AFRL-HE-WP-JA-2006-0001

AIR FORCE RESEARCH LABORATORY

The Role of Measurement Error in Familiar Statistics

Malcolm James Ree

Our Lady of the Lake University
San Antonio TX 78207-4689

Thomas R. Carretta

Human Effectiveness Directorate
Warfighter Interface Division
Wright-Patterson AFB OH 45433

January 2006

Human Effectiveness Directorate
Warfighter Interface Division
Wright-Patterson AFB OH 45433

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
January 2006	Journal Article	

4. TITLE AND SUBTITLE
The Role of Measurement Error in Familiar Statistics

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)
*Malcolm James Ree, **Thomas R. Carretta

5d. PROJECT NUMBER
7184

5e. TASK NUMBER
09

5f. WORK UNIT NUMBER
72

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) AND ADDRESS(ES)
*Our Lady of the Lake University
San Antonio TX 78207-4689

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
**Air Force Materiel Command
Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Interface Division
Wright-Patterson AFB OH 45433-7022

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/HECI

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-HE-WP-JA-2006-0001

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
This is a journal article published in Organization Research Methods. The clearance number is AFRL/WS-05-2072 and was cleared 8 September 2005.

14. ABSTRACT
Measurement error, or reliability, affects many common applications in statistics, such as correlation, partial correlation, analysis of variance, regression, factor analysis, and others. Despite its importance, the role of measurement error in these familiar statistical applications often receives little or no attention in textbooks and courses on statistics. The purpose of this article is to examine the role of reliability in familiar statistics and to show how ignoring the consequences of (less than perfect) reliability in common statistical techniques can lead to false conclusions and erroneous interpretation.

15. SUBJECT TERMS
Reliability, Statistics, Inference, Interpretation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
Unclassified					Thomas R. Carretta
a. REPORT	b. ABSTRACT	c. THIS PAGE	SAR	16	19b. TELEPHONE NUMBER *(include area code)*
UNC	UNC	UNC			(937) 656-7014

Organizational Research Methods
Volume 9 Number 1
January 2006 99-112

10.1177/1094428105283192
http://orm.sagepub.com
hosted at
http://online.sagepub.com

The Role of Measurement Error in Familiar Statistics

Malcolm James Ree
Our Lady of the Lake University
Thomas R. Carretta
Air Force Research Laboratory

Measurement error, or reliability, affects many common applications in statistics, such as correlation, partial correlation, analysis of variance, regression, factor analysis, and others. Despite its importance, the role of measurement error in these familiar statistical applications often receives little or no attention in textbooks and courses on statistics. The purpose of this article is to examine the role of reliability in familiar statistics and to show how ignoring the consequences of (less than perfect) reliability in common statistical techniques can lead to false conclusions and erroneous interpretation.

Keywords: *reliability; statistics; inference; interpretation*

Measurement error is an integral part of measurement and is frequently indexed by reliability. The reliability of a measure is the ratio of true variability to total variability. In simple nontechnical language, reliability means precision. Most educational researchers and psychologists learned about reliability in courses in psychometrics. Statistical techniques such as descriptive statistics, regression, or analysis of variance are taught in separate courses. In some instances, the two are combined, such as when Spearman's true correlation is introduced; this, however, is frequently the only time. More recently, courses covering meta-analytic techniques frequently bridge that gap. The purpose of this article is to show the role of reliability in familiar statistics and to show how ignoring the consequences of (less than perfect) reliability in common statistical techniques can lead to false conclusions and erroneous interpretation. Because of their widespread use in applied research and application, we will illustrate the role of reliability with examples from descriptive statistics, *z* tests and *t* tests, correlation, partial correlation, linear regression, test bias analysis, factor analysis, analysis of variance, and analysis of covariance.

Authors' Note: The opinions expressed are those of the authors and are not necessarily those of the U.S. government, the Department of Defense, or the U.S. Air Force. Please address correspondence to Malcolm James Ree, Our Lady of the Lake University, 411 SW 24th Street, San Antonio, TX 78207-4689; e-mail: mree@satx.rr.com.

Measurement Error Model

The true score model is the most frequently used measurement error model (see Fuller, 1987). In this model, the basic equation states that the observed score is equal to the true score plus an error score. Furthermore, the error score is assumed to be random and therefore independent of the true score. The true score and error score are not correlated. If the error score is not random, the result is called bias and the consequences may be situationally specific. There are measurement models for nonrandom error, but the current article is limited to random error. The basic true score equation is O = t + e, or Observed = True + Error.

