EDITORIAL

# Theory Testing and Measurement Error

Frank L. Schmidt
*University of Iowa, Iowa City, IA, USA*

John E. Hunter
*Michigan State University, East Lansing, MI, USA*

Accurate empirical tests of theories and hypotheses are not possible unless the inevitable biases induced into data by measurement error are controlled for. Yet despite 90 years of recommendations from measurement theory and methodology, some still do not control for these biases in their research. This paper presents simple and direct demonstrations showing why basic measurement principles require that biases in data created by measurement error be removed and refutes commonly heard objections to the corrections for these biases. One factor contributing to resistance on the part of some researchers is the fact that most psychologists are not aware that measurement error is produced by real psychological processes that can be studied and understood. This paper describes those substantive psychological process and shows how each generates a different type of measurement error. We also show how different types of reliability estimates assess and calibrate different error processes and types of measurement error, leading directly to conclusions about which types of reliability estimates are appropriate for measurement error corrections in different research settings. Failure to control for biases induced by measurement error has retarded the development of cumulative research knowledge. It is our hope that this paper will contribute to removing these hobbles from psychological research.

In the physical sciences, measurement error has long been the focus of sustained attention and examination (Fuller, 1987; Hedges, 1987). The reason for this is simple: It is not possible to have accurate empirical tests of theories and hypotheses unless the biases introduced into data by measurement error are controlled and corrected for. In particular, a major goal in all areas of science is the calibration of construct level relations. These relations are the building blocks of theory. All measures of constructs—whether in the physical or social sciences—contain measurement error. There is no such thing as

Direct all correspondence to: Dr. Frank Schmidt, Department of Management and Organization, College of Business, University of Iowa, Iowa City, IA 52242, USA. E-mail: frank-schmidt@uiowa.edu

**INTELLIGENCE 27(3)**: 183–198
ISSN: 0160-2896

errorless measurement. As a result, all observed relations are relations between specific measures, not relations between constructs, and are therefore biased estimates of relations between constructs. Hence, the need in research to address the ubiquitous problem of measurement error and to correct for the biases created by measurement error.

The increasing use of structural equation modeling (SEM) in general differential psychology (including behavior genetics) has helped to bring this point home to researchers. In fact, the built-in correction for the distorting effects of measurement error is often cited as one of the major advantages of SEM. But, in other research, there is often still a lingering reluctance to eliminate biases in data created by measurement error. Yet, if it is essential to control for biases induced by measurement error when using SEM to test causal models (i.e., theories), how could it be unneeded or inappropriate in the calibrating of bivariate relationships?

This journal is devoted to differential psychology—the differential psychology of human cognitive abilities. More so than in most other areas of psychology, the fundamental research tool in differential psychology is psychological measurement and measurement theory. Indeed, without measurement theory and methods, differential psychology as we know it would scarcely even be possible. This fact points up an anomaly: There is a contradiction between the prescriptions of psychometric theory and methods and the everyday practices of many differential psychology researchers. Since the early 1900s, the psychometric methodological literature has consistently stated that correction for measurement error is critical to accurate calibration of scientific quantities and to the evaluation of scientific theories. Yet, many currently published studies still do not address or discuss measurement error and its effects on the reported research results. There have long existed major pockets of resistance within differential psychology to this essential prescription. Hence, the need for the present editorial.

There are, of course, other artifacts in addition to measurement error that cause biases and distortion in research data. Two of the most common ones are range restriction or range enhancement and dichotomization of continuous measures. Range restriction and dichotomization both cause downward biases in observed correlations, while (artificial) range enhancement causes an upward bias. These problems have been addressed in the literature (e.g., Linn, Harnisch, & Dunbar, 1981; Hunter & Schmidt, 1990a,b; Ree, Carretta, Earles, & Albert, 1994), but, nevertheless, it might be useful for a future editorial to address the corrections needed for these biases. However, for reasons that appear to be more emotional than rational, these corrections do not seem to be resisted as much as corrections for biases induced by measurement error—the subject of the present paper.

## Basic Principles

Later, we will examine some of the objections that have been advanced against eliminating biases in data caused by measurement error. But, first, we want to illustrate the nature of the problem created for theory testing by the biases produced by measurement error. In classical measurement theory, the fundamental general formula for the observed correlation between any two measures $x$ and $y$ is:

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

Where $r_{xy}$ is the observed correlation, $r_{x_t y_t}$ is the correlation between the true scores underlying the measures, $r_{xx}$ and $r_{yy}$ and are the reliabilities, respectively, of the $x$ and $y$ measures. Eq. (1) is called the *attenuation formula*, because it shows how measurement error in the $x$ and $y$ measures reduces the observed correlation ($r_{xy}$) below the true score correlation ($r_{x_t y_t}$).

Solving Eq. (1) for $r_{x_t y_t}$ yields the *disattenuation formula* (Eq. [2]):

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

If the sample size is infinite (i.e., in the population), both these formulas are perfectly accurate. In the smaller samples used in actual research, there are sampling errors in the estimated values 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 often used to indicate that all value are estimates:

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

The $\hat{r}_{x_t y_t}$ is the estimated correlation between the construct underlying the measure $x$ and the construct underlying the measure $y$. Alternatively, it is an estimate of the (uncorrected) correlation that would be observed between the measures $x$ and $y$ if both measures could be made free of measurement error.

