Psychological Methods
1996, Vol. 1, No. 2, 199–223

1082-989X/96/$3.00

Measurement Error in Psychological Research: Lessons From 26 Research Scenarios

Frank L. Schmidt
University of Iowa

John E. Hunter
Michigan State University

As research in psychology becomes more sophisticated and more oriented toward the development and testing of theory, it becomes more important to eliminate biases in data caused by measurement error. Both failure to correct for biases induced by measurement error and improper corrections can lead to erroneous conclusions that retard progress toward cumulative knowledge. Corrections for attenuation due to measurement error are common in the literature today and are becoming more common, yet errors are frequently made in this process. Technical psychometric presentations of abstract measurement theory principles have proved inadequte in improving the practices of working researchers. As an alternative, this article uses realistic research scenarios (cases) to illustrate and explain appropriate and inappropriate instances of correction for measurement error in commonly occurring research situations.

There are two central questions as to error of measurement: (a) Should we correct empirical findings for the biases produced by imperfect measurement? (b) If we decide to correct, then how do we use different estimates of reliability to make corrections in various situations? This article addresses both questions.

Should empirical findings be corrected for the distortions produced by imperfect measurement? There are two very different answers to this question suggested by the current literature. On the one hand, the methodological literature since 1910 has been virtually universal in stating that correction is not only desirable but critical to both accurate estimation of scientific quantities and to the assessment of scientific theories. We summarize that informed opinion later in this article. On the other hand, examination of currently published studies shows that most studies—especially in laboratory research areas—still make no mention either of error of measurement or of reliability. Does this mean that there are many areas in which it is safe to assume perfect measurement in psychology? We note later that empirical findings have shown this to be false. Every psychological variable yet studied has been found to be imperfectly measured, as is true throughout all other areas of science. Furthermore, even in articles in which reliabilities are reported, the majority of studies do not use those reliabilities to correct findings for the distortion produced by error of measurement. Does this mean that if one reports the reliability, then the distortions produced by error of measurement will no longer occur? That is, is it true that a scientist need only mention error of measurement in order to make it go away? Alas, this too is known to be false. If one knows the reliability, then one can compute the extent of distortion and correct for that distortion. However, the distortion in the primary findings is still there even if one reports the reliabilities in the Method section.

There has never been either mathematical or substantive rebuttal of the main findings of psychometric theory. Random error of measurement distorts virtually every statistic computed in modern

Frank L. Schmidt, Department of Management and Organization, College of Business, University of Iowa; John E. Hunter, Department of Psychology, Michigan State University.

Correspondence concerning this article should be addressed to Frank L. Schmidt, Department of Management and Organization, College of Business, 108 Pappajohn Business Administration Building, University of Iowa, Iowa City, Iowa 52242-1000. Electronic mail may be sent to frank-schmidt@uiowa.edu.

studies. Accurate scientific estimation requires correction for these distortions in all cases.

However, while the basic correction formulas are very simple, application of those formulas is often not simple. Reliability can be estimated in different ways, and different methods capture different components. If a large error component is missed by the method used in a given study, then correction using the reliability estimate may be only partial correction. Furthermore, the distortion produced by error of measurement differs for different statistics and differs across situations depending on the specific error processes that enter a particular computation. Many researchers have found it very difficult to translate psychometric textbooks into useful procedures for their particular study. The bulk of this article is devoted to presenting these issues in what we believe to be a much more understandable manner than that used in current measurement texts.

Should We Correct Empirical Findings for Distortions Due to Imperfect Measurement?

In recent years, changes have taken place that make it more important for researchers to understand the effects of measurement error in their research and to correct properly for these effects. Research has increasingly moved away from the mechanically empirical approach of the past to an increasing emphasis on the development and testing of theoretical explanations of behavior (e.g., see Klimoski, 1993; Schmitt & Landy, 1993). This change has led to the realization that in theory-based research, the real interest is in the relationships that exist between actual traits or constructs rather than between specific measures of traits or constructs. That is, the interest is in correlations at the true-score (or construct) level rather than in the observed-score correlations that are different for different measures, depending on the amount of measurement error they contain (J. P. Campbell, 1990; Schmidt, 1993; Schmidt & Hunter, 1992). As a result, psychological researchers have increasingly come to realize the need in theory testing to correct observed relationships for the biasing effects of measurement error. J. P. Campbell (1990) has explored this change in some detail and has indicated that it will continue into the future. Because of the emphasis in structural equation modeling on correcting biases caused by measurement error, the increasing use of structural equation modeling has also contributed to this change in thinking. These developments indicate that it is becoming increasingly important to make the appropriate corrections for biases induced in research data by measurement error.

At the same time, there has developed a broader emphasis on precision in estimating correlations and effect sizes, both corrected and uncorrected and in both applied and theory-based research. A better understanding of the magnitude of the impact of sampling error has led to the use of larger samples in individual studies and has thereby led to increased precision in estimates of correlations. In addition, the use of meta-analysis has provided more precise and generalizable estimates of relationships among different constructs and measures (Schmidt & Hunter, 1992). These advances have led to the demise of the older belief that population relations among variables are unique to settings, organizations, or subgroups and have therefore encouraged attempts to develop general theories and explanations (Schmidt, 1992). This increasing emphasis on precision has made it more important that corrections for biases due to measurement error be as precise and accurate as possible and not merely approximately correct.

The amount of measurement error variance in some measures used in psychological research is large, often in the neighborhood of 50% of the total variance of the measure. This is especially likely to be true of the short scales often used in field research to save time and thereby allow a larger number of measures to be included. Yet many researchers today still do not correct their data at all for biases induced by measurement error. In fact, in many research areas, failure to even acknowledge biases created by measurement error is still a more important problem today than is the use of inappropriate corrections.

Problems in Applying Reliability Theory and Methods

Corrections for biases due to measurement error are used increasingly frequently in the research literature. We know from our experience as reviewers and from reading the published literature that inappropriate attenuation corrections continue to be made by researchers. Some of these errors lead to large downward or upward biases. Researchers need to know how to make these corrections correctly and how to evaluate them properly when they are made and presented by others in the research literature. How can this best be accomplished? Many extended technical treaties on the abstract principles of reliability theory have appeared over the years, especially in the educational domain (e.g., Cronbach, 1947; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Feldt & Brennan, 1989; Stanley, 1971; Thorndike, 1951; Traub, 1994). Virtually all of these have discussed the biasing effects of measurement error on correlations, and most have discussed corrections for these biases. Yet these abstract psychometric dissertations appear to have had little impact on working researchers. These discussions rarely illustrate the abstract technical principles with concrete research applications. Without such examples, it is difficult for researchers to apply the abstract technical principles to concrete research problems that appear in a wide variety of forms and with a myriad of obscuring features, details, and idiosyncracies. This article is based on a different approach: examination of a series of concrete research scenarios that we have encountered in our work as researchers, advisors to researchers, or reviewers. This article examines 26 real-world "case studies" and explicates them on the basis of the principles of reliability theory found in Cronbach (1947, 1951) and Cronbach et al. (1972). Although we cite and rely on these sources, other sources (e.g., Thorndike, 1951) would yield identical resolutions; only the terminology would be sometimes (slightly) different.

In classical measurement theory, the fundamental general formula for the observed correlation between any two measures, x and y, is

$$r_{xy} = r_{x_t y_t}(r_{xx} r_{yy})^{1/2} \qquad (1)$$

where r_{xy} is the observed correlation, $r_{x_t y_t}$ is the correlation between the true scores of the measures x and y, and r_{xx} and r_{yy} are the reliabilities of x and y, respectively. This is called the *attenuation formula*, because it shows how measurement error in the x and y measures reduces the observed (computed) correlation (r_{xy}) below the true score correlation ($r_{x_t y_t}$). Solving this equation for $r_{x_t y_t}$ yields the dissattenuation formula:

$$r_{x_t y_t} = r_{xy}/(r_{xx} r_{yy})^{1/2}. \qquad (2)$$

If the sample size is infinite, both these formulas are perfectly accurate. That is, they are perfectly accurate in the population. In the smaller samples used in real research, there are sampling errors in the estimates of r_{xy}, r_{xx}, and r_{yy}, and therefore there is also sampling error in the estimate of $r_{x_t y_t}$. Because of this, a circumflex is sometimes used to indicate that all values are estimates:

$$\hat{r}_{x_t y_t} = \hat{r}_{xy}/(\hat{r}_{xx} \hat{r}_{yy})^{1/2}. \qquad (3)$$

The $\hat{r}_{x_t y_t}$ is the estimated correlation between the construct underlying the measure x and the construct underlying the measure y.[1] Alternatively, it is an estimate of the (uncorrected) correlation that we would observe between the measures x and y if both measures could be made free of measurement errors. This much is familiar to some researchers. What confuses many, however, is the question of what type of reliability estimate should be used under different circumstances to provide the appropriate estimate of the true-score correlation. Indeed, because there are several types of reliability coefficients, each with different properties, this question can be confusing and can lead to erroneous estimates, as is illustrated in the research scenarios that follow.

Corrections for measurement error are also included in meta-analysis methods as described by Hunter and Schmidt (1990) and Hunter, Schmidt, and Jackson (1982). In our advising of researchers conducting meta-analyses, we have repeatedly observed uncertainty and confusion as to the kinds of reliability estimates that are appropriate for inclusion in meta-analyses. The principles that apply to the selection of the appropriate reliability

[1] Item response theory can be used to demonstrate that the true scores for many scales based on classical measurement theory are monotonically but not perfectly linearly related to the underlying trait (construct) in question (Lord & Novick, 1968). However, the relationship is usually close enough to linear that the effect on the estimated construct-level correlations of taking true scores as colinear with the construct is negligible. However, the trend toward increased precision in psychological research means that someday even this factor may have to be routinely taken into account.

coefficient for use in single studies also apply to meta-analyses. Hence the guidance provided in the single-study scenarios in this article can be used to determine the appropriate type of reliability coefficients to use in meta-analyses.

There are two approaches in meta-analysis to correcting biases due to measurement error. In the first, individual correlations or *d* values from each primary study are first corrected, and then the meta-analysis is performed on these corrected correlations or *d* values (Hunter & Schmidt, 1990, chap. 3 & 7). In this approach to meta-analysis, the entire process of making corrections for measurement error is identical to that described in this article.