This yields[1]

$$\sigma^2_{observed} + \sigma^2_{true} + \sigma^2_{error}. \quad (1)$$

By definition, reliability (Stanley, 1971) is the ratio of true score variance to observed score variance or reliability = $r_{XX'} = \sigma^2_{true}/\sigma^2_{obs}$. This is equivalent to $r_{XX'} = 1 - (\sigma^2_{error}/\sigma^2_{obs})$, or reliability equals 1 minus the proportion of error variance to observed variance. The effects of measurement error on familiar statistical techniques can be determined from these equations. The purpose of the current effort is to explain the consequences of the use of less than perfectly reliable measures in statistical analyses.

The Variable Is Not Reliable

It is not unusual to hear people say, "That test is reliable" or "That is a reliable measure" of some construct. However, Thompson (2003) has forcefully made the point that a test (or other measured variable) is neither reliable nor unreliable. Reliability concerns the scores of the measure and is a consequence of the sample at hand. "It is important to evaluate score reliability in *all* studies, because it is the reliability of the data in hand that will drive study results, and not the reliability of the scores described in the test manual" (Thompson, 2003, p. 5). Your sample will almost surely differ from the normative sample reported in the test manual. It may differ in composition by gender, ethnicity, age, experience, education, testing circumstances, or many other variables. These differences will cause the reliability of your sample to be different from the reliability reported in the manual, and your results will be driven by the reliability of your sample.

Descriptive Statistics

Mean

The effect of unreliability on the mean is benign. Because the error score is random, the mean of the error score is expected to be zero. Therefore, the expectation of the observed mean equals the true mean. The bias caused by measurement error on the observed mean is nil.

Variance and Standard Deviation

The effects of measurement error on the variance and standard deviation are not so agreeable. Returning to the true score equation, we see that the observed variance is the sum of the true and error variances ($\sigma^2_{observed} = \sigma^2_{true} + \sigma^2_{error}$). Consequently, the observed variance is greater when error increases. Observed score variance is always greater than true score variance when the variable has been measured with less than perfect reliability. If the true variance (σ^2_{true}) is 100 and the error variance (σ^2_{error}) is 10, the observed variance (S^2_{obs}) will be 110. If the true variance remains 100 and the error variance is 20, the observed variance will be 120. Note that the effect on the standard deviations will appear to be less at 10.49 ($\sqrt{110}$) and 10.95 ($\sqrt{120}$). In these cases, the reliability of the two scores would be .91 ($\sigma^2_{true}/\sigma^2_{obs} = 100/110$) and .83 ($\sigma^2_{true}/\sigma^2_{obs} = 100/120$), respectively. The biasing influence on effect size will be discussed in a subsequent section.

The *z* Test and the *t* Test

The basic form of the *z* test and the *t* test is a sample statistic minus a population parameter in the numerator, divided by a standard error in the denominator. In the case of the *z* or *t* test of a mean, the benign effect on the mean precludes changes in the numerator. The effect of reliability is found in the denominator. The standard error (or estimated standard error in the *t* test) is the standard deviation divided by the square root of *n*, the sample size.

Consider the two-sample (or independent-samples) two-tailed *z* test using a .05 Type I error rate. With a difference between the means of 3.6, a sample size of 30, and a true standard deviation (i.e., measured without error, $r_{xx'} = 1.0$) of 10, the computed *z* value would be significant at 1.972. If the standard deviation were increased to 11 by unreliability ($r_{xx'} = .83$), the *z* test statistic would not be significant with a value of 1.793. If the reliability were reduced further, the *z* value also would be reduced. For example, with the same sample size and mean difference, but reliability reduced to $r_{xx'} = .625$ (observed standard deviation = 16), the *z* value is 1.232 and would not be significant.

Confidence Intervals

Another way to evaluate the effects of unreliability is to look for differences in the width of confidence intervals. Addition of error variance to true variance causes the confidence intervals to increase. With a sample size of 30 and a true standard deviation of 10, when the reliability is 1.0, the true standard error is 1.826. If the reliability were reduced to .830 or to .625, the standard errors become 2.008 and 2.921, respectively. The confidence interval becomes wider.

Effect Size

Less than perfect reliability also will have an influence on effect size ($(\mu_1 - \mu_2)/\sigma_e$) (see Baugh, 2002, for an insightful discussion of the issue). Russell and Peterson (2002) reported the effect size for African American means versus White means on a series of tests in a research project. They discuss a spatial test called Reasoning, which had an effect size of .77. Russell and Peterson noted that many tests show an African American versus White effect

size of 1.0, and their experimental tests were on average less than 1.0. The Reasoning test had a test-retest reliability of .65, and correcting the effect size for this unreliability, the true effect size becomes .96, very close to the size reported frequently for such differences. This change in effect size occurred because of the change in the estimate of the standard deviation when unreliability was accounted for. A different conclusion about the Reasoning test would have been reached had unreliability been taken into account.