These are fundamental equations in classical measurement theory, the measurement model used in probably 95% of research in differential psychology. The more complex alternative measurement model, item response theory (IRT), can be used to show 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 relation is usually close enough to linear that the effect on the estimated construct level correlations of taking true scores as collinear with the construct is negligible. This fact is very fortunate; if this were not the case, it would be necessary to use the more complicated and difficult IRT measurement model in research. The two models agree in stating that corrections for measurement error are essential for accurate research results. However, the correction process is much more difficult to understand and apply in the IRT model.

Consider a question that occurs in the study of intelligence: What is the correlation in the general population between perceptual speed (PS) and general mental ability (GMA)? Suppose two different researchers set out to answer this question, each using an $N = 3000$ representative sample from the general population. The first researcher reports $r = 0.45$, while the second reports $r = 0.30$. The value reported by the first study is 50% larger than that yielded by the second. Since both samples are large and representative, this huge difference is not due to sampling error. Obviously, this sort of conflict in the research literature is troubling. Yet, problems of this sort can be produced by simple failure to consider and control for measurement error.

Consider the following explanation. The first researcher is careful to use only the most reliable scales: both his scales have reliability of 0.90. The second researcher uses much shorter—and hence less reliable—scales, in order to be able to measure more constructs in the limited testing time available. His scales both have reliability of 0.60. This

**Table 1.** Average Value of $r_{xy}$ as a Function of $r_{xx}$ and $r_{yy}$ When the Actual Correlation between Constructs ($r_{x_t y_t}$) is 0.50

| *Reliability of Measure y* | *Reliability of Measure x* | | | | | |
|---|---|---|---|---|---|---|
| | *0.40* | *0.50* | *0.60* | *0.70* | *0.80* | *0.90* |
| 0.40 | 0.20 | | | | | |
| 0.50 | 0.22 | 0.25 | | | | |
| 0.60 | 0.25 | 0.27 | 0.30 | | | |
| 0.70 | 0.27 | 0.30 | 0.32 | 0.35 | | |
| 0.80 | 0.28 | 0.32 | 0.35 | 0.37 | 0.40 | |
| 0.90 | 0.30 | 0.34 | 0.37 | 0.40 | 0.42 | 0.45 |

contradiction between these two findings disappears if we apply Eq. (3) in each case to correct for the downward bias due to measurement error:

$$0.45/[0.90(0.90)]^{1/2} = 0.50$$

$$0.30/[0.60(0.60)]^{1/2} = 0.50$$

We now see that the construct level correlation—which is the theoretically relevant correlation—is 0.50 in both studies. That is, the correlation between the two *abilities*—as opposed to *measures*—is 0.50 in both studies.

Suppose that instead of two studies we had a large number of such studies. Then, each study could be characterized by a different combination of reliabilities. Hence, every study could report a different observed correlation—producing even more confusion and contradiction. Table 1 illustrates this situation. Table 1 shows that if reliabilities for both scales vary from 0.90 down to 0.40, reported correlations in individual studies will vary anywhere between 0.20 and 0.45. It is obviously unscientific to say that the relation between PS and GMA depends on what specific scales happen to be used in a particular study.

Table 1 is produced by applying Eq. (1)—the attenuation formula—to a constant value of $r_{x_t y_t}$, using different combinations of $r_{xx}$ and $r_{yy}$. Applying Eq. (3)—the disattenuation formula—to Table 1 shows that all values of $r_{x_t y_t}$ are 0.50.

Table 1 shows 21 different values assumed by $r_{xy}$. Actually, if the reliability of each measure varies between 0.40 and 0.90 in steps of 0.01, then there are 1275 different expected values of $r_{xy}$. Some of these values will be identical, but most will be different. Thus, Table 1 actually understates the problem.

This example (and Table 1) illustrates the simple bivariate case. In the bivariate case, the effect of measurement error is always in the same direction: it reduces the size of the observed correlation; i.e., it produces a downward bias. But, in the multivariate case, in which more than one independent variable is measured, the biases created by measurement error can be in either direction. Path coefficients or regression weights can be biased upward or downward—and there is no way to tell in advance what the direction of bias will be. This means that it is even more important to correct for these biases.

For example, Schmidt, Hunter, and Caplan (1981) found a positive bias in several regression weights and a negative bias in another. These researchers asked the following question: Do specific aptitudes, such as verbal ability, spatial ability, and quantitative