Often the reliability information necessary to correct each correlation or *d* value individually is not presented in most of the primary studies. In those cases, the "artifact distribution" meta-analysis method is used (Hunter & Schmidt, 1990, chaps. 4 & 7). In this approach, reliability values are taken from individual studies, test manuals, and other sources that do report the appropriate reliability estimates. These values are then compiled into reliability distributions characteristic of that research literature, and these distributions are then used with appropriate calculational procedures to correct for the biases induced by measurement error. In this case it is again important to know what types of reliability estimates are appropriate for use in these artifact distributions. This article provides that information for the most commonly occurring research situations.

Failure to Correct and Resulting Problems

Scenario 1

Situation. A researcher is interested in theories relating job satisfaction to perceptions of fairness in the reward for performance. She decides to test one such theory by comparing the level of satisfaction following performance under two different incentive schemes. In a laboratory experiment, she randomly assigns 20 students to one of two conditions; 10 to each cell of the design. The 10 subjects assigned to Condition A perform with reward determined by Incentive Scheme A. The 10 subjects assigned to Condition B perform with reward determined by Incentive Scheme B. The dependent variable is the subject's report of satisfaction with the reward. Subjects were asked to record their feeling about the experiment as "satisfied" or "not satisfied."

The researcher was quite troubled when she found no significant difference between the means for the two groups. When asked whether she assessed the reliability of the dependent variable, she claimed that there is no error of measurement in her study. She stated she was careful to use perfectly random assignment to conditions and was equally careful in correctly recording exactly which response option was chosen by the subject.

Problem. There are two errors in this researcher's beliefs. First, the random assignment of subjects to conditions is irrelevant to this issue. If subjects were nonrandomly assigned to conditions, there might be a potential problem with the construct validity of the treatment variable but there is little likelihood that the nonrandom assignment would have any effect on the extent of random measurement error in the measure of the dependent variable. Thus random versus nonrandom assignment is irrelevant to the issue of imperfect reliability in the measurement of satisfaction.

The deeper and more pervasive error in the researcher's thinking is the identification of the concept "error of measurement" with the process of correctly recording the data. Random error in the measure of satisfaction is produced if the response recorded is subject to any random influence. It is true that error in recording the responses would be one such source of error, a source that is carefully controlled in most psychological studies. But many studies have shown that error in recording responses is usually trivial in comparison to other sources of random error (see Hunter & Schmidt, 1990, pp. 117–125, for a full discussion of such sources).

The largest source of random error in the recorded response is randomness in the response process itself. Many studies have shown that most human responses have a large random component. In the case of very well rehearsed responses such as answering "Are you male or female?," the response is often highly predictable. For example, in the famous "American Voter" study (A. Campbell, Converse, Miller, & Stokes, 1960), the test–retest reliability of the variable sex was .95. In these extreme cases, random response error is small in magnitude, and ignoring it causes little error in causal inference.

But the essence of most field research and virtually all laboratory research is to study responses that are *not* well rehearsed. Studies are conducted on the basis of subject responses that should vary in order to reflect the causal processes to be studied. Thus in most research contexts—and in laboratory studies in particular—any observed response will have a very large random component. If a subject is observed in an exactly repeated situation, the correlation between "replicated" responses is rarely any higher than .25. That is, for unrehearsed single responses, it is unwise to assume a test–retest reliability higher than .25.

In field research, researchers are taught to be concerned about measurement. They are taught that randomness can be reduced if multiple responses are observed and averaged to generate the final measurement. This is the principle of the "test" or "scale" of measurement.

Consider the researcher in this scenario. In using a single item to measure satisfaction, it is likely that the reliability of that measure is no higher than .25. For purposes of this scenario, let us assume that value. Had the researcher used a seven-item scale to measure satisfaction, then the Spearman-Brown formula indicates that reliability would climb to .70, that is, nearly three times more reliable measurement. With a 15-item scale, the reliability would have been .84.

Consider now the effect size for the study and the implications of differences in quality of measurement for statistical power. Lipsey and Wilson (1993) recently reviewed over 300 meta-analyses on a very wide variety of psychological treatments. The average effect size was $d = .46$ (corresponding to a point-biserial r of .22, if sample sizes are equal in experimental and control groups). Suppose we take this to be the effect size for perfect measurement in the present researcher's study.

With a reliability of .25, the effect sizes in her study would be reduced from a potential $d = .46$ to an actual $d = .23$ (i.e., $\sqrt{.25}(.46) = .23$). The probability of finding a significant result would then be only 12% (using a one-tailed test). That is, the probability of an error for the significance test in this study is 88%. So it is no great surprise that she failed to find a significant difference.

Had she used a seven-item scale to measure satisfaction, the effect size would have been $d = .31$. The probability of finding a statistically significant difference in this study would be 17% for a one-tailed test. For a sample size using 10 subjects per cell, this means that the significance test still has an error rate of 83%. But the odds using a seven-item scale improve from 12% to 17%, a 42% increase in the odds of drawing the correct conclusion from the study.

For a 15-item scale, the effect size rises to .40—almost the maximum value achieved using perfect measurement—and the probability of observing a significant difference rises to 23% using a one-tailed test. That is, with a 15-item scale the odds that the significance test will produce the wrong result drop to 77% using a one-tailed test. By comparison, the odds of drawing a correct conclusion rise from 12% for a 1-item scale, to 17% for a 7-item scale, to 23% for a 15-item scale.

For perfect measurement, the effect size is $d = .46$ and the probability that the significance test is right is 26% using a one-tailed test. This is not a lot higher than the power for the 15-item scale. Thus there is only limited improvement in going from a reliability of .84 to a reliability of 1.00.

This researcher would have been well rewarded to have correctly learned the nature of random error in measurement of human responses (animals are even more random). Even in laboratory studies it is critical to measure the reliability of the dependent variable. It is also important to correct the study value for attenuation using the measured reliability, since this will make it possible to compare the results across studies using measurement scales with differing reliabilities.

Scenario 2

Situation. A researcher is testing several hypotheses from a theory of work adjustment. These hypotheses predict certain relationships among the constructs of role ambiguity, role stress, and overall work adjustment. The researcher is aware that measurement error can distort research findings, so she decides to use scales that are somewhat longer than average to attain adequate levels of reliability for her measures. After her data are gathered, she finds that coefficient alpha is .76 for the measure of role ambiguity, .85 for the measure of role stress, and .81 for the measure of overall work adjustment. She has read in a text on research methods that a reliability of .70 or above is adequate for research purposes. She concludes that there is no need to correct for measurement error, because the fact that she has adequate levels of reliability means that measurement error is not a problem in her research data.

Problem. This case is an example of the magic number belief. This researcher falsely believes that if reliability is as high as the magic number .70, then there will be no downward bias in observed correlations due to attenuation resulting from measurement error. But the bias introduced into estimates of correlations between constructs does not magically cease to exist when reliability hits the .70 level. In fact, when reliability is .70, the bias factor is $\sqrt{.70} = .84$; that is, the observed correlation will be on average 16% below its correct value. But that is the effect of measurement error in only *one* of the two variables. If both variables have reliability of .70, the bias factor is $\sqrt{.70(.70)} = .70$. This means that the observed correlation would on average be 30% below its correct value, a very large bias. In this researcher's data, the bias factor for the observed correlation between the measures of role ambiguity and overall work adjustment would be $\sqrt{.76(.81)} = .78$, a downward bias of 22%.

It is true that larger reliabilities produce smaller downward biases than do smaller reliabilities. But some bias continues to exist unless and until reliability reaches 1.00, which never happens. Thus there is never a magic value for reliability beyond which estimates of correlations from research data are unbiased and can be taken at face value as estimates of correlations among the constructs of interest.

Scenario 3

Situation. A researcher is studying a job in which mental arithmetic must be used to make on-the-spot decisions about warehouse storage. His task is to determine the best way to assure adequate mental arithmetic skills on the job to prevent costly decision errors by employees. He first evaluates a training program designed to improve this skill; comparing the experimental and control groups, he obtains a *d* value of .41. That is, he finds that the training program increases skill in mental arithmetic by 41% of a standard deviation. He next conducts a predictive criterion-related validity study (on a different group) of a general mental ability test, using as the criterion the same measure of mental arithmetic given at the end of the training program. He corrects the observed criterion-related validity coefficient of .32 for measurement error in this criterion. However, he does not correct the *d* value from the training evaluation study, stating that measurement error corrections are applicable only to correlations.

Problem. Measurement error corrections are just as applicable to *d* values as to correlations. The *d* value is the standardized difference between the control group and experimental group means; it is biased downward by measurement error in the dependent variable (here the measure of mental arithmetic skills) in exactly the same way as the correlation, and the correction should be made in exactly the same manner (Hunter & Schmidt, 1990, pp. 241–247).

Scenario 4

Situation. An industrial–organizational psychologist conducts a large-scale validity study of an assessment center used to select first-line supervisors in a large manufacturing firm. Six different criterion measures of job behavior are used; two of these measures are promotion rate and job tenure. The researcher makes corrections for criterion unreliability for the other measures, but she argues that there is no unreliability in promotion rate and tenure. People are either promoted or not, and there is no error in the recording of the promotions. The records also clearly show whether each participant left the company and, if so, how long tenure was before leaving. She is confident there are no recording errors in tenure, either. In her research report, she therefore states that the uncorrected validity coefficients for the criteria of promotion rate and tenure are unbiased estimates of the true validities.

Problem. For purposes of this scenario, we accept the researcher's claim that there are no recording errors in these data, although in reality this is unlikely to be the case. But even if there were no recording errors, the promotion and tenure measures cannot be assumed to be perfectly reliable. As a general matter, there are no measures that are completely free of measurement error. Even in the hard sciences, whenever measures have been examined, measurement error has been found.

Although it may be difficult to assess the reliability of promotion rate, it is not impossible. Promotions occur over considerable time periods and are based on assessments of employee job performance and capabilities. For most employees, there are multiple opportunities for promotion. For employees who have stayed with the organization for a reasonable period of time, personnel records can be used to obtain their promotion rates during two periods of time. The correlation between these two promotion rate measures can be corrected using the Spearman-Brown reliability to be appropriate for the time period used in the validity study.