Clearly, unreliability causes a reduction in statistical power, an artifactual increase in confidence intervals, and a bias in estimating effect size. Ignoring the effects of unreliability will lead to inappropriate conclusions and inferences about the tests and the constructs being studied.

Correlation

With the increased popularity of the meta-analytic technique of validity generalization (Hunter & Schmidt, 1990, 2004), the correction for attenuation has become well known again, at least by industrial/organizational psychologists (see Ree & Earles, 1993). Spearman (1904) demonstrated that the correlation between the observed scores of two variables was a function of the reliability of the two variables. The well-known formula that expresses this is

$$r_{xy} = r_c(\sqrt{r_{XX'}}\sqrt{r_{YY'}}), \tag{2}$$

where r_c is the estimate of the true correlation (sometimes written $r_{\tau x \tau y}$, where τ indicates true score), r_{xy} is the observed attenuated correlation, and $r_{xx'}$ and $r_{yy'}$ are the reliabilities of X and Y, respectively.

For example, if two measures of the same construct (true score correlation of 1.0) each have a reliability of .8, the maximum correlation between the two (r_{XY}) is .8. If one of the measures has a reliability of .6 and the other .8, the maximum observed[2] correlation would be .69. Ignoring the consequences of reliability of the measures, the conclusion would be that there is a moderate to strong correlation rather than the perfect correlation obtained at the true score level. A practical consequence of this might be the search for new predictors to close the (specious) gap between .69 and 1.0.

Observed correlations can be corrected for the unreliability of the variables by using an algebraic manipulation of Equation 2 to yield

$$r_C = r_{XY} / (\sqrt{r_{XX'}}\sqrt{r_{YY'}}). \tag{3}$$

Consider an observed correlation of .72, where both variables X and Y have reliabilities of .8. Using the equation above for correcting the correlation, the true correlation between X and Y is .9. That is, $r_c = .72/(\sqrt{.8} \times \sqrt{.8}) = .9$.

Correlations between variables that change from low or moderate to moderate or high after correction for (less than perfect) reliability suggest that the variables' utility could be improved if they were made more reliable. In addition, low to moderate correlations that do not increase in magnitude after correction for unreliability suggest that the variables contain other sources of variance.

Partial Correlation

Partial correlation is the correlation between two variables, *X* and *Y*, while holding a third variable, *Z*, constant. Whether used for control or for selecting variables for stepwise regression, the role of reliability in partial correlation can be large. Consider the following example with three variables, *X*, *Y*, and *Z*, which measure the same construct with perfect reliability. The true correlation between *X* and *Y*, *X* and *Z*, and *Y* and *Z* would be 1.0. The partial correlation between any pair holding the other constant (i.e., partialing it out) would be .0. If the reliability of all measures were .8, the partial correlation can be given as .44.

$$r_p = \frac{r_{XY}(.8) - r_{XZ}(.8)\, r_{YZ}(.8)}{\sqrt{1-(r_{XZ} \times .8)^2}\sqrt{1-r_{YZ} \times .8)^2}} = .44.$$

Note that the value goes from no partial relationship (.0) to a moderate (.44) partial relationship. This is a big difference and might have substantial implications for theory, application, and policy. Caution is urged in interpretation.

If *Z* were a variable used for control by partialing it out, its reliability would be influential in the estimation. For example, researchers partialed out age (*Z* in this example) to estimate the true correlation between leg length (*X*) and running speed (*Y*). Suppose that age (*Z*), the variable to be partialed out, had a reliability of .4 and the triad of correlations among *X*, *Y*, and *Z* was truly 1. The observed partial correlation between leg length (*X*) and running speed (*Y*) would be .29 rather than .0. The observed correlation of .29 is a poor estimate of the true correlation, and the researcher would make erroneous conclusions about the relationship between the variables.