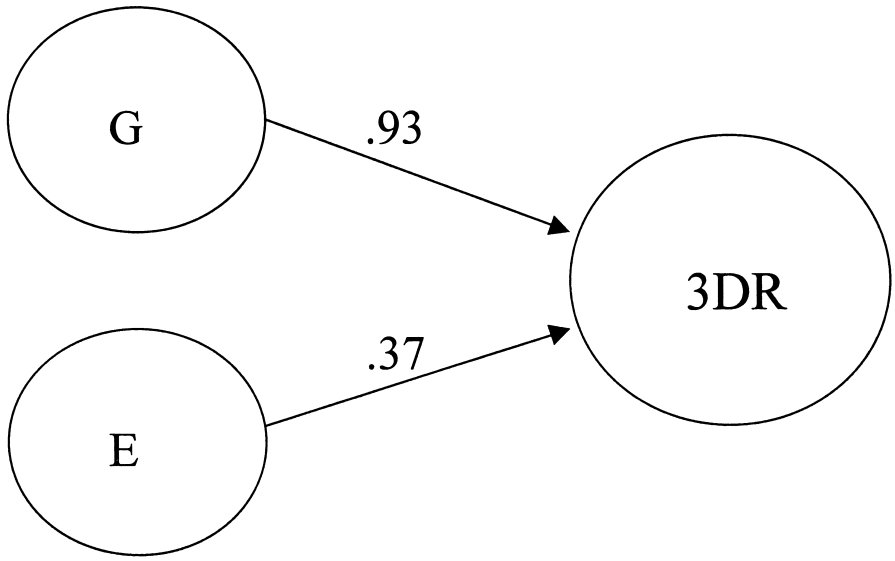


**Figure 1.** Causal impact of genes (G) and environment (E) on the ability 3DR when $h^2 = 0.86$ and $e^2 = 0.14$.

ability, add to the prediction of job performance over and above the prediction produced by general mental ability (GMA)? If these specific aptitudes get positive standardized regression weights in a multiple regression equation that includes GMA, then the answer is yes. At the observed score level, this was in fact the case, and it appeared that specific aptitudes made a contribution over and above that of GMA. However, they found that these positive regression weights were produced solely by positive bias caused by measurement error. When all zero order correlations were corrected for measurement error [using Eq. (3)] before running the regression, the regression weights on the specific aptitudes were all zero, indicating that at the construct (ability) level, these specific abilities made no contribution over and above that of GMA. In this case, measurement error created a downward bias in the regression weight on GMA and an upward bias in the weights for the specific aptitudes. Failure to correct for measurement error would have led to substantively erroneous conclusion about the causal operation of abilities in the real world of work.

The effect of measurement error on heritabilities is especially large, and failure to correct for measurement error can substantially distort conclusions (Schmidt & Hunter, 1996, Scenario 8). Suppose, for example, that a researcher, using a standard measure of three-dimensional mental rotation (3DR), reports an observed heritability of 0.60 and concludes that 60% of the variance of this ability is due to genetic effects and 40% to environmental effects. This conclusion is erroneous. The 0.60 is not the heritability of the trait or construct of 3DR, it is the heritability of scores on that particular *measure* of 3DR. There are many other possible such measures of 3DR, each with its own reliability; each such measure will produce a different heritability estimate. Obviously, this is a recipe for a

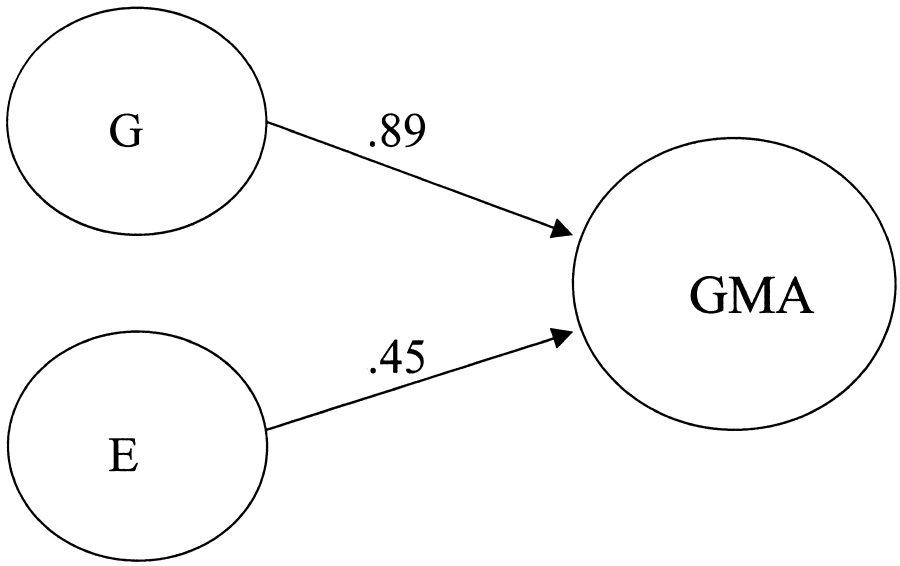


**Figure 2.** Causal impact of genes (G) and environment (E) on general mental ability (GMA) when $h^2 = 0.80$ and $e^2 = 0.20$.

confused research literature. Because the 0.60 is the heritability of the measure, the 40% of variance attributed to "environmental effects" includes measurement error variance—hardly what is meant theoretically by environmental effects on ability!

Suppose in our example the reliability of the 3DR scale is 0.70. Then, the unbiased estimate of the heritability of this trait is 0.60/0.70 = 0.86. Hence, the heritability of the trait is 43% larger than the heritability of the measure. Alternatively, failure to correct the observed heritability for measurement error results in a 30% underestimation of heritability. It also results in a 285% overestimation of the variance due to environmental effects (0.40/0.14 = 2.85)! These are not minor errors.