It is also possible to determine whether the relative standing of employees on promotion rate changes over time. Doing that requires obtaining measures of promotion rate for three time periods during careers. The three correlations among these three periods are then computed. If there has been no change, then these three correlations will be equal (within the limits of sampling error). If systematic change is found to be occurring, then the reliability of the promotion rate measure (r_{yy}) is computed as

$$r_{yy} = r_{12}r_{23}/r_{13}.$$

A similar procedure can be used to estimate the reliability of the measure of tenure. However, that determination would be more difficult, because it would require tenure records on some individuals from several organizations, and organizational differences might bias the reliability estimates. Determining the reliability of promotion rate and tenure might be difficult or even infeasible in any one study. However, it should be possible to obtain estimates of this general sort from the research literature. It should be, but it currently is not, because to our knowledge the needed studies have not been conducted, or if they have, they have not been published. Thus this is an information gap in the research literature that needs to be filled. Until this gap is filled, the best option when such reliability estimates cannot be made is to state that the observed validity coefficients for promotion rate and tenure are downwardly biased to some unknown extent. It is erroneous to pretend that these measures are perfectly reliable.

Scenario 5

Situation. On the basis of his own research and findings by others, a researcher hypothesizes that a major cause of job dissatisfaction is negative affectivity, the general tendency to experience negative emotions. He further hypothesizes that the trait of negative affectivity is highly stable over time. To calibrate the degree of stability, he administers a measure of negative affectivity to a large group and 1 year later administers a parallel form to the same group. Both forms have been carefully developed and are reliable; one has an estimated alpha coefficient of .86 and the other an alpha of .84. The correlation over the 1-year period is .80. On the basis of this 1-year correlation of .80, the researcher concludes that the trait (construct) of negative affectivity is fairly but not completely stable. He concludes that 20% of the variance of the trait is unstable and 80% is stable.

Problem. This estimate of the stability of the trait of negativity is downward biased. The researcher has confused the stability of the measures with the stability of the trait. The 1-year stability of the measures is .80; the 1-year stability of the trait is higher. Random errors of measurement at each time period bias the correlation between the two time periods downward. These errors of measurement are assessed by the alpha coefficients. Correcting for measurement errors at Times 1 and 2 provides an estimate of the stability of the trait:

$$\text{Estimated stability} = .80/(.86 \times .84)^{1/2} = .94.$$

Thus the trait of negativity is actually considerably more stable than this researcher concluded.

Actually, this trait may be nearly perfectly stable. The .94 value above is at least to some small degree an underestimate. This conclusion stems from the fact that coefficient alpha consigns transient error (Hunter & Schmidt, 1990, pp. 123–125) to true variance and, as a result, somewhat underestimates measurement error (and overestimates reliability) at Time 1 and at Time 2. In this study, subjects respond to the measure of negative affectivity at Time 1 and Time 2. At each time, that particular occasion is characterized for each subject by a certain mood, level of emotion, type of feeling, and so on. These factors are transient: A day later or a week later, they may be different, and hence the individual's responses may be somewhat different. These transient influences on the scores are not part of the construct (negative affectivity) that the researcher is trying to measure. They are a part of measurement error variance, not true variance. The ideal research design would be as follows. At Time 1 administer parallel form measures a week or so apart and estimate reliability at Time 1 as the correlations between these two measures taken at these two times. This estimate of reliability controls for transient error: Transient moods and other factors will not reproduce themselves on the second occasion. Do the same thing at Time 2. These reliability estimates should be slightly smaller than the alpha coefficients. Then correct the .80 correlation across the 1-year period using these reliability estimates. The resulting stability figure will probably be somewhat larger than the .94 above; it might be 1.00, indicating perfect stability of the trait.

Scenario 6

Situation. A researcher is testing specific aptitude theory against the *g* factor theory of mental ability. He uses multiple regression to test his hypothesis that the specific aptitudes of quantitative, verbal, and spatial ability contribute to the prediction of job performance over and above the prediction from general mental ability (*g*). All three of his ability measures are reliable; each has reliability of at least .80. With a large N ($N = 3{,}312$) to ensure stable results, his regression results appear to confirm his hypothesis: Although the standardized regression weights for the specific aptitudes are much smaller than the weight for *g*, some are nevertheless statistically significant and seemingly large enough to be of practical value in some situations. In addition, the multiple R is .02 larger when the specific aptitudes are added.

Problem. Unless the measure of *g* has perfect reliability (never the case and not the case here), the addition of the specific aptitude measures to the measure of general ability serves to raise the reliability (and hence the validity) of the predictor set as a whole (the weighted sum of the independent variable scores) as an overall measure of *g*. The effect of this can be to create the false appearance of nonzero beta weights for the specific aptitudes and the false appearance of an increment to the multiple correlation by the specific aptitudes. Schmidt, Hunter, and Caplan (1981) discussed this phenomenon in some detail. After being informed of this possibility, the researcher tested this hypothesis by correcting all (the criterion related) validities and intercorrelations for measurement error and then recomputing the standardized regression weights. This time he found that all the weights for the specific aptitudes were zero and that the increment to the multiple correlation from adding the two specific aptitudes as predictors was zero. These findings confirm the alternative hypothesis that the trait of general intelligence (*g*) alone was responsible for the prediction.

Scenario 7

Situation. A researcher has developed a 30-item self-report measure of organizational citizenship behavior that she believes has more construct validity than the self-report measure that is currently used in most studies in the literature. She hypothesizes that her measure correlates higher with supervisory evaluations of citizenship behavior than does the older measure. She tests this hypothesis on a group of 48 laboratory technicians who complete both measures and are rated on citizenship behaviors by their supervisors. She reports that her measure correlates significantly with the supervisory ratings ($r = .27$, $p > .05$), while the other measure does not ($r = .22$, *ns*). She uses coefficient alpha to estimate the reliability of the organizational citizenship measures and uses the appropriate measures of interrater reliability to estimate the reliability of the ratings. Corrected for unreliability in both the measure and the ratings, the correlation for her measure increases from .27 to .46. She does not correct the correlation for the other measure because it is not significant, stating that a correlation that is not significantly different from zero should not be corrected. She concludes that the best estimate of the true-score correlations with supervisory judgments of organizational citizenship behaviors is .46 for her measure and zero for the other measure. Therefore she concludes that her measure has construct validity and that the other measure does not.

Problem. This comparison is quite biased. Consider the observed correlations of .22 and .27 for the two measures. The 95% confidence intervals for these two correlations show an overlap of .51 correlation points: 95% confidence interval for the old measure, $-.06 < .22 < .50$; 95% confidence interval for the new measure, $-.01 < .27 < .55$. Thus the confidence intervals provide no foundation for concluding that these correlations are different. (A significance test of the difference between these two correlations would also indicate that they are not significantly different; but the confidence interval provides much more direct information [Berenstein, 1994; Schmidt, 1994].) Thus there is no basis for this researcher's conclusion that the correlation of .27 should be corrected for measurement error while the correlation of .22 should not.

This researcher has used a false decision rule that is often used in data analysis (Schmidt, 1992). This decision rule states that if a difference or correlation is not statistically significant, then the best point estimate of its value is zero. But the best point estimate of the population value of an observed correlation or difference is the observed value of the correlation or difference, regardless of whether it is statistically significant or not. Hence in this case the best estimates of the measurement error–attenuated population correlations are .22 for the old measure and .27 for the new measure. The best estimates of the true-score correlations for these measures are these values corrected for measurement error. For the new measure, correction for the bias introduced by measurement error increases the estimate from .27 to .46. For the old measure, this correction increases the estimate from .22 to .37. Comparison of these two corrected values *suggests* that the older measure may have less construct validity than does the new measure. However, this is only a weak suggestion, as can be seen by comparison of the confidence intervals for the corrected correlations.

The confidence intervals for the corrected correlations can be computed by correcting the endpoints of the confidence intervals for the observed correlations for the biasing effects of measurement error (Hunter & Schmidt, 1990, pp. 121–122). When we do this we get the following true-score correlation confidence intervals: 95% confidence interval for old measure, $-.10 < .37 < .85$; 95% confidence interval for new measure, $-.02 < .46 < .94$. Comparison of these confidence intervals makes it clear that each correlation lies well within the confidence interval for the other. Thus the difference between the .46 and the .37 is well within the limits of differences expected from sampling error. Thus even if the data in this example are analyzed correctly, the original question cannot be answered here with any certainty. A sample size of 48 does not contain enough information to provide an adequate test of the researcher's hypothesis. This is true whether or not statistical significance tests are used. The key point illustrated in this scenario, however, is that the practice of correcting only statistically significant correlations for measurement error leads to serious biases. It is akin to the practice of reporting only statistically significant correlations, leading to an upward bias in reported correlations.

Scenario 8

Situation. A researcher is interested in the heritability of the personality traits of extraversion and emotional stability (neuroticism, when scored negatively). Using a well-known personality inventory, he measures these traits in samples of identical and fraternal twins. On the basis of the scores obtained, his observed heritabilities are .43 for extraversion and .45 for emotional stability. He concludes that for the trait of extraversion, 43% of the variance is due to genetic effects and 57% to environmental effects. For the trait of emotional stability (neuroticism), he concludes that 45% of the variance is explained by genetic differences between individuals, while 55% is due to differing environmental influences.

Problem. These conclusions are incorrect. The heritabilities that are the basis of these conclusions are observed heritabilities; they have not been corrected for the biases caused by measurement error in the personality scales. Thus, for example, .43 is a downwardly biased estimate of the heritability of the trait of extraversion. The .43 figure estimates the heritability not of the trait but of the scores on the measure of extraversion. Furthermore, some of the 57% of variation attributed to environmental effects is actually the variance of measurement errors; thus this researcher has overestimated the size of environmental effects in addition to underestimating heritability of the traits.

What type of reliability should be used to correct these observed heritability estimates? We need a measure of reliability that assigns a measurement error not only random errors of measurement but also transient errors (see Scenario 5) and specific factor variance. Thus what is needed is the correlation between two parallel forms of these measures separated by, say, 2 weeks in time. In the personality domain, such reliabilities average about .65. Suppose the values in this case are .67 for extraversion and .64 for emotional stability. Then the unbiased estimate of heritability for the trait of extraversion would be .43/.67 = .69. And the unbiased estimate of the heritability for the trait of emotional stability is .45/.64 = .70. Thus the heritabilities of the traits are considerably higher than are the heritabilities of the observed scores on the measures. The amount of variation that can be attributed to environmental effects is considerably less, about 30%.

The reader probably noticed that in making these corrections for measurement error, we divided by the reliabilities themselves instead of by the square roots of the reliabilities. This is because the heritability is an estimate of a squared correlation: the squared correlation between genetic differences and the scores, in the case of observed heritabilities, and the squared correlation between genetic differences and the actual trait, in the case of corrected heritabilities. (This is the reason that heritabilities are represented by the symbol h^2 and are given percent-variance accounted-for interpretations.) Therefore the square root of the observed heritabilities is the estimated correlation between the genetic differences (genotypes) and the observed scores; for example, for extraversion this correlation is $\sqrt{.43}$ = .66. Likewise, the square root of the corrected heritability is the estimated correlation between genetic differences (genotypes) and the trait itself; for example, for extraversion this correlation is $\sqrt{.69}$ = .83. For the trait of emotional stability, this estimated correlation is $\sqrt{.70}$ = .84. Thus the correlations underlying heritabilities are considerably larger than are the heritabilities themselves. A trait with a heritability of .70 correlates .84 with genetic differences between individuals.

Scenario 9

Situation. A researcher hypothesizes that behaviorally anchored rating scales (BARS; Smith & Kendall, 1963) will lead to reduced halo error in comparison with Likert scales. Halo error in job performance ratings refers to unrealistically high correlations between differently rated dimensions of job performance. For example, if the population correlation between personal appearance and computer programming skills is .10, a correlation of .85 between personal appearance and computer programming skills would be evidence of halo error. To test his hypothesis, the researcher develops BARS and Likert scales to measure the same 10 dimensions of job performance (e.g., skills in dealing with customer complaints). The BARS scales are of the usual type, and each Likert subscale has seven items. He has a group of supervisors do performance appraisals with both sets of scales and computes scale intercorrelations. The average correlation between performance dimensions is .72 for the Likert scales but only .53 for the BARS scales. The researcher concludes that BARS scales do indeed result in reduced halo.

Problem. The researcher failed to control for the reliability of the ratings. Each BARS rating is based on only one item; that is, for each performance dimension, there is only one judgment and only one response. Each Likert rating, on the other hand, is based on seven items; the rater makes seven judgments for each performance dimension. Just as in the case of tests, longer measures are usually more reliable than are shorter measures. Higher reliability means higher correlations among dimensions, because there is less downward biasing of these correlations due to measurement error. Thus the lower intercorrelations between performance dimensions for the BARS scale may be due merely to lower reliability.

The following procedure can be used to test this hypothesis. First, develop a BARS scale that has two items per dimension. Within each dimension, the correlation between these two items, after correction using the Spearman-Brown formula, is the reliability of the sum of the two items on that dimension. These summed two-item scores should then be correlated across the 10 dimensions, and these 45 correlations should be corrected using these reliability estimates. Then the average of these corrected correlations is computed for the BARS scale.

The same procedure should then be followed for the Likert scales. Coefficient alpha should be computed for each of the 10 seven-item Likert scales; this provides an estimate of reliability comparable with that computed for the BARS scales. Next, the correlations should be computed between total scores on the 10 Likert scales. Then these 45 correlations should be corrected for unreliability using the alpha coefficients for the Likert scales.

The average of the corrected correlations for the Likert scales should then be compared with the corresponding average for the BARS scales to determine which has more halo. Correcting for bias due to unreliability eliminates the possibility that differences between the two scale types in average dimension intercorrelation are due to differences in the reliability of ratings. Only if the BARS average is smaller than that for the Likert scales could one conclude that BARS scales are less affected by halo error.

Despite almost 30 years of discussion and debate about whether the BARS rating format reduces halo error, the analysis described here has never, to our knowledge, been conducted. Yet it is the only approach that can provide an unambiguous answer to the question.

Correcting Using the Wrong Reliability Coefficient

Scenario 10

Situation. A researcher conducts a criterion-related validity study of an assessment center using supervisory ratings as her criterion measure of job performance. Job performance is rated by the first-line supervisor (manager) on 10 dimensions of job performance. The sum of standardized rating across these 10 dimensions is used as the final index of overall job performance. On the basis of the correlations among these rated dimensions, she computes coefficient alpha ($\alpha = .84$), which she uses to correct the observed validity coefficient for attenuation due to unreliability in the criterion measure.

Problem. Coefficient alpha as used here is a measure of intrarater reliability. It is an estimate of what the correlation would be if the same rater rerated the same employees (Cronbach, 1951). What is needed is a measure of interrater reliability. The problem with intrarater reliability is that it assigns specific error (unique to the individual rater) to true (construct) variance. Each rater is analogous to a different form of the rating instrument; the specific error for each rater is that rater's idiosyncratic perceptions of employee job performance. This specific error is known to be very large in ratings of job performance (King, Hunter, & Schmidt, 1980; Rothstein, 1990). Average intrarater reliability can be measured using coefficient alpha, as is done here, or by having the same rater rate the same group of employees at two different times and correlating the ratings. With either method, intrarater reliability for the total rating score from multiple subscale rating instruments is usually found to be in the .80 to .90 range. Interrater reliability, however, is much smaller, averaging about .50 for the same type of rating scales (Rothstein, 1990). That is, the average correlation of total ratings between two different knowledgeable raters who rate the same employees is only about .50. The difference between these two figures is quite large, indicating that somewhere between 30% and 40% of the variance of the ratings of the average rater is specific factor error variance.

Use of intrarater reliabilities to correct criterion-related validity coefficients for criterion unreliability produces substantial downward biases in estimates of actual validity. This bias results from the fact that intrarater reliability assigns specific factor error variance to true job performance variance. In Scenario 12, we see that random response error variance is very large in employment interviews. Random response error is somewhat smaller in total job performance ratings that are the sum of rating across, say 10 or more, rating judgments (10 or more rating subscales). Such job performance ratings vary less from day to day, as is shown by intrarater reliabilities in the .80 to .90 range. But specific factor error is a very large component of ratings, perhaps an even larger component than in the case of personality tests (see Scenario 13). With both ratings and personality inventories, failure to control for specific factor error leads to especially serious errors in research conclusions.

Scenario 11

Situation. Many researchers have hoped that they could measure different dimensions of job performance by asking supervisors to rate employees on those dimensions. However, the ratings on different dimensions are typically so highly correlated that most researchers have adopted a halo hypothesis about ratings. The extreme form of this hypothesis holds that in rating different dimensions of job performance, raters are essentially rating the dimension of overall performance over and over.

A researcher is studying halo in ratings of employee job performance. He defines 12 dimensions of job performance, and each employee is rated on each of those 12 dimensions. Each of 12 dimensions is measured by a separate nine-item Likert scale.

For each rater, a 12 × 12 correlation matrix between dimensions is formed by correlating the ratings made by that rater judging the subordinates who work under that rater. The researcher found these matrices to be essentially identical in form for different raters, and so he averaged across raters to reduce the impact of the sampling error.

For every rater, the correlations between the rated work dimensions are quite large, with all correlations falling in the .70 to .80 range. When averaged across raters, the correlation matrix appeared to be flat with an average correlation of .75. The researcher then considered the hypothesis that halo is so strong that judges make no distinction between dimensions at all. All nominal dimensions are really the same as the rater's overall evaluation.

If this extreme halo hypothesis were true, then the correlations between dimensions would differ from 1.00 only because of error of measurement. That is, if the extreme hypothesis were true, then once the correlations are corrected for the downward bias due to measurement error, the correlations would differ from 1.00 only by sampling error.

The researcher knows that in the personnel selection literature, the proper reliability for ratings by a single supervisor is the interrater reliability, that is, the correlation between ratings of workers made by different supervisors. For well-constructed multiple-item rating scales scored by adding across items, the average correlation between two raters is .47 (King et al., 1980; see also the aggregate study by Rothstein, 1990, which found an average correlation of .50). The researcher used this value as the interrater reliability to correct for attentuation.

When corrected for attentuation using the average interrater reliability of .47, the average correlation between dimensions of job performance rose from .75 to 1.60. This was quite shocking to the researcher.

Problem. The error in this analysis is a failure to define the concept error of measurement in substantive terms. The definition of error of measurement used in the dimensionality research on ratings made by a single rater is different from the definition used in personnel selection. The corresponding definition of "reliability" will also differ.

In personnel selection research, the goal is to measure actual job performance. Research shows that supervisors are quite idiosyncratic in their evaluation of workers. This idiosyncracy is the largest known source of error of measurement in using ratings to evaluate performance. Thus as the first step in defining error of measurement, we define the "true score" for performance ratings to be the worker's average evaluation across a population of raters. One kind of error of measurement is the difference between the rating made by one supervisor and the consensus rating that would be obtained by averaging ratings made by the population of supervisors. This is the kind of error of measurement assessed by the interrater reliability.

In the dimensional research, the focus is on ratings made by a single supervisor. Because the researcher is studying the perceptual process in single raters, the idiosyncracy of perception is part of the construct being studied. Because we are not trying to estimate true performance, the idiosyncracy is not an error process in this study. Thus the interrater reliability is irrelevant to the issue of error of measurement in this research.

What is the relevant definition of error of measurement for ratings made by a single rater on a single dimension? Consider the problems that the rater has in responding to the items on the rating instrument. The rater must translate their memories, thoughts, and perceptions into a specific response to a specific question. Research in all areas has shown a large element of randomness in this translation process. This is random response error in the rating made on a specific item.

The researcher implicitly took this aspect of random response error into account by devising nine items for each dimension. Adding across the nine item responses is equivalent to averaging the individual-item random response errors. However, empirical research has shown that the size of the random response error to single items is so large that even averaging across nine responses will not eliminate the random error in the final scale score. The averaged error in the final scale score will still be large enough to require correction.