Linear Regression Coefficients

Simple Linear Regression With One Predictor

Consider a simple linear regression of *Y* on *X*. In the explanation of this regression, many statistics texts contain a single line such as "it is assumed that all predictors are fixed variables measured without error." The role of measurement error in estimation of raw score regression weights, *b*, is given by

$$b = \beta\, r_{XX'}, \tag{4}$$

and for the regression constant (or intercept), we have

$$a = \bar{Y} - (b/r_{XX'})\bar{X}. \tag{5}$$

In the case of a one-predictor regression, the effect is direct and easy to understand. The *b* coefficient is biased toward zero, and the *a* coefficient is inflated. They are biased estimates of the population parameters. Unreliability in the criterion has no biasing effect on the regression coefficients; however, it does attenuate the correlation between the predictor and criterion. There is a simple method to correct these biased estimates. The *b* coefficient is divided

by the reliability of the predictor variable, and this b coefficient is then placed in Equation 5 for the intercept.

Suppose job performance criterion Y is regressed on test X, yielding the regression equation $Y = 2.0 + 1.6X$ and that the reliability of test X is .80. Correcting the b coefficient gives (1.6/.8 = 2), and assuming means of 5 for the X variables and 10 for the Y variables, correcting the intercept gives $10 - 2(5) = 0$. The corrected regression equation is $Y = 0 + 2X$.

Multiple Regression

When there are multiple predictors, the effects on the regression coefficients become complex and difficult to specify simply. The effect of reliability is a function of the reliability magnitudes and the true score correlations among the predictors. Unreliability in the criterion has no biasing effect on the regression coefficients; however, it does attenuate the multiple correlations between the predictors and criterion. Aiken and West (1991) provided an instructive example for the case of two independent variables, X and Z, used to predict the criterion Y. In this case, the standardized regression weight being estimated is the partial regression coefficient of Y on X holding out the effect of Z. The effect of the unreliability of the variable being partialed out has a substantial effect on the partial regression coefficient of the other variable. Even if one independent variable in a regression were measured with perfect reliability, the unreliability of the other independent variables will have a biasing effect on the regression coefficient associated with the independent variable measured without error. The standardized regression coefficient is given by

$$b_{YX.Z} = (r_{YX} - r_{YZ}r_{XZ}/(1 - r_{XZ}). \tag{6}$$

To correct this equation for unreliability of variable Z, it is necessary to write it as

$$cb_{YX.Z} = (r_{YX}r_{ZZ'} - r_{YZ}r_{XZ})/(1 - r_{XZ}). \tag{7}$$

For example, if X is measured without error, the reliability of Z is .64, and $r_{YX} = r_{YZ} = r_{XZ} = .5$, the corrected standardized coefficient is

$$cb_{YX.Z}((.5 \times .64) - (.5 \times .5))/(1 - .5) = .07/.5 = .14.$$

The two-variable case can be extended to the case of many independent variables.

Interpretation of Regression Coefficients

The failure to include reliability in the interpretation of the regression equation causes problems in several ways depending on the use made of the regression equation and its coefficients. The first is in the interpretation of the relative importance of the constructs related to the predictors. Frequently, researchers compare weights and derive meaning of the relative importance of the constructs represented by the observed variables such as verbal or mathematical ability. The uncorrected regression weights are not dependable indicators of the importance of the independent variables; therefore, interpretation of them can lead to erroneous conclusions.

Consider an aptitude test with three equally reliable measures representing reading skill, mathematics knowledge, and space perception. Furthermore, the source of validity is limited

to the common first factor (i.e., *g*) underlying each test in the battery and no validity, in this example, is due to the specific measurement (i.e., *s*) of each test. Under these conditions, each test should have the same regression weight when used in a regression equation to predict the criterion. However, if there are differences in test reliabilities, the regression coefficients will vary differentially from their true population values. Suppose these three example tests have reliabilities of .65, .70, and .85, respectively. In estimation, the three regression coefficients will differ because of their reliability. For example, if the three uncorrected regression coefficients were .195, .210, and .255, some might interpret this to mean that space perception is 1.3 times (.255/.195) as important as reading skill. In reality, the only difference is in the reliability of the measures.

In his computer programs called "Package," John Hunter (personal communication, May 1, 1995) has a regression procedure that allows for explicit correction for unreliability and corrects the regression coefficients.

Even if the reliability of the measures starts out the same, prior selection leads to reduction of reliability in the sample. Prior selection refers to the process of selecting a sample using some method, such as minimum qualification scores, that changes the variability of the scores in that sample. Gulliksen (1950, 1987, p. 124, Equation 5) provides the following equation to show the relationship between prior selection and reliability.

$$R_{xx} = 1-(s_x^2 / S_x^2)(1-r_{XX}). \tag{8}$$

Consider the previous example with the three tests in which the sample has been selected on the basis of scores on the reading skill test, which has caused indirect selection (Thorndike, 1949) to occur on the mathematics and space-perception tests. This indirect selection is the result of the correlation of the variables. Given the same true regression coefficients and reduction in variance of 50%, 30%, and 20%, respectively, for reading skill, mathematics, and space perception, the reliabilities of the tests have changed differentially. The regression coefficients thus become differentially biased and poor estimates of the population values. Some would interpret these coefficients, and clearly, erroneous conclusions would be drawn.