Notice that in correcting the observed heritability for measurement error, we divided by $r_{xx}$, not by $(r_{xx})^{1/2}$. 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. In this sense, the measurement error in the tests creates a larger downward bias in heritability estimates than in correlations.

Actually, heritabilities are less informative than their square roots. The square root of the corrected heritability is the correlation between genotype and the trait. In our example, $(0.86)^{1/2} = 0.92$. Hence, genetic differences between individuals correlate 0.92 with true scores (i.e., with the *trait* 3DR). If genetic and environmental effects are uncorrelated, this 0.92 is the standardized path coefficient from genotype (G) to the trait 3DR, as shown in Fig. 1. In addition, the square root of the environmental variance proportion is the

standardized path coefficient from environment (E) to 3DR. These path coefficients are much more scientifically informative and the percent variance figures, because path coefficients reveal the actual casual leverage of each independent variable. For example, we can see that each 1 SD increase in G produces a 0.92 SD increase in 3DR, while each SD increase in E produces a 0.37 SD increase in 3DR. Hence, it is apparent that the causal impact of G is 2.59 times greater than that of E (0.92/0.37 = 2.59). By contrast, the variance interpretation gives the false impression that G is over six times as important as E (0.86/0.14 = 6.14).

Consider this same question with respect to GMA. Based on studies comparing identical and fraternal twins, the trait level broad heritability of GMA (i.e., the broad heritability corrected for measurement error) is approximately 0.80. Hence, 80% of the variance in the actual trait of GMA is due to genetic differences between individuals and 20% is due to differences in environments. This presentation creates the strong and definite impression that genes are five times as important in the determination of general intelligence as environmental differences (0.80/0.20 = 5). But, the fact that genetic differences account for five times as much *variance* as environmental differences does *not* mean genes are five times as important in the determination of GMA. As shown in Fig. 2, the ratio of their path coefficients, which index the true causal leverage each has on GMA, is 0.89/0.45 = 1.98 ≅ 2. Hence, genes are only *twice* as important causally as environments—not five times as important.

## Objections to Correcting for Measurement Error

Objections to the corrections that eliminate the biases induced by measurement error are difficult to address because no statement of them can be found in the scientific literature. They can be found neither in the methodological literature (which in fact calls for such corrections) nor the substantive research literature. In this sense, these objections are part of the "underground" or informal culture of psychological research. They are the kind of thing one hears orally and informally but never sees in written form. Hence, it is not surprising that some of these objections seem to be emotional and unthinking rather than rationally based. An example is the contention that disattenuation corrections produce "hydrolytic correlations"—correlations that have been artificially "jacked up." This statement is a rejection of the basic measurement model as explicated in Eqs. (1)–(3). Another objection is that corrected correlations are "hypothetical correlations," that they are "not real data." It would seem apparent that data with biases removed are *at least* as real as biased data. Again, this objection reflects either rejection of basic measurement principles, or ignorance of them. These objections also reflect an inability to think theoretically; that is, to think in terms of scientific constructs, the underlying and simplifying abstractions that are the essence of scientific theories. If one has no concept of traits or constructs as theoretical dimensions that underlie data, then one can make no distinction between observed scores and construct scores (true scores). The failure to make this distinction leads to a sort of reification of initially observed data and a failure to understand the ways in which data are deceptive when accepted naively at face value (Schmidt, 1992, p. 1179).

However, certain other objections are not nearly so mindless. One such objection is that measurement error corrections should not be used because they can result in

correlations greater than 1.00. And in fact, $\hat{r}_{x_t y_t} > 1.00$ do sometimes appear, for two reasons: sampling error and use of an inappropriate reliability coefficient.

Consider sampling error. The discussion in this paper has focused on measurement error; in our examples, we have minimized sampling errors by assuming large sample sizes (e.g., $N = 3000$). But, in most real data, there is substantial sampling error as well as measurement error. Researchers sometimes find it difficult to think about sampling error and measurement error simultaneously, but it is necessary to do so. [One of the strengths of meta-analysis is that it corrects simultaneously for sampling error and measurement error (Hunter & Schmidt, 1990a; Schmidt, 1992). This is also one of the things that make it challenging for students to learn.]

Suppose that two constructs correlate 1.00 at the true score level (i.e., $\hat{r}_{x_t y_t} = 1.00$). That is, suppose they really are the same construct under two different labels. For example, suppose that "job involvement" is really just job satisfaction under a different name. Then the expected (i.e., average) corrected correlation will be 1.00. But, due to sampling error, about half the corrected values will be less than 1.00. By the same token, half will be larger than 1.00. This is not cause for upset and alarm; it is simple sampling error, and is expected and predicted (Schmidt & Hunter, 1996, Scenario 25). In the introduction to this paper, we made a distinction between Eq. (2) and Eq. (3), the difference being that Eq. (3) allows for sampling error. We noted that 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. Hence, the basic measurement model anticipates this objection.

When sample sizes are small (the usual case in most research), sampling errors are much larger than most researchers realize. Observed correlations can even be negative when their population values are positive (and vice versa). Examples illustrating this point can be found in Hunter and Schmidt (1990a, Chaps. 1 and 2), Schmidt (1992), and Schmidt, Ocasio, Hillery, and Hunter (1985). At every sample size, researchers underestimate the size of sampling errors.

If the actual true-score correlation is a large value but less than 1.00 (say, 0.85), sampling error will cause some of the corrected values to be above 1.00, but the percentage will be less than 50%. Since the correlation cannot by definition exceed 1.00 (in absolute value), such estimates are simply set to 1.00. A similar thing occurs in estimating the size of components of variance in analysis of variance (ANOVA) when variances are estimated by subtraction. Sampling error sometimes causes these estimates to be negative, and a variance cannot be negative. By accepted convention, such estimates are just set to zero (Hunter & Schmidt, 1990a, pp. 412–414).

What would be an example of $\hat{r}_{x_t y_t} > 1.00$ because of use of the wrong reliability coefficient? Suppose split half reliabilities are computed. For example, for each scale the odd–even split half correlation is computed. This correlation is the reliability of a test half as long as the test in question, and so this correlation must be corrected upward using the Spearman–Brown formula. But, suppose the researcher neglects to do this. The reliability estimates are then too small, resulting in overcorrection when Eq. (3) is applied. If the underlying $r_{x_t y_t}$ is fairly large, this error can result in $\hat{r}_{x_t y_t} > 1.00$. Another example of this error is given in Schmidt and Hunter (1996, Scenario 11).

The above is an example of an erroneous application of the correction for measurement error. Errors of this sort are sometimes referred to as "abuses" of such corrections. Some argue that corrections for measurement error should not be made because they are "subject to abuse" or "subject to misuse"; i.e., because errors can be made in applying such corrections. (Typically, only errors of application that result in overestimates of construct level correlations are referred to as "abuses" or "misuses." Errors of application that result in *underestimates* of construct level correlations seem to trigger no alarms.)

What about this objection? Over the years, this objection has been brought against virtually every important methodology used in psychological research. It has been argued that factor analysis has been "misused" and therefore should not be used in research. Some have said that many users make errors in conducting meta-analyses, and so meta-analysis should not be used. More recently, some have maintained that the literature contains many defective applications of SEM, and that therefore the journals should not publish SEM-based studies. In fact, there is no method that has not been misused in some study. Thus, this line of reasoning leads to the conclusion that no methods of data analysis and interpretation can be used in psychological research! Obviously, any research tool can be misused. But, that is hardly a legitimate reason for banning its use!

Another point is that this objection actually advocates a misuse (or abuse) of measurement error corrections. Those who endorse this objection conclude that measurement error corrections should not be made. But, this is in itself a misuse, since omission of such corrections results in biased estimates of scientific parameters. Surely any use or omission that results in major biases in data and conclusions is a misuse.

Another objection holds that corrections for measurement error should not be made because they *reduce the precision* of the correlation estimates. That is, the corrections increase the amount of sampling error and increase the width of the confidence intervals. Wider confidence intervals mean greater uncertainty and this is undesirable. What about this objection?

It is certainly true that corrections for unreliability increase sampling error. There is more sampling error in $\hat{r}_{x_t y_t}$ than in $\hat{r}_{xy}$. It is also true that the corrections increase the width of the confidence intervals. The confidence interval around $\hat{r}_{x_t y_t}$ is wider than the confidence interval around $\hat{r}_{xy}$. In fact, it is wider by precisely the factor of correction. For example, if the correction increases the estimated correlation by 30%, the confidence interval width will also be increased by 30%. So, is this objection correct? No, because it ignores systematic error. When correction is made for measurement error, nonsystematic error—sampling error—is increased; but it is also true that systematic error is eliminated. The systematic error is the downward bias caused by measurement error. This systematic error cannot be eliminated without corrections for measurement error. In science, systematic error is more important than nonsystematic error—because nonsystematic (random) error can be averaged out. In fact, this is precisely what meta-analysis does: After correcting for systematic error, it averages across corrected rs ($\hat{r}_{x_t y_t}$) to average out random sampling error (Hunter & Schmidt, 1990a; Schmidt, 1992). Systematic error, on the other hand, cannot be averaged out. It can be removed only through the appropriate systematic correction.

A final objection goes something like this: "Instead of relying on corrections for measurement error, we should be putting more effort into developing reliable measures." The assumption underlying this objection is that the use of reliable measures eliminates the need for corrections. This assumption is a variation on the "magic number belief" (Schmidt & Hunter, 1996, Scenario 2). This belief holds that if reliability is above some magic number—usually 0.70—then it is "adequate," and so there is no need to correct for measurement error. However, the bias introduced into estimates of correlations between constructs does not magically disappear when reliability reaches a particular level. In fact, when reliability is 0.70, the bias factor is $(0.70)^{1/2} = 0.84$; that is, the observed correlation will average 16% below its actual value. And that is the effect of measurement error in only *one* of the two measures. If both measures have reliabilities of 0.70, the bias factor is $[0.70(0.70)]^{1/2} = 0.70$. That is, the observed correlation is on average 30% below its actual value, a large bias. (One example of this occurs in the estimation of heritabilities, discussed earlier.)

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 always a need to eliminate the downward biases induced by measurement error.

Another response to this objection is this: What about all the studies in the literature that fell below your magic acceptable reliability level? Should they just be discarded? If we eliminate the bias in these correlations by correcting for measurement error, we can include these studies in meta-analyses, making use of the information they contain. Hence from this point of view also the corrections are critical to the development of cumulative knowledge (Hunter & Schmidt, 1990a).

## The Substantive Meaning of Measurement Error

In psychological research, all sorts of entirely different things are referred to by the ambiguous term "error variance"; things as different as individual differences variance, sampling variance, coding errors, and many other things, with no substantive definition provided of the type of "error" being referred to. The same thing often occurs in treatments of measurement error: measurement error is not defined substantively. That is, the actual psychological processes that produce measurement error are not described. As a result, one is left with the impression that measurement error springs from hidden and unknown sources and its nature is mysterious. It is our belief that this impression contributes to the objections discussed above to making corrections for measurement error. That is, researchers who have no idea what measurement error substantively is have difficulty accepting that they should correct for it. In fact, they may think of measurement error as hypothetical; that is, they may feel that the existence of measurement error is not an incontrovertible fact but merely an hypothesis—an hypothesis that may or may not be correct.

The processes that produce measurement error are not mysterious. Measurement error is produced by real psychological processes that can be studied and understood (Feldt & Brennan, 1989). There are three major substantive error processes that must be considered in *all* psychological measurement: random response error, transient error, and specific factor error. In addition, for some kinds of measurement, error due to disagreement between scorer must also be considered.

### Random Response Errors of Measurement

Random response error occurs within occasions; it is caused by variations in attention, mental efficiency, momentary distractions, etc. Research in cognitive psychology and human information processing has shown that the human central nervous system contains considerable noise at any given moment. This "neural noise" can, for example, cause a person to answer two semantically identical questions differently—because of misreading a single word, because of a stray worry that popped up, etc.

Random response error is controlled (reduced) by averaging (or summing) across items within occasions. Other things equal, the larger the number of items, the greater the extent to which random response error is averaged out in the final score (which, after all, is essentially, the average across all the items).

Random response error has been found to be important in all areas of measurement studied, and it is important in the measurement of human abilities.

### Transient Errors of Measurement

While random response error occurs across moments on a single occasion, transient error occurs across occasions. Transient error is produced by a cause that affects all measurements taken at the same time but which varies randomly from one time to the next. Transient errors are caused by variations in mood, feeling, mental efficiency, or general mental state across occasions. For example, any given day is characterized for each person by a certain mood, level of emotion, level of mental clarity, etc. These psychological factors are transient. Someone having an anxious day or a mentally sluggish day today will probably not be in that mental state tomorrow or the next day. Again, these processes are not mysterious; they are real psychological processes that we are all familiar with. Most everyone has experienced a mentally sluggish day, an irritable day, a depressed day, an elated day, or a hangover. Transient error processes, unlike random response error processes, do not vary across moments within occasions. They vary across occasions.

The amount of transient error is assessed by correlating performance or responses across occasions. It is not possible to estimate or control for transient error when a measure is administered on only one occasion.

Research suggests that transient error may not be a major problem in the measurement of human abilities (Schmidt & Hunter, 1996, Scenarios 14 and 17). However, this research base is sketchy and more and better data are needed. We currently have such research underway.

### Specific Errors of Measurement

Specific error arises from the subject's idiosyncratic response to some aspect of the measurement situation. For example, on a questionnaire item, different subjects may give different meanings to the same word. In an assessment of fear reactions toward animals, a subject may have had some odd experience that causes him to be less afraid of snakes than would be expected on the basis of his general fear reactions toward animals. Or, when job tasks are sampled to set up work stations in a work sample measure, a worker may by chance get the tasks that he is particularly poor at performing. If different measurements taken vary these irrelevant aspects of the situation, then the errors due to these random

causal factors will be independent of one another. That is, the specific errors will then be independently sampled across measurements and will tend to cancel each other out. However, if the same measurement scale is administered at another time, the idiosyncratic responses will not change and the specific error of measurement will be repeated. That is, for specific error, the random elements of the response are independently sampled across situations (e.g., items or problems) but not across time. Thus specific errors tend to cancel each other out across different items, problems, or questions; but for any one item, problem, or question, they replicate themselves across occasions.

These specific factors are not random response error because they can be stable across time. But, they are not part of the trait or construct being measured—because they do not correlate at all with any other item measuring that trait. So, they function as measurement error, and must be so treated.

It may appear that specific factors are not psychological in the sense that random response error and transient error are. That is, it may appear that specific factors are properties of items (or scales) per se and are not based on psychological processes. However, specific factor error is produced by the interaction of people with items (or scales), and these interactions are psychological processes. Consider the vocabulary item "capon." A capon is a castrated rooster. The knowledge that capon refers to poultry may be indicative of one's general vocabulary, while the knowledge that a capon is a castrated rooster may represent a specific error reflecting the extent to which the respondent has an agricultural background. This specific error is irrelevant to verbal ability but it does reflect a real psychological or experiential process.

Within a test or measuring instrument, specific error is controlled by averaging (or summing) across items; this process averages the influence of specific item errors out of the total score. Specific factors can be averaged out because they correlate zero with each other.

Just as an individual item within a measure may have a specific error factor, so different scales may contain specific error factors. For example, most major verbal ability tests to at least some extent measure specific factors unique to that scale. From a broader theoretical perspective, we may consider the trait of verbal ability to be defined by what is measured in common by different verbal ability tests. Therefore, the specific factor in each test is specific measurement error. Across such verbal ability scales, then, specific error of measurement is controlled by averaging (or summing) across scales. This point is developed further below.

Specific factor error of measurement is important in the measurement of human abilities and in the personality domain. However, additional research is needed to more accurately calibrate the size of the specific error variance component in these and other domains.

### Measurement Error Due to Scorer Disagreement

Scorer error is disagreement between scorers or scorings of the same instrument. For example, two different trained evaluators scoring the same set of essay examinations might correlate 0.50. Two judges scoring a projective test (for example, the Thematic Apperception Test; TAT) for achievement motivation may correlate 0.70 with each other. Two different judges observing and rating people as they perform a hands-on work sample test might correlate 0.85. Scorer error can be substantial when scoring involves subjective judgment, as in the case of essay examinations or the TAT.

Scorer errors of measurement in the above examples are generated by psychological processes. In the case of essay examinations and the TAT, the task of the scorers is to apply complex (and perhaps somewhat ambiguous) rules to examinee responses to arrive at final scores. Research on human information processing indicates that tasks of this sort are cognitively difficult for people, resulting in considerable disagreement between individuals. (In fact, there is considerable disagreement within the same individual when the same person performs the same task on two different occasions.) So, once again, measurement errors are produced by real psychological processes—processes that can be identified, studied, and understood.

A key point about scorer error is that controlling for it does *not* control for random response error, transient error, or specific factor error in the responses of the examinees. For example, consider the scorer agreement reliability of 0.70 for the TAT, above. Thirty percent of the variance is due to scorer error alone. But, the remaining 70% of the variance is not true score variance. Some—perhaps much—of the remaining 70% of the variance is due to random response error on the part of the *examinees*, transient error on the part of the *examinees*, and specific factor error (specific to that particular form of the TAT). To control for these other kinds of measurement error, in addition to scorer error, one must give *parallel forms* of the TAT to the same people *on different days*. The correlation between different scorers of parallel forms administered on different days detects and assesses all four types of measurement error (Schmidt & Hunter, 1996, Scenario 12). This correlation might be very small—0.30 or less, illustrating the fact that scorer agreement reliability alone greatly overestimates actual reliability. Correcting for measurement error using scorer agreement reliabilities almost always results in gross undercorrections and therefore biased estimates of trait or construct level correlations.

Because objectively scored tests are usually used, scorer errors of measurement are often not important in the measurement of human abilities. However, scorer measurement error is important in the measurement of writing ability and other constructs using essay tests.

## Calibrating Error Processes in Reliability Coefficients

For purposes of the following discussion, we will assume objective and accurate scoring procedures, and therefore will ignore scorer reliability. Different reliability coefficients calibrate or assess the magnitude of different measurement error processes. The extent or magnitude of an error process is assessed by a reliability estimate only when that error process reduces the size of the reliability estimate; the amount of this reduction is the measure of the size of the impact of the error process. Only one type of reliability coefficient, the coefficient of equivalence and stability (CES; Cronbach, 1947), assesses the extent of all three types of measurement error. The CES is estimated by correlating two parallel forms of the measure administered on two different occasions. The use of two parallel forms assesses scale—specific measurement error factors, and the use of two occasions assesses for transient error. The assessment of transient error automatically assesses the extent of random response error (as does use of multiple items). Because of this, the CES is the ideal reliability estimate. Its magnitude is appropriately reduced by all three sources of measurement error. The quantity of one minus the CES is the proportion of total variance that is due to all three measurement error sources together. The CES is the

optimal estimate of reliability to use in making corrections for biases induced by measurement error. Based on the logic of measurement theory, the other types of reliability coefficients discussed below correct only partially for measurement error.

If the same form of a measure is correlated across two different occasions, the result is an estimate of the coefficient of stability (CS; Cronbach, 1947). The CS assesses the extent of random response error and transient error but does not assess specific factor error. Hence, use of the CS in corrections for measurement error leads to undercorrection; that is, some but not all of the downward bias induced by measurement error is eliminated.

There are many ways in which the reliability of a test can be estimated from administration at one point in time. In the abilities domain, the most common method used is the familiar Kuder–Richardson-20 formula (KR-20). If the items are scored on a continuum rather than dichotomously, this becomes Cronbach's Alpha (Cronbach, 1951). Another method is odd–even split half corrected using the Spearman–Brown formula. Cronbach (1947) refers to all such estimates as coefficients of equivalence (CE), because they estimate the correlation between that scale and a parallel form of that scale administered at the same time.

The CE assesses random response error and specific factor error, but not transient error. Since the scale is administered on only one occasion, there is no possibility of assessing the extent of transient error. Hence, use of CE in correcting for measurement error leads to incomplete corrections (undercorrection) if the measure in question is affected by transient errors.

At this point, we need to add a note about specific factor errors of measurement. In the above discussion, specific factor error refers to factors specific *to parallel forms* of the measure, as parallel forms are defined in classical measurement theory. However, as we noted earlier in discussing the construct of verbal ability, it is often scientifically desirable in developing general explanatory theory to broaden the concept of specific factor error beyond the limits of parallel forms as defined in classical measurement theory. For example, the literature contains numerous measures of verbal ability that are not parallel forms of each other as defined in classical measurement theory (Lord & Novick, 1968). These measures were constructed by different researchers at different times, with no attempt having been made to make them parallel in the classical sense. We may want to define the theoretical construct of verbal ability as the factor that all such tests have in common. Suppose we use five such scales simultaneously to measure verbal ability, with the total or average score across these scales being the final observed score. (For completeness, assume also each scale is administered on a different occasion, thus capturing the effects of any transient error.)

What is the appropriate reliability of this scale? Each scale should be treated as an item in a five item test, and reliability should be computed using Cronbach's Alpha. This ensures that factors specific to each scale are assigned to measurement error. The resulting reliability is the generalizability coefficient (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). This coefficient will usually be slightly smaller than would have been the case if all five scales have been constructed to be classical parallel forms, but the resulting research findings will be widely generalizable. These findings will generalize to the population of all such non-parallel measures of verbal ability. The theory constructed in this manner will hence have more scientific generality and in that important sense will be a better theory. At this point, we have stepped beyond the bounds of classical measurement theory into

generalizability theory. In situations of this sort—which arise with some frequency—generalizability theory makes contributions beyond what is strictly possible within classical measurement theory.

## The Role of Substantive Knowledge in Measurement Error Corrections

In the discussion above, we made a point of the fact that the psychological processes producing measurement errors are not mysterious and unknowable but can be examined and understood. A related fact is that knowledge of how these processes function differently in different research domains can be obtained through research, and this knowledge can be applied in making the needed corrections. For example, in the domains of ability and aptitudes, research on measurement error suggests that transient error, at least for adults, is small—perhaps small enough to be ignored in many cases. This is shown by the fact that CES is typically only slightly smaller than the coefficient of equivalence (CE). This means that estimates of the CE, such as KR-20, can be used in lieu of CES estimates to correct for measurement error with only marginal loss of accuracy.

Research on measurement error also suggests that specific factors are not large in the abilities domain. This is seen in the fact that test–retest reliabilities (CS estimates) are often similar to CES estimates computed over the same time interval (Schmidt & Hunter, 1996, Scenario 13). Specific factors as defined by generalizability theory are somewhat larger, however. Specific factor error in the abilities domain may be larger than transient error, and it is advisable to assess and quantify specific factor measurement error whenever possible.

In ratings of job performance and other performances, factors specific to particular raters account for approximately 30% of the variance of the ratings. This large idiosyncratic specific factor can be removed only by averaging ratings across several raters (Rothstein, 1990; Schmidt & Hunter, 1996, Scenario 10; Viswesvaran, Ones, & Schmidt, 1996). On the other hand, transient error appears to be quite small in this area.

Different substantive research domains, with their different kinds of measuring scales, are plagued in differing degrees by the different error-generation processes. All of these research areas should devote more effort to calibrating and understanding the size and nature of these error processes in their respective domains. In any substantive research area, understanding the error processes that operate in measurement is part of understanding the actual subject matter. Understanding error processes contributes to be understanding of the substantive phenomenon being investigated.

In this paper, we have presented the basic principles and general procedures for correcting the biases induced in data by measurement error. However, in specific research situations, additional questions of a more detailed nature typically arise. In fact, each study is unique, and it is often not clear to researchers how the general principles presented in this paper should be applied in a particular case. To address this need, we (Schmidt & Hunter, 1996) have examined in some detail a series of 26 representative, concrete research scenarios that we have encountered in our work as researchers, advisors to researchers, or reviewers of research. These real world ''case studies'' provide additional specific and detailed guidance in making corrections for measurement error. Among the 26 scenarios most researchers will be able to find at least one that is similar or identical to their situation.

## Conclusion

The accurate and meaningful testing of hypotheses and theories is not possible without corrections for the distortions induced into data by unavoidable measurement errors. Failure to make such corrections retards the development of cumulative knowledge in research. Although there are objections to such corrections, these objections are not found in either the methodological or substantive research literatures and are more in the nature of emotionally based folk beliefs about research that are part of the informal or underground culture. All such objections can be shown to be unfounded. Although many substantive researchers would apparently and understandably rather not be bothered by the methodological complexities involved in removing biases created by measurement error, there is no way around the need to address this problem. Failure to do so has far too often in the past retarded the advance of cumulative research. It is time to remove these unnecessary hobbles from psychological research.

## References

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 edn. (pp. 105–146). New York: Macmillan.

Fuller, W. A. (1987). *Measurement error models*. New York, NY: Wiley.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science: The empirical effectiveness of research. *American Psychologist, 42*, 443–455.

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

Hunter, J. E., & Schmidt, F. L. (1990b). Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology, 75*, 175–184.

Linn, R. L., Harnisch, D. L., & Dunbar, S. B. (1981). Correction for range restriction: An empirical investigation of conditions resulting in conservative correction. *Journal of Applied Psychology, 66*, 655–663.

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

Ree, M. J., Carretta, T. R., Earles, J. A., & Albert, W. (1994). Sign changes when correcting for range restriction: A note on Pearson's and Lawley's formulas. *Journal of Applied Pschology, 79*, 298–301.

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

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.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*, 557–560.