What is the relevant statistical definition of "random error" for this research, and how do we measure the relevant scale reliability? For each dimension, the rater is asked to consider nine items that are all considered to measure the same dimension. If this hypothesis is correct, then the nine responses to those nine items differ from the dimension score only by random response error. If this is true, then the relevant reliability for the dimension score is the usual Spearman-Brown reliability (or coefficient alpha or Kuder-Richardson 20 [KR-20]) for the dimension computed across the nine items. This scale reliability would be computed for each of the raters separately. The scale reliability for each rater could then be averaged across raters to reduce the impact of sampling error.

Consider the nine items for a given dimension and a given rater. Suppose the average correlation between the items for that rater is .39. The "internal consistency" reliability for the scale is computed using the Spearman-Brown formula with n equal to nine. So computed, the scale reliability for that scale for that rater is .85.

In principle, the average scale reliability across raters could be different for each dimension. This is usually not true, and for simplicity we assume that it is not true here. Assume that all dimensions have the same scale reliability of .85. We can then use that scale reliability to correct the correlation between dimensions. The average correlation between dimensions for a single rater in the research project was found to be .75. If we correct this using a reliability of .85 for each dimension, the corrected correlation is .75/.85 = .88. This correlation is less than the 1.00 predicted by the extreme halo hypothesis, though still far larger than the correlation between actual work dimensions (at least for some dimensions such as physical appearance vs. report writing, or either of these vs. peer rapport). This finding of a true-score correlation of .88 between ratings of different dimensions is similar to the average estimated true-score correlation in the 13 studies reviewed by King et al. (1980).

Scenario 12

Situation. A researcher is testing a theory that postulates that managerial motivation increases with age. The test of this hypothesis depends on the size of the true-score (construct level) correlation between managerial motivation and age. She uses the Incomplete Sentences Test (IST), a projective instrument, to measure managerial motivation. The responses of the managers to the IST are scored independently by two trained response coders; intercoder agreement (correlation) for total score is found to be .85. The researcher uses this reliability figure to correct the observed correlation (r = .20) between managerial motivation and age for attenuation due to measurement error. She finds that the corrected correlation (r = .22) is much smaller than predicted by the theory and concludes that this prediction from the theory is disconfirmed.

Problem. The procedure used here to estimate reliability is a commonly used one, but the resulting reliability estimate is not the appropriate one for testing the hypothesis under consideration. There are three sources of measurement error that bias the observed correlations downward from its true-score value in this study and that are not taken into account by this researcher.

First, there is purely random response error. This is similar to transient error (see Scenario 5) except that it varies within occasions as well as across occasions. The factors that produce random response error vary across moments within occasions. Administering the test twice on a single occasion would control for purely random response error but not for transient error. In this study, the two coders score exactly the same responses for each subject; hence the protocols that they evaluate contain exactly the same random response errors. Hence random response errors act to inflate the correlation between the two raters.

Second, another source of measurement error that inflates the .85 reliability estimate in this study is specific error. Specific error results from the fact that only one form of the IST is used. Each real or potential parallel form of the IST has some specific factor variance; that is, each form, in addition to measuring the construct of interest, also measures factors specific to that form; these factors result from the peculiarities of the content of that form. Specific factors are not part of the construct and hence are a form of measurement error. When two different parallel forms are correlated, the specific factors are assigned to measurement error in the resulting estimate of reliability because they are different in the two forms and do not correlate with each other and hence drop out.

Third, there is transient error. In this study, subjects respond to the IST on only one occasion; for each person, that particular moment in time is characterized by a certain mood, level of emotion, type of feeling, and so on. These factors are transient: A week later, or even a day later, they may be different, and hence the individual's responses may be somewhat different. The critical point is that these transient influences on the scores are not part of the construct (managerial motivation) that the researcher is trying to measure; they are a part of measurement error variance, not true variance.

Random response error and transient error can be controlled by administering the test to the subjects on two different occasions; the correlation between coders across occasions then provides an estimate of reliability that properly removes both transient error and purely random error from the estimate of construct variance (true variance). This more accurate reliability estimate is likely to be much smaller, as has been found to be the case for the employment interview (McDaniel, Whetzel, Schmidt, & Mauer, 1994). When two interviewers jointly interview each interviewee (i.e., both interviewers are present during the same interview), the average correlation between interviewers is .81. However, when the two interviewers interview the same applicants on two different occasions, the average correlation drops to .52. This is a large difference; apparently, about 29% of the variance of employment interview evaluation scores is due to random variations in questions and topics, random variations in interviewee responses, and perhaps transient errors. Only 19% (1.00 − .81) is due to scorer (coder) disagreement. The best estimate of the reliability of the typical employment interview is .52. The best way to increase this reliability is to have different interviewers interview candidates on different occasions and then average the interview scores across the occasions and interviewers.

The correlation between scorers across two occasions for the IST is the reliability for a score from one administration of the test. If the score to be used is the average across two occasions, the familiar Spearman-Brown formula can be used to compute this (higher) reliability.

Finally, specific error can be controlled by administering parallel forms of the IST on two occasions and computing the correlation between forms, across occasions, and across scorers. This estimate of reliability controls for all three types of measurement error.

These considerations are not just technicalities; they are critical to meaningful theory testing. For example, the study described here tested the hypothesis that there is a substantial true-score correlation between managerial motivation and age. The uncorrected correlation is .20, and use of the inflated .85 reliability estimate leads to a false estimate of .22 for the true-score correlation. The accurate estimate of the reliability of the IST scores, described above, would perhaps be around .40. This reliability figure would indicate that the true-score correlation is .32. This value might be large enough to lead the researcher to conclude that the hypothesis was supported, the opposite of her original conclusion.

Scenario 13

Situation. A researcher examining a theory of mental ability wants a precise estimate of the true-score correlation between vocabulary (one facet of verbal ability) and three-dimensional spatial ability (one facet of perceptual ability). He administers his test of vocabulary on one occasion, and 1 month later he administers his test of three-dimensional spatial ability. The researcher believes that because there is a 1-month interval between the administrations, he should make his corrections for attenuation using the coefficient of stability (test–retest using the same test; Cronbach, 1947) estimated with a 1-month interval between administrations. These reliability figures are presented in the test manual for each scale. The group on which these reliability estimates were computed has the same test standard deviations as the researcher's group, so he uses them to correct for the bias induced by measurement error.

Problem. This procedure controls for both transient error and purely random measurement errors, but it fails to control for specific error. The researcher's reliability estimate is the correlation between the same form of the test administered on two different occasions. Because the same forms are used both times, the influence of specific factors on the scores will be the same on both occasions. They will correlate with each other and hence inflate the reliability estimate, overestimating the proportion of variance in (both sets of) the scores that is due to construct variance. As a result, this researcher undercorrects, and his estimate of the true score correlation between vocabulary and spatial ability is downwardly biased. The correct estimate would be made using parallel forms of each test on each occasion. In some research domains, specific factor variance is quite large. For example, in the measurement of personality traits, it may account for 20% or more of score variance. In our research in the area of personality, we have observed that the correlation between the same form of a measure of a personality trait administered twice is often .20 or more higher than is the correlation between parallel forms of the measures (with the same time interval). In such domains, it is particularly important to control for specific factor error when conducting construct-based, theory-oriented research. In the measurement of specific abilities (such as verbal or spatial ability), specific factor error is usually smaller, typically accounting for 3% to 5% of score variance. However, if measures of different specific abilities are used and interpreted as measures of general mental ability, specific factor error variance in these measures of specific abilities in relation to the trait of general mental ability is larger, typically 10% to 20%. Hence specific error is important enough to require careful attention and control.

Scenario 14

Situation. A researcher is interested in the correlation between the constructs of mechanical ability and knowledge of electronics. She administers tests of each during a single session. On the basis of these data, she computes the KR-20 reliabilities of each test and the correlation between the tests. She then uses the KR-20 reliability estimates to correct the observed correlation and estimate the correlation between the constructs.

Problem. This procedure results in a close approximation to the correct answer. KR-20 reliability is the special case of coefficient alpha (Cronbach, 1951) that exists when item responses are dichotomous (i.e., items are scored correct or incorrect, agree or disagree, etc.). The KR-20 reliability formula assigns specific factor variance in items (and hence in the measure) to error; it also assigns purely random measurement errors to error (Cronbach, 1951). Transient error is not assigned to error in KR-20 reliability estimates, but transient error has repeatedly been found to be very small or nonexistent in the mental abilities domain. Thus when used to correct for the bias created by measurement error, the KR-20 reliability coefficient can be expected to generate accurate estimates of the true-score correlation.

Scenario 15

Situation. An applied researcher is developing a selection system for clerical workers at a large insurance company. He is aware of validity generalization findings from both civilian and military settings, showing that perceptual speed (clerical speed and accuracy) predicts performance in clerical jobs with substantial validity (Pearlman, Schmidt, & Hunter, 1980). He feels that the Minnesota Clerical Test is the best commercially published test of perceptual speed, but the company has asked him to develop its own measure. He reasons that he can have confidence in the scale he has developed if he can show that its true-score correlation with the Minnesota Clerical Test is 1.00 or nearly 1.00. Accordingly, he administers both his new perceptual speed test and the Minnesota Clerical Test to approximately 1,800 applicants to the company for clerical jobs and finds that the two tests correlate .81. He next computes the KR-20 reliabilities of each test on this same sample, obtaining a value of .96 for his test and .94 for the Minnesota Clerical Test. He then computes his estimate of the true-score correlation between the two tests as follows:

$$\hat{r}_{t_1t_2} = .81/\sqrt{.96(.94)} = .85.$$

He concludes that the true-score correlation of .85 departs too much from 1.00 to indicate that the two tests are measuring the same underlying construct. He states that because 15% of the true-score variance of each is not common to the other, the two tests are to some important extent measuring different constructs. He is puzzled by this finding, because the two tests appear to have very similar items and general content. Nevertheless, he decides that he must construct a second new test of perceptual speed that he hopes will show a larger true-score correlation with the Minnesota Clerical Test.

Problem. The true-score correlation of .85 is an underestimate. Tests of perceptual speed are highly speeded; in fact, they are often used as examples of the quintessential speed test. Perceptual speed tests are made up of very easy items, and it is usually assumed that with unlimited time every examinee could answer every question correctly. Hence they have very stringent time limits. Most perceptual speed tests used in clerical selection comprise name- and number-matching items.

The examinee must, for example, indicate whether two names are exactly identical or slightly different. The KR-20 reliability estimate cannot be used for measures that are substantially speeded, because speeding biases the reliability estimate upward. To estimate the reliability of speed tests, one must administer separately timed forms and correlate these. This provides an estimate of parallel forms reliability. Alternately, one can administer separately timed halves of the test and correct the resulting correlation using the Spearman-Brown formula. Unlike the KR-20 estimates, these reliability estimates are not upwardly biased.

Informed of these facts, this researcher estimated the reliability of each test on a new group of 1,150 clerical applicants by administering separately timed halves and correcting their intercorrelation with the Spearman-Brown formula. The resulting reliability estimate was .79 for his test and .87 for the Minnesota Clerical Test. His estimate of the true-score correlation between the two tests was then

$$\hat{r}_{t_1t_2} = .81/\sqrt{.79(.87)} = .98.$$

The researcher then concluded that his new test of perceptual speed measured the same underlying construct as the Minnesota Clerical Test, and he proceeded to incorporate his test into his new clerical selection battery for the company.

Scenario 16

Situation. A researcher is studying the relationship between measures of the personality trait conscientiousness and evaluations of "organizational citizenship behaviors" on the job. Coefficient alpha for the self-report measure of conscientiousness is .86. He uses supervisory ratings to measure the frequency of citizenship behaviors. He has the same two supervisors rate each subject on 20 different citizenship behaviors; the sum across both raters of these standardized ratings is the total citizenship score used in the study. The ratings are made independently, and the average correlation between raters (interrater reliability) for total score is .48. The researcher uses this figure of .48 to correct the observed correlations for the effects of measurement error in the measure of citizenship behavior. He uses the coefficient alpha value of .86 to correct for unreliability in the measure of conscientiousness. The uncorrected correlation is found to be .35, and the correlation corrected for measurement error in the ratings is .54.

Problem. Assuming that transient error is nonextant or minimal, the use of coefficient alpha for the measure of conscientiousness is correct. However, the use of .48 as the reliability of the ratings is in error. The value .48 is the reliability for ratings from one rater, while the researcher has used the sum (or average) of ratings from two independent raters. From the Spearman-Brown formula, the interrater reliability for this sum is $2(.48)/(1 + .48) = .65$. When the observed correlation of .35 is divided by $\sqrt{.65(.86)}$, the result is .47 rather than .54. Hence use of the improper interrater reliability inflated the true-score correlation estimate by 15%.

This researcher was testing a theory of conscientiousness and therefore was interested in construct-level relationships. Thus unbiased estimation required correction for measurement error in both variables. If the researcher's purpose had been to estimate the true validity of the conscientiousness measure as a selection device, there would be no correction for measurement error in the test, because observed scores from the test would have to be used in selection. The estimate of operational validity for the observed test scores is $\hat{r}_{xy_t} = .35/\sqrt{.65} = .43$.

Scenario 17

Situation. A researcher has developed her own measure of general mental ability for use in her research, and she has constructed two parallel forms of this measure. To estimate the reliability of these scales, she would like to administer the two parallel forms with 8 weeks between administrations but finds it impossible to arrange for this because of practical difficulties. However, she can arrange to have the test administered to a large group (of the appropriate composition) on one occasion, along with an older, well-established measure of general ability. On the basis of this administration, she finds that the two forms correlate .87. The KR-20 values for the two parallel forms are .86 and .88. One form correlates .88 with the established ability test, and the other correlates .86. Although these figures are very encouraging, she is disappointed that she was not able to obtain a direct estimate of the coefficient of equivalence and stability (Cronbach, 1947) by correlating the two forms over a time interval. She nevertheless concludes in her published article that the equivalence and stability reliability for her two forms are probably at least .85 and may be as high as .87.

Problem. On the basis of measurement theory alone, this conclusion may not be justified, because the estimates of reliability obtained by this researcher do not control for transient error or for any real changes in general intelligence that might occur over the 8-week period. However, on the basis of substantive cumulative knowledge from research on abilities, her conclusion is justified. It is a well-established fact that in the case of abilities and aptitudes, transient error (see Scenario 5) is small or nonexistent. It is also an established fact that no real changes in mental abilities occur over such short time periods. This can be inferred from the fact that in this domain estimates of the coefficient of equivalence—provided by KR-20 and parallel forms correlations at a single point in time—are usually very similar to estimates of the stability and equivalence reliability. That is, the insertion of a time interval, so long as it is of a reasonable length (e.g., 6 months or so for adults), does not reduce the reliability estimate below equivalence estimates.

This scenario illustrates the important fact that substantive research findings and cumulative knowledge can modify the interpretation of reliability estimates. Specifically, if enough evidence accumulates that a particular source of measurement error is negligible in a particular domain, then that source can subsequently be ignored for certain purposes. Research evidence indicates that transient error is apparently not very important in the abilities area, especially for general mental ability.

Scenario 18

Situation. A researcher is using a large primary database developed by a national polling organization to study the attitudes of the American public toward labor unions. He is particularly interested in the construct of general favorability (general evaluative attitude) and its relation to the value that people place on egalitarianism. The database contains six questions assessing the respondents' general evaluative attitude toward unions, and the reported coefficient alpha for these items is .83. After examining these six items, the researcher concludes that they are all similar to each other and are in fact redundant with each other. To eliminate this redundancy in his own research, he decides to use only the best question: the item that he judges on the basis of content to have the greatest construct validity. He then correlates this measure of general favorability toward unions with the measure of egalitarianism and uses the alpha coefficients to correct these correlations for the downward bias created by measurement error in the measurement of attitudes. The value of egalitarianism is measured by a 10-item scale that has an alpha coefficient of .73, and the researcher uses this value in making the corrections for measurement error in the dependent variable. The researcher's hypothesis is that people who place a strong emphasis on the value of egalitarianism will have more favorable attitudes toward unions.

Problem. The reliability (coefficient alpha) used for the measure of general attitude toward unions is the reliability of the sum or average of the six questions. But the observed correlations are based on measurement of this attitude using only the "one best item." Hence the reliability estimate is too large, and the corrected correlations are downwardly biased estimates of the true-score correlation. Because the reliability coefficient is too large, it undercorrects the observed correlation.

What can be done to solve this problem? The Spearman-Brown formula can be used (in reverse) to compute the reliability of one item from the known reliability (.83) of six items, and this estimate could be used to correct the observed correlations. Although this procedure will produce unbiased estimates, it is not the optimal approach. The best approach would be to go back to the data and recompute the observed correlations using the sum or average of scores from all six questions. Because the observed correlations are then based on all available data, observed correlations will be larger and will have smaller standard errors. Because the summed scores that are based on all six questions have higher reliability, there will be a smaller correction, and hence the corrected correlation will have less sampling error. That is, both the observed and the corrected correlations will have less sampling error (and therefore a smaller standard error). In short, although the practice of dropping items, questions, or raters is seen with some frequency, it is bad practice to discard data. This researcher dropped five of the six questions, not because he thought they were invalid, but just to make data analysis ostensibly easier by eliminating "redundant" data. These items are not redundant; they increase the reliability of the measure used and thus contribute to increasing the stability of the final research findings.

Scenario 19

Situation. A researcher seeks to test the hypothesis that organizational commitment and job involvement are distinct constructs. He seeks to demonstrate this by showing that the true-score correlation between measures of these two variables is less than 1.00. His argument is that, while they may be substantially—even highly—correlated, they are not perfectly correlated, even at the true-score level and even in large, representative, and heterogeneous groups of employees. He arranges to administer measures of both variables to a large group (total $N = 1{,}983$) of employees in a major service firm. On the basis of responses from the administration at one point in time, he computes coefficient alpha for each scale and uses these reliability values to correct the correlation between these measures computed in the total group of nearly 2,000 subjects. The corrected correlation is .82; on the basis of this finding, he concludes that organizational commitment and job involvement are similar but distinct constructs.

Problem. Coefficient alpha is an alternate-form reliability estimate. It treats each scale item as if that item were an alternate form of a test for the construct. This analytic strategy captures the effects of random response error and specific error but does not control the effects of transient error. That is, transient error is treated as part of true scores. If either organizational commitment or job involvement is subject to transient error, then coefficient alpha will overestimate the reliability of that measure. The corrected correlation would thus be underestimated. Thus the corrected correlation of .82 could be lower than the actual correlation between these constructs. If the true-score correlation is actually 1.00, then the conclusion this researcher reached would be false.

There are variables whose measurement has been found to be virtually unaffected by transient error; examples include aptitudes, abilities, and job performance ratings. For such variables, the reliability estimate used by this researcher would be a good way to handle error of measurement. However, it is not clear whether this is true of organizational commitment or job involvement.

There are those who believe that transient error can be ignored. They would say, "But if we want to know the person's organizational commitment on a given day or at a given moment, then transient variation would not be error and coefficient alpha would be the correct reliability estimate." However, examination of the phrase *organizational commitment* reveals that this criticism depends sharply on the substantive nature of the transient error process.

Consider one hypothesis of transient error that all investigators would regard as "error variance" if true. Suppose that organizational commitment feelings on a given day depend on the person's emotional state at that moment. In particular, suppose that commitment depends on the person's level of physical energy. Suppose that when people feel energetic, they look forward to work and have higher commitment feelings and that when they are sick and feel lethargic, they have lower commitment feelings toward work. Would any theorist really want to define the important construct of organizational commitment in such a way that commitment is lower if the person has a cold than if the person is well? Probably not. That is, it would be obvious that transient variation is measurement error and not part of the construct itself.

By contrast, there are constructs for which we are interested in day-by-day fluctuations. For such constructs, day-to-day instability is not an error process and would thus not contribute to the attenuation of correlations with other constructs, especially if the other construct is also defined on a day-by-day basis. For example, consider the correlation between sleep deprivation and negative mood. If a person says, "I'm irritable when I don't get enough sleep," that person is specifically talking about day-to-day fluctuations in both sleep level and mood state. In this case, the day-to-day fluctuations are not random error but rather the focus of the correlation.

On the other hand, consider the use of the item "Do you sleep well?" as a measure of anxiety on a personality scale. The use of this item stems from the known fact that people suffering from anxiety have trouble sleeping. If one is measuring the stable trait of anxiety, then the time reference to the item is "Generally speaking, do you sleep well?" If one asks instead "Did you sleep well last night?," then the day-to-day fluctuations in sleep will act as transient error in measuring anxiety.

In most contexts, transient error is usually due to "trivial" factors that we would want to exclude from our definition of the construct. Thus we cannot dismiss transient error as irrelevant without knowing the substantive nature of the transient error.

In summary, in this research scenario, if the measurement of organizational commitment and job involvement is subject to little or no transient error, then the procedures used by this researcher to estimate the true-score correlation between these two constructs are correct and the researcher's estimate of .82 is appropriate. If transient error is a significant factor in these measures, then the two reliabilities will be overestimates and the estimated true-score correlation will be an underestimate.

What research could be done to find out whether transient error is important here? Consider the organizational commitment scale. If parallel forms of that scale were administered a week or so apart, then the correlation between the two administrations would be an estimate of reliability in which transient variation is controlled for (i.e., is assigned to measurement error). If this value is the same (or nearly the same) as the coefficient alpha computed earlier by the researcher, then we know that transient error is unimportant in this measurement domain and can be ignored. In this case, we would conclude that the researcher's procedures were correct.

On the other hand, if the correlation between the two administrations is considerably smaller than coefficient alpha (say, .60 vs. .85), then we know that transient error is substantial here. We also know that the alpha coefficients are inflated by transient error and cannot be used in making the corrections for measurement error. Instead, the corrections between the two administrations of the parallel forms must be used in making the corrections. If this is the case, then the actual true-score correlation between organizational commitment and job involvement is larger than the .82 estimated by this researcher.

There are many areas like this one in which the research required to find out whether transient error is important has not been conducted. Researchers in these areas typically assume without any evidence that it is of no importance. This is bad research practice. Research of the kind described here is needed in these areas to provide the answers needed to improve research practices.

Scenario 20

Situation. An industrial–organizational psychologist develops a new type of employment interview, the Contextual-Behavioral Interview (CBI). Because the process of conducting this interview is complex, the first question she seeks to answer is whether even trained interviewers can conduct these interviews reliably. She is also concerned that the CBI might be suspected of being an orally administered intelligence test. She decides to proceed by correlating CBI total scores with a well-respected intelligence test and correcting this correlation for attenuation due to measurement error to estimate the correlation with the construct (trait) of intelligence. She trains two psychologists in the use of CBI interviews and then has them interview 243 job applicants. Both interviewers are present at each interview, and each alternately asks questions for different segments of the interview while the other observes. The interviewers tabulate their final interview scores independently (without discussion between them). The correlation between the two interviewers is .81, and on the basis of this finding, she concludes that the CBI interview has adequate reliability.

The correlation between the intelligence test and interview scores is .67 for each interviewer. The KR-20 reliability of the general mental ability test is .86. Her estimate of the true-score correlation between interview scores and intelligence is

$$\hat{r}_{x_t y_t} = .67/(.81 \times .86)^{1/2} = .80.$$

On the basis of these findings, she concludes that (a) CBI scores are reliable and (b) the CBI measures more than just mental ability. She states that her results indicate that 36% of the true-score variance of interview scores ($1.00 - .80^2$) reflects factors independent of mental ability. Thus the CBI interview is more than just an orally administered ability test.

Problem. The estimate of reliability used for the CBI interview is inflated, leading to down-the two constructs. As is explained in Scenario 12, much of the variance of employment interview scores consists of random response measurement error or transient error. On the basis of data from 210 samples, McDaniel et al. (1994) reported that the average correlation between interviewers sitting in on the same interview was .81, the value obtained here. However, when the interviewers conducted their interviews on different days, the average correlation was only .52. Thus random response error (and perhaps transient and specific factor error) causes considerable instability in interviewees' responses from occasion to occasion. These unstable influences on scores cannot be considered to be part of the construct or constructs measured by the interview; they must be considered error of measurement. The best available estimate of the real reliability of CBI scores is the .52 figure from McDaniel et al. (1994). Thus the CBI is probably not nearly as reliable as the researcher concluded. When .52 is used in the attenuation correction formula as the reliability of the CBI, the estimated true-score correlation is

$$\hat{r}_{x_t y_t} = .67/(.81 \times .52)^{1/2} = 1.00.$$

This suggests that the CBI really is nothing more than an orally administered test of general mental ability.

Scenario 21

Situation. A researcher is concerned that industrial–organizational psychology has drifted away from its previous focus on real job behaviors and has come to rely too heavily on ratings and questionnaire measures (often self-reports) as a substitute for actual behavior on the job. She believes that research should be based more heavily on observation of actual job behavior than is the case today. She argues that such measures would not only be more valid, they would also be more reliable. She reported the following research to support the hypothesis that reliability would be higher. She assigned pairs of trained observers to small groups of workers; these observers recorded eight different categories of actual worker behaviors for a period of 4 hr. Across the behavior categories, the average correlation between the two observers was found to be .89. On the basis of this finding, she concluded that focusing on actual job behaviors leads to much higher reliabilities than can be obtained with supervisory ratings of job performance.

Problem. The reported figure of .89 is intercoder reliability, often referred to as *conspect reliability* or *scorer reliability.* Some of the problems associated with this form of reliability were discussed in Scenario 12. For purposes of the present scenario, the critical problem is that scorer reliability fails to control for random response error. The two observers in this research coded the behavior of the same workers at the same time. They appear to have coded these behaviors quite accurately, but these behaviors themselves might be quite unstable from day to day. On the basis of findings in other areas, we can predict that if these two observers had coded the behaviors on different days, the correlation between their evaluations would have been considerably below the reported .89, perhaps in the neighborhood of the average correlation between job performance ratings by two supervisors (.50; Rothstein, 1990).

Scenario 22

Situation. A psychologist seeks to introduce "multiskilling" for two jobs performed on the production floor. His goal is to have each employee currently performing Job A learn to perform Job B, and vice versa, so that any worker can fill in for people in either job when necessary. At present, employees in each job have observed the other job but not performed it. Before the cross-training, he decides to collect baseline data on how much job knowledge each employee currently has for both jobs, so he constructs and administers content-valid job knowledge tests for both jobs to all employees on the two jobs. Because knowledge of either job depends on cognitive ability, he hypothesizes that those employees whose job knowledge is highest for their own job will also tend to have acquired more knowledge about the other job (even though they have only observed the other job and have not done it). To test this hypothesis, he correlates scores on the two tests, finding a correlation of .52 for the Job A incumbents. The correlation for the Job B incumbents is also .52. He knows that the real relationship is stronger than this because unreliability in his two job knowledge tests has reduced this correlation somewhat. To estimate the real relationship, he computes the KR-21 reliabilities of each test and corrects for measurement error:

$$r_{x_t y_t} = .52/(.81 \times .79)^{1/2} = .65.$$

Problem. The .65 is an upwardly biased estimate of the true extent of the relationship, because the KR-21 reliability formula underestimates the reliability of the job knowledge tests. The KR-21 formula, unlike the KR-20 formula, assumes that all test items are equal in difficulty (i.e., are all answered correctly by the same percentage of examinees). This is rarely true in any domain, and it is almost never true for job knowledge tests. Job knowledge tests almost always contain a wide range of item difficulties. When items vary in difficulty, KR-21 underestimates reliability, leading to overcorrections for measurement error. Told about this, the researcher computed the KR-20 reliabilities and used these to correct the observed correlation:

$$r_{x_t y_t} = .52/(.85 \times .84)^{1/2} = .61.$$

Thus the appropriate (unbiased) estimate of the true-score correlation is .61. The .65 obtained earlier was an overestimate by almost 7%.

Computing the KR-20 reliability estimate requires access to item-level data; one needs to know the difficulty (or endorsement rate) for each item. This information is not required for KR-21 (only the average item difficulty need be known). KR-21 is sometimes used because item-level data are not available. In such cases, the researcher should note that the KR-21 estimates are underestimates of reliability.

Measurement Error in More Complex Cases

Scenario 23

Situation. A researcher is attempting to validate predictors of voluntary absenteeism from work, which he measures over a period of 1 year. He finds that the year-to-year correlation of absenteeism is .70; that is, the reliability of absenteeism over 1 year is .70. However, the researcher decides not to correct the observed validity coefficients for unreliability in the absenteeism criterion. He argues that the departure of the .70 from 1.00 does not reflect error of measurement but rather is due to real changes in the behavior. That is, he reasons that the .70 merely reflects reality. He states that corrected validities would be overestimates, because in the real world there is no such thing as perfectly stable (perfectly reliable) absenteeism behavior.

Problem. This researcher argues that the instability observed in the measure of absenteeism over 1 year is due to real (relative) change among employees and not to measurement error. However, he presents no evidence to support this contention, and therefore his statement is merely a hypothesis. There is an empirical test that can be applied to determine whether there has been real change. This test requires that absenteeism be measured at three points in time. For example, we could measure absenteeism in each of 3 subsequent years. If no real change has taken place, then if there were no sampling error, we expect that $r_{13} = r_{12} = r_{23}$. For example, if we find that $r_{13} = .70$, $r_{12} = .70$, and $r_{23} = .70$, the data would indicate that no real change has taken place. This would indicate that all of the instability in the measure is due to measurement error. This means that the researcher's hypothesis is wrong and that his validity coefficient should be corrected using his .70 reliability figure (obtained for a 1-year period).

If employees' relative standing on absenteeism is really changing over time, then we expect to find that $r_{13} < r_{12}$ and $r_{13} < r_{23}$. For example, suppose we find that $r_{13} = .61$, while $r_{12} = .70$ and $r_{23} = .70$. This finding suggests that some real change in absenteeism may be occurring over time. However, this finding does not imply that all of the instability of absenteeism scores is due to real change. The departure of the researcher's .70 figure from 1.00 is to some extend due to purely random errors of measurement: failures to record every absence, absences by one employee erroneously recorded for another employee, involuntary absences mistakenly recorded as voluntary absences, and so on. These measurement errors are artifacts, and their effect is to downwardly bias the observed validity. A correction should be made for this bias. Under these conditions of real change, the reliability for Time Period 2 is

$$r_{12} = \frac{r_{12}r_{23}}{r_{13}} = \frac{.70(.70)}{.61} = .80.$$

This reliability coefficient should be used in making the correction for measurement error. Notice that this reliability is higher than the .70 that holds for a situation in which there is no real change over time.

Since real change has taken place, the true validity could be different for different periods. Therefore, observed validities should be computed separately for each time period and then corrected using the reliability computed for that time period. If the estimated true validities are essentially identical for each time period, then we conclude that the changes over time in absenteeism do not affect the validity of the predictor for predicting absenteeism.

Finally, we note in passing that real change over time in individuals' relative standing on a trait or other dimension has no necessary effect, positive or negative, on the correlation between that dimension and another variable or measure (cf. Ackerman, 1989; Rogosa & Willett, 1985; Schmidt, Ones, & Hunter, 1992, pp. 645–646). This means that real change over time in job performance does not necessarily reduce the validity of predictors of job performance.

Scenario 24

Situation. A social psychologist testing a theory of personal values hypothesizes that each of 10 different "peripheral" values are positively related to the "core" value of individual freedom. Each of these 11 values is measured by its own nine-item Likert scale; each of these 11 scales has a coefficient alpha reliability of .75 or higher. The researcher computes the correlations between the 10 peripheral values and the core value in his research sample. However, his statistical analysis shows that only 2 of the 10 correlations are statistically significant. His interpretation of this is that 8 of the peripheral values are apparently unrelated to the core value and that the theory is therefore not supported. However, after some thought the researcher concludes that in the context of the theory under study, it is really the true-score correlations that are important, not the observed correlations. He then corrects the 10 correlations for measurement error. Applying the same significance test, he finds that the corrected correlations are all significant beyond the .01 level. He concludes that these findings of significant relationships between peripheral values and the core values provide strong support for his theory.

Problem. The conventional statistical significance tests applied to the estimated true-score correlations are not legitimate, because the correction for attenuation increases the standard error of the correlation coefficients. As a general rule, corrections for measurement error cannot make a correlation, or other zero-order statistic, statistically significant if it is not statistically significant before the correction. For example, correcting a correlation coefficient from .25 to .35 increases it by 40%. However, its standard error also increases by 40%, so its level of statistical significance (p value) remains constant. If the appropriate standard error for the corrected coefficient is computed and used in the significance test, that significance test is statistically correct. However, it yields the same outcome as the conventional significance test conducted on the observed (uncorrected) correlation. Hence in this study, only the significance tests conducted on the uncorrected correlations are legitimate. In this analysis, only 2 of the 10 correlations were statistically significant.

However, this does not mean that the other 8 are zero or are best considered to be zero. The operational rule of thumb in the social sciences has long been, "If it is not significant, it is zero." However, this rule of thumb is erroneous and leads to frequent errors in data interpretation (Schmidt, 1992, 1994), especially in situations of low statistical power. The researcher in this case appears to follow this erroneous rule of thumb. He seems to assume that his nonsignificant correlations indicate the absence of a relationship. This researcher should look at the magnitude of the estimated true-score correlations. Although the significance test he conducted is erroneous, the estimated true-score correlations themselves are unbiased estimates of the magnitude of the relationships. If these correlations are small, the theory is indeed not supported, but if they are substantial, that finding would support the theory. The absence of statistical significance is in that case probably due to low statistical power. Further research can then test this hypothesis.

Scenario 25

Situation. A researcher seeks to determine whether two different measures of satisfaction with pay are interchangeable. The manuals for each of these instruments provide estimates of coefficient alpha computed on large samples (*N*s > 1,500) of appropriate composition, and the alphas are similar in magnitude. She administers both measures to a sample of 31 employees and uses the alphas from the manuals to correct this correlation to estimate the construct-level correlation. To her surprise, the corrected correlation is 1.23. She decides that there might be serious problems with attenuation corrections and vows never to make such corrections in her research again.

Problem. The problem here is not that the wrong reliability estimates were used to correct for bias induced by measurement error; assuming no transient error, the use of coefficient alpha is correct. Rather, this is a case of failure to understand sampling error. When the sample is as small as $N = 31$ (and even for much larger *N*s), sampling error will often cause the observed correlation to depart from its population value by a substantial amount. (Sampling errors are much larger than most researchers realize. Examples illustrating this point can be found in Hunter & Schmidt, 1990, chaps. 1 & 2; Schmidt, 1992; and Schmidt, Ocasio, Hillery, & Hunter, 1985.) If the observed correlation happens to be a low random bounce, the corrected correlation will be underestimated. In this case, it might have been .90. If this had happened, this researcher would probably not have been shocked. However, if the observed correlation happens to be a high random bounce, the estimate of the true-score correlation can be greater than 1.00. If the true-score population correlation is actually 1.00 (as is probably the case here), this is to be expected about half the time. That is, purely because of ordinary sampling error, about 50% of the sample estimates of the true-score correlation should be above 1.00. If the actual true-score population correlation is lower, say .80, then the expected percentage of sample estimates above 1.00 is less than 50%. In either case, such sample estimates pose no problem. Since the correlation cannot exceed 1.00 by definition, such estimates are simply rounded down to 1.00. As we noted in the introduction to this article, the correction for attenuation, properly applied, is perfectly accurate in the population (i.e., when $N = \infty$). When N is less than infinite, corrected correlations, like uncorrected correlations, are estimates and contain sampling error. Sampling error can cause computed estimates to be larger than 1.00.

Scenario 26

Situation. A researcher hypothesizes that there is a positive relation between job satisfaction and the construct of "role adjustment." She obtains measures of both constructs on a sample of 243 appliance salespeople. The observed correlation is .25, and the 95% confidence interval (CI) is .13 to .37. Using coefficient alpha reliabilities for each scale, she corrects the observed correlation for measurement error:

$$\hat{r}_{x_t y_t} = .25/(.66 \times .73)^{1/2} = .36.$$

She concludes that her hypothesis is supported. She states that the best estimate of the actual relationship is .36 but that the true value could be as low as .13 or as high as .37.

Problem. The researcher is correct in stating that the best point estimate of the true-score correlation is .36. However, her statement about the confidence interval is erroneous. The confidence interval used applies to the uncorrected correlations; it does not apply to the corrected correlation. The confidence interval for the corrected correlation can be estimated by correcting the endpoints of the confidence interval for the observed correlation:

$$\text{Lower CI endpoint} = .13/(.66 \times .73)^{1/2} = .19,$$
$$\text{Upper CI endpoint} = .37/(.66 \times .73)^{1/2} = .53.$$

Thus the correct statement is that the best estimate of the true-score correlation is .36 but that the 95% confidence interval is from .19 to .53.

Concluding Remarks

These scenarios involving measurement error problems in psychological research reflect real situations that we have encountered in our own research and that of others. In addition, they reflect the situations we have encountered most frequently. Because of this, we feel that they are more realistic, informative, and useful than abstract psychometric discussions of measurement theory and reliability principles. Over the years, many bad methodological practices have retarded progress in the development of cumulative knowledge in psychology and the other social sciences (Hunter & Schmidt, 1990). Failure to attend properly to measurement error is perhaps not the worst of these; that honor probably goes to the practice of relying on statistical significance tests in interpreting research data (Hunter & Schmidt, 1990; Schmidt, 1992, 1994). But measurement error is at least second in importance. Research progress in theory development and cumulative knowledge is impossible without careful attention to problems created by measurement error. We believe that these real-life, concrete examples provide useful guides to working researchers grappling with the real-world complexities of their own research programs.

References

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

Berenstein, M. (1994). The case for confidence intervals in controlled clinical trials. *Controlled Clinical Trials, 15,* 411–428.

Campbell, A., Converse, P. E., Miller, W. E., & Stokes, D. E. (1960). *The American voter.* New York: Wiley.

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

Cronbach, L. J. (1947). Test reliability: Its meaning and determination. *Psychometrika, 12,* 1–16.

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

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: Macmillan.

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

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Newbury Park, CA: Sage.

King, L. M. Hunter, J. E., & Schmidt, F. L. (1980). Halo in a multidimensional forced-choice performance evaluation scale. *Journal of Applied Psychology, 33,* 507–516.

Klimoski, R. J. (1993). Predictor constructs and their measurement. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 99–134). San Francisco: Jossey-Bass.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48,* 1181–1209.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Mauer, S. (1994). The validity of employment interview: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79,* 599–616.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training criteria in clerical occupations. *Journal of Applied Psychology, 65,* 373–407.

Rogosa, D., & Willett, J. B. (1985). Satisfying a simplex structure is simpler than it should be. *Journal of Educational Statistics, 10,* 99–107.

Rothstein, H. R. (1990). Interrater reliability of job performance ratings: Growth to asymptote with increasing opportunity to observe. *Journal of Applied Psychology, 75,* 322–327.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47,* 1173–1181.

Schmidt, F. L. (1993). Personnel psychology at the cutting edge. In N. Schmitt & W. Borman (Eds.), *Personnel selection in organizations* (pp. 497–516). San Francisco: Jossey-Bass.

Schmidt, F. L. (1994, August). *Quantitative methods and cumulative knowledge in psychology: Implications for the training of researchers.* Division 5 presidential address presented at the Annual Convention of the American Psychological Association, Los Angeles, CA.

Schmidt, F. L., & Hunter, J. E. (1992). Development of a causal model of processes determining job performance. *Current Directions in Psychological Science, 1,* 89–92.

Schmidt, F. L., Hunter, J. E., & Caplan, J. R. (1981). Validity generalization results for two jobs in the petroleum industry. *Journal of Applied Psychology, 66,* 261–273.

Schmidt, F. L., Ocasio, B. P., Hillery, J. M., & Hunter, J. E. (1985). Further within-setting empirical tests of the situational specificity hypothesis in personnel selection. *Personnel Psychology, 38,* 509–524.

Schmidt, F. L., Ones, D. S., & Hunter, J. E. (1992). Personnel selection. *Annual Review of Psychology, 43,*627–670.

Schmitt, N., & Landy, F. (1993). The concept of validity. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 275–309). San Francisco: Jossey-Bass.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47,* 149–155.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 356–442). Washington, DC: American Council on Education.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 560–620). Washington, DC: American Council on Education.

Traub, R. E. (1994). *Reliability for the social sciences.* Newbury Park, CA: Sage.

Received February 21, 1995
Revision received June 19, 1995
Accepted June 28, 1995 ■