Another use of regression coefficients is in production of individual job-specific regression equations for personnel classification. Johnson and Zeidner (1991) have called for the use of linear programming to achieve optimal assignment of individuals to jobs by such systems. When the regression coefficients are computed in several range-restricted samples of job incumbents, the prior selection of the job incumbents causes the reliabilities of the tests to vary from sample to sample (Gulliksen 1950, 1987, p. 124, Equation 5). These varying reliabilities cause biases in the regression coefficients. In addition, the effect of the potential removal of homoscedasticity because of range restriction induced by prior selection also biases the regression coefficients. When samples are preselected and homoscedasticity is maintained, the regression coefficient in the selected sample will not show bias due to heteroscedasticity (Cohen & Cohen, 1983). The benefits of the use of these biased coefficients in optimization (Johnson & Zeidner, 1991) may be illusory and due to nothing more than the reliability artifact.

Any technique that uses regression coefficients such as clustering, profile analysis (Nunnally & Bernstein, 1994) or policy capturing (Ward & Jennings, 1973) must take the unreliability of the variables into account or inappropriate inferences will be made.

Test Bias Detection

Jensen (1980, pp. 383-386) and others (Cohen & Cohen, 1983; Crocker & Algina, 1986; Fuller, 1987) have shown that less than perfect reliability can influence the interpretation of models of test bias that rely on examination of regression slope, intercept, and standard error of estimate. What may be mistakenly interpreted as test bias may in fact be due solely to unreliability. As Jensen noted, "Before concluding that a test is intrinsically biased, it should be determined how much of the apparent bias is attributable to the unreliability of the test" (p. 383).

Test unreliability disadvantages high-scoring individuals, regardless of their group (e.g., ethnicity/race, gender, socioeconomic) membership. Therefore, any group with proportionally fewer high-scoring members will benefit (as a group) from a test's unreliability. As noted by Jensen (1980), Hunter and Schmidt (1976, p. 1056) suggested that test unreliability by itself might account for half of the overprediction of grade point average for Blacks reported in the literature.

In an unbiased test with perfect reliability, by definition, the slope, intercept, and standard error of estimate are the same for the groups being compared. Through several illustrative examples, Jensen (1980) showed that even in an unbiased test, unreliability reduces the regression slope, produces group differences in the Y intercept, and increases the standard error of estimate.

Regression Slope

In a perfectly reliable test, the observed slope will be b_{YX}. When reliability is less than 1, the slope becomes $r_{XX}b_{YX}$. If the reliability of the test were zero, the regression line would be horizontal (no slope). There is no group-difference effect of test unreliability on the slope, unless the reliabilities differ in the two groups.

Regression Intercept

Interpretation of regression intercepts is hazardous when the predictor is not perfectly reliable. Jensen (1980) showed that if the test's reliability is less than perfect and there are two groups and a single regression line, there must be two intercepts found solely because of the unreliability of the predictors. The difference in intercepts for the two groups will increase by an amount equal to $\Delta(k_A - k_B) = b_{YX}(1 - r_{XX})(\overline{X}_A - \overline{X}_B)$, where k_A and k_B are the intercepts for groups A and B and $\overline{X}_A$ and $\overline{X}_B$ are the means. Furthermore, b_{YX} is the raw score regression coefficient for the regression of Y on X, and r_{XX} is the reliability of predictor X. The expected difference in intercepts is a function of group means, regression coefficient, and predictor reliability. For example, if the regression coefficient were 1 and $\overline{X}_A$ and $\overline{X}_B$ were 10 and 5 for a test (X) with reliability of .9, the expected difference in intercepts would be 0.5. If the reliability were decreased to .7, the expected difference in intercepts would increase to 1.5. If the reliability decreased further to .5, the expected intercept difference would increase to 2.5. The nature and magnitude of the artifact is made clear when we contrast this to the circumstance in which reliability is perfect and a zero difference in intercepts is found. The uncritical interpretation of different intercepts as bias is unwarranted.

Standard Error of Estimate

Test unreliability increases the standard error of estimate (SE_Y) by an amount equal to $\Delta SE_{Y'} = \sigma_Y(\sqrt{1-(r_{XY}^2 r_{XX})}\sqrt{1-r_{XY}^2})$. Test unreliability increases the amