course, this is also problematic in that such experiments often lose the generalizability associated with field studies. Nonetheless, it should be noted that our methodology was developed in part to assess the test–retest reliability in previous studies.

Another limitation of our study involves our dependence on the extent of available data. Because of limitations in the type and amount of information provided in each study and the present state of the art in methods, we could not correct the correlations for range restriction. Because range restriction would attenuate the observed correlations, our estimate of test–retest reliability may overstate the actual amount of error attributed to this specific methodological artifact. Sturman and Trevor (2001) showed that turnover causes range restriction when examining job performance over time; it would be valuable for future researchers to investigate how range restriction specifically influences the stability and temporal consistency of job performance ratings. Ideally, this will lead to research that prescribes how to correct for such attenuation. We wish to note, though, that although the amount of error represented by δ_0 may overestimate the amount of test–retest reliability, it still is an accurate estimate of the amount of measurement error attributable to methodological artifacts. That is, both range restriction and unreliability would lower the estimated intercept. Our correction for this factor is thus an appropriate estimate of attenuation, and thus we have confidence on the appropriateness of our estimates of stability and our estimate of unreliability at least as a measure of overall measurement error. Nonetheless, it would be valuable to know exactly how much of this attenuation is attributable to test–retest unreliability versus range restriction.

It would also be valuable for future researchers to consider the potential moderating effects of additional job characteristics. Like many other studies, we focused exclusively on job complexity; however, it is likely that considering additional job characteristics could help explain the level of performance stability. Our study is also limited in that we compared different methods of performance evaluation but did not consider the multidimensional nature of job performance. We purposely chose to focus our study on the moderating effects of measurement. Although our results are informative for research on dynamic performance, performance prediction, and performance assessment, our study makes little contribution to an understanding of the construct domain captured by performance measures. When understanding the nature of job performance, it would be valuable for future researchers to consider the longitudinal nature of the various dimensions of performance—task performance, citizenship performance, counterproductive performance (Rotundo & Sackett, 2002)—individually. To date, there has been very little longitudinal research on organizational citizenship behaviors. Given the attention that dynamic performance has received historically, and the recognition of citizenship behaviors as a critical dimension of job performance, more longitudinal research in this area is critical.

In sum, this study goes beyond simply supporting the idea that performance is dynamic to (a) make a point to differentiate between three key concepts: temporal consistency, stability, and test–retest reliability; (b) predict moderators to these relationships; and (c) provide estimates of these effects. The findings from this study contribute to the literature on dynamic performance and provide more insight into the nature of individual job performance ratings.

References

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

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometrics and information processing perspectives. *Psychological Bulletin*, *102*, 3–27.

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General*, *177*, 288–318.

Ackerman, P. L. (1989). Within-task intercorrelations of skilled performance: Implications for predicting individual differences? *Journal of Applied Psychology*, *74*, 360–364.

*Adkins, C. L., & Naumann, S. E. (2001). Situational constraints on the achievement–performance relationship: A service sector study. *Journal of Organizational Behavior*, *22*, 453–465.

Alvares, K. M., & Hulin, C. L. (1972). Two explanations of temporal changes in ability–skill relationships: A literature review and theoretical analysis. *Human Factors*, *14*, 295–308.

Alvares, K. M., & Hulin, C. L. (1973). An experimental evaluation of a temporal decay in the prediction of performance. *Organizational Behavior and Human Performance*, *9*, 169–185.

Austin, J. T., Humphreys, L. G., & Hulin, C. L. (1989). A critical reanalysis of Barrett, Caldwell, and Alexander. *Personnel Psychology*, *42*, 583–596.

Barrett, G. V., & Alexander, R. A. (1989). Rejoinder to Austin, Humphreys, and Hulin: Critical reanalysis of Barrett, Caldwell, and Alexander. *Personnel Psychology*, *42*, 597–612.

Barrett, G. V., Caldwell, M. S., & Alexander, R. A. (1985). The concept of dynamic criteria: A critical reanalysis. *Personnel Psychology*, *38*, 41–56.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*, 1–26.

*Bass, B. M. (1962). Further evidence on the dynamic character of criteria. *Personnel Psychology*, *15*, 93–97.

Bayley, N. (1955). On the growth of intelligence. *American Psychologist*, *10*, 805–818.

Bazerman, M. H. (1998). *Judgment in managerial decision making* (4th ed.). New York: Wiley.

Becker, G. (2000). How important is transient error in estimating reliability? Going beyond simulation studies. *Psychological Methods*, *5*, 370– 379.

Blumberg, M., & Pringle, C. D. (1982). The missing opportunity in organizational research: Some implications for a theory of work performance. *Academy of Management Review*, *7*, 560–569.

Bommer, W. H., Johnson, J. L., Rich, G. A., Podsakoff, P. M., & Mac- Kenzie, S. B. (1995). On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. *Personnel Psychology*, *48*, 587–605.

*Breaugh, J. A. (1981). Predicting absenteeism from prior absenteeism and work attitudes. *Journal of Applied Psychology*, *66*, 555–560.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 1, pp. 687–732). Palo Alto, CA: Consulting Psychologists Press.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment.* Newbury Park, CA: Sage.

Cascio, W. F. (1998). *Applied psychology in human resource management* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Conway, J. M. (1999). Distinguishing contextual performance from task performance for managerial jobs. *Journal of Applied Psychology*, *84*, 3–13.

Costa, P. T., Jr., & McCrae, R. R. (1988). Personality in adulthood: A six-year longitudinal study of self-reports and spouse ratings on the NEO Personality Inventory. *Journal of Personality and Social Psychology*, *54*, 853–863.

Costa, P. T., Jr., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences*, *13*, 653–665.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Deadrick, D. L., Bennett, N., & Russell, C. J. (1997). Using hierarchical linear modeling to examine dynamic performance criteria over time. *Journal of Management*, *23*, 745–757.

*Deadrick, D. L., & Madigan, R. M. (1990). Dynamic criteria revisited: A longitudinal study of performance stability and predictive validity. *Personnel Psychology*, *43*, 717–744.

De Shon, R. P., Ployhart, R. E., & Sacco, J. M. (1998). The estimation of reliability in longitudinal models. *International Journal of Behavioral Development*, *22*, 493–515.

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people most of the time. *Journal of Personality and Social Psychology*, *37*, 1097–1126.

Erez, A., Bloom, M. C., & Wells, M. T. (1996). Using random rather than fixed effects models in meta-analysis: Implications for situational specificity and validity generalization. *Personnel Psychology*, *49*, 275–306.

Farrell, J. N., & McDaniel, M. A. (2001). The stability of validity coefficients over time: Ackerman's (1988) model and the general aptitude battery. *Journal of Applied Psychology*, *86,* 60–79.

Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology*, *66*, 127–148.

Fisher, R. A. (1932). *Statistical methods for research workers.* London: Oxford University Press.

Gandy, J. A. (1986, June). *Job complexity, aggregated samples, and aptitude test validity: Meta-analysis of the General Aptitude Test Battery Data Base.* Paper presented at the meeting of the International Personnel Management Association, San Francisco.

Ghiselli, E. E. (1956). Dimensional problems of criteria. *Journal of Applied Psychology*, *40*, 1–4.

Ghiselli, E. E., & Haire, M. (1960). The validation of selection tests in the light of the dynamic character of criteria. *Personnel Psychology*, *13*, 225–231.

Green, S. B. (2003). A coefficient alpha for test–retest data. *Psychological Methods*, *8*, 88–101.

*Griffin, R. W. (1991). Effects of work redesign on employee perceptions, attitudes, and behaviors: A long-term investigation. *Academy of Management Journal*, *34*, 425–435.

Gutenberg, R. L., Arvey, R. D., Osburn, H. G., & Jeanneret, P. R. (1983). Moderating effects of decision-making/information-processing job dimensions on test validities. *Journal of Applied Psychology*, *68*, 602– 608.

*Hanges, P. J., Schneider, B., & Niles, K. (1990). Stability of performance: An interactionist perspective. *Journal of Applied Psychology*, *75*, 658– 667.

*Harris, M. M., Gilbreath, B., & Sunday, J. A. (1998). A longitudinal examination of a merit pay system: Relationships among performance ratings, merit increases, and total pay increases. *Journal of Applied Psychology*, *83*, 825–831.

*Harrison, D. A., Virick, M., & William, S. (1996). Working without a net: Time, performance, and turnover under maximally contingent rewards. *Journal of Applied Psychology*, *81*, 331–345.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* San Diego, CA: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504.

Heise, D. R. (1969). Separating reliability and stability in test–retest correlation. *American Sociological Review*, *34*, 93–101.

Heneman, R. L. (1986). The relationship between supervisory ratings and results-oriented measures of performance: A meta-analysis. *Personnel Psychology*, *39*, 811–826.

Henry, R. A., & Hulin, C. L. (1987). Stability of skilled performance across time: Some generalizations and limitations on utility. *Journal of Applied Psychology*, *72*, 457–462.

*Hoffman, C. C., Nathan, B. R., & Holden, L. M. (1991). A comparison of validation criteria: Objective versus subjective performance measures and self- versus supervisor ratings. *Personnel Psychology*, *44*, 601–619.

*Hofmann, D. A., Jacobs, R., & Baratta, J. E. (1993). Dynamic criteria and the measurement of change. *Journal of Applied Psychology*, *78*, 194– 204.

Hofmann, D. A., Jacobs, R., & Gerras, S. J. (1992). Mapping individual performance over time. *Journal of Applied Psychology*, *77*, 185–195.

Hulin, C. L., Henry, R. A., & Noon, S. L. (1990). Adding a dimension: Time as a factor in the generalizability of predictive relationships. *Psychological Bulletin*, *107*, 328–340.

Humphreys, L. G. (1960). Investigations of the simplex. *Psychometrika*, *25*, 313–323.

Hunter, J. E. (1983). A causal analysis of cognitive ability, job knowledge, job performance, and supervisory ratings. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), *Performance measurement and theory* (pp. 257–266). Hillsdale, NJ: Erlbaum.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*, 72–98.

Hunter, J. E., & Schmidt, F. L. (1990). *Method of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Jackson, S. E., & Schuler, R. S. (1985). A meta-analysis and conceptual critique of research on role ambiguity and role conflict in work settings. *Organizational Behavior and Human Decision Processes*, *36*, 16–78.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

Judge, T. A., Higgins, C. A., Thoresen, C. J., & Barrick, M. R. (1999). The Big Five personality traits, general mental ability, and career success across the life span. *Personnel Psychology*, *52*, 621–652.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge, England: Cambridge University Press.

Keil, C. T., & Cortina, J. M. (2001). Degradation of validity over time: A test and extension of Ackerman's model. *Psychological Bulletin*, *127*, 673–697.

Kerlinger, F. N. (1986). *Foundations of behavioral research.* New York: Holt, Rinehart & Winston.

Lance, C. E., Teachout, M. S., & Donnelly, T. M. (1992). Specification of the criterion construct space: An application of hierarchical confirmatory factor analysis. *Journal of Applied Psychology*, *77*, 437–452.

McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988). Job experience correlates of job performance. *Journal of Applied Psychology*, *73*, 327– 330.

*McEvoy, G. M., & Beatty, R. W. (1989). Assessment centers and subordinate appraisals of managers: A seven-year examination of predictive validity. *Personnel Psychology*, *42*, 37–52.

Mendoza, J. L. (1993). Fisher transformations for correlations corrected for selection and missing data. *Psychometrika*, *58*, 601–615.

*Mitchel, J. O. (1975). Assessment center validity: A longitudinal study. *Journal of Applied Psychology*, *60*, 573–579.

Mitchell, T. R. (1997). Matching motivational strategies with organizational contexts. *Research in Organizational Behavior*, *19*, 57–149.

Mitchell, T. R., & James, L. W. (2001). Building better theory: Time and the specification of when things happen. *Academy of Management Review*, *26*, 530–547.

Motowidlo, S. J., & Van Scotter, J. R. (1994). Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology*, *79*, 475–480.

Muckler, F. A., & Seven, S. A. (1992). Selection performance measures: "Objective" versus "subjective" measurement. *Human Factors*, *34*, 441– 455.

Murphy, K. R. (1989). Is the relationship between cognitive ability and job performance stable over time? *Human Performance*, *2*, 183–200.

Murphy, K. R., & De Shon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology*, *53*, 873–900.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Oldham, G. R., & Cummings, A. (1996). Employee creativity: Personal and contextual factors at work. *Academy of Management Journal*, *39*, 607–634.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (randomeffects) models for meta-analysis tests of moderator variable effects. *Psychological Methods*, *3*, 354–379.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measure, design, and analysis: An integrated approach.* Hillsdale, NJ: Erlbaum.

Peters, L. H., & O'Connor, E. J. (1980). Situational constraints and work outcomes: The influence of a frequently overlooked construct. *Academy of Management Review, 5*, 391–397.

*Ployhart, R. E., & Hakel, M. D. (1998). The substantive nature of performance variability: Predicting interindividual differences in intraindividual performance. *Personnel Psychology, 51*, 859–901.

Raju, N. S., & Brand, P. A. (2003). Determining the significance of correlations corrected for unreliability and range restriction. *Applied Psychological Measurement, 27*, 52–71.

Raju, N. S., Burke, M. J., Normand, J., & Langlois, G. M. (1991). A new meta-analytic approach. *Journal of Applied Psychology, 76*, 432–446.

Rambo, W. W., Chomiak, A. M., & Price, J. M. (1983). Consistency of performance under stable conditions of work. *Journal of Applied Psychology, 68*, 78–87.

*Ravlin, E. C., Adkins, C. L., & Meglino, B. M. (1994, August). Organizational definition of performance and individual value orientation: Interactive effects on performance and absence. In T. Welchans (Chair), *New directions in fit: Interaction and process.* Symposium presented at the meeting of the Academy of Management, Dallas, TX.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than g. *Journal of Applied Psychology, 79*, 518–524.

*Reilly, R. R., Smither, J. W., & Vasilopoulos, N. L. (1996). A longitudinal study of upward feedback. *Personnel Psychology, 49*, 599–612.

*Rothe, H. F. (1947). Output rates among machine operators: Distributions and their reliability. *Journal of Applied Psychology, 31*, 484–489.

*Rothe, H. F. (1970). Output rates among welders: Productivity and consistency of following removal of a financial incentive system. *Journal of Applied Psychology, 54*, 549–551.

Rotundo, M., & Sackett, P. R. (2002). The relative importance of task, citizenship, and counterproductive performance to global ratings of job performance: A policy-capturing approach. *Journal of Applied Psychology, 87*, 66–80.

*Russell, C. J. (2001). A longitudinal study of top-level executive performance. *Journal of Applied Psychology, 86*, 560–573.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., & Fruyt, F. (2003). International validity generalization of GMA and cognitive abilities: A European community meta-analysis. *Personnel Psychology, 56*, 573– 605.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 scenarios. *Psychological Methods, 1*, 199– 223.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Metaanalyses of validity studies published between 1864 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37*, 407– 422.

Schwab, D. P. (1999). *Research methods for organizational studies.* Mahwah, NJ: Erlbaum.

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*, 956–970.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* Thousand Oaks, CA: Sage.

Steel, R. P., & Mento, A. J. (1986). Impact of situational constraints on subjective and objective criteria of managerial job performance. *Organizational Behavior and Human Decision Processes, 37*, 254–265.

*Steel, R. P., & Van Scotter, J. R. (2003). The organizational performance cycle: Longitudinal assessment of key factors. *Journal of Business and Psychology, 18*, 31–50.

Sturman, M. C. (2003). Searching for the inverted U-shaped relationship between time and performance: Meta-analyses of the experience/performance, tenure/performance, and age/performance relationships. *Journal of Management, 29*, 609–640.

*Sturman, M. C., & Trevor, C. O. (2001). The implications of linking the dynamic performance and turnover literatures. *Journal of Applied Psychology*, *86*, 684–696.

Trevor, C. O., Gerhart, B., & Boudreau, J. W. (1997). Voluntary turnover and job performance: Curvilinearity and the moderating influences of salary growth and promotions. *Journal of Applied Psychology*, *82*, 44–61.

Tubre, T. C., & Collins, J. M. (2000). Jackson and Schuler (1985) revisited: A meta-analysis of the relationships between role ambiguity, role conflict, and job performance. *Journal of Management*, *26*, 155–169.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207–232.

*Tziner, A., Ronen, S., & Hacohen, D. (1993). A four-year validation study of an assessment center in a financial corporation. *Journal of Organizational Behavior*, *14*, 225–237.

U.S. Department of Labor. (1991). *Dictionary of occupational titles* (4th ed.). Washington, DC: Author.

Vance, R. J., MacCallum, R. C., Coovert, M. D., & Hedge, J. W. (1988). Construct validity of multiple job performance measures using confirmatory factor analysis. *Journal of Applied Psychology*, *73*, 74–80.

Van Scotter, J. R., Motowidlo, S. J., & Cross, T. C. (2000). Effects of task performance and contextual performance on systematic rewards. *Journal of Applied Psychology*, *85*, 526–535.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*, 557–574.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2002). The moderating influence of job performance dimensions on convergence of supervisory and peer ratings of job performance: Unconfounding construct-level convergence and rating difficulty. *Journal of Applied Psychology*, *87*, 345–354.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology*, *90*, 108–131.

*Warr, P., & Bunce, D. (1995). Trainee characteristics and the outcomes of open learning. *Personnel Psychology*, *48*, 347–375.

Williams, C. R., & Livingstone, L. P. (1994). Another look at the relationship between performance and voluntary turnover. *Academy of Management Journal*, *37*, 269–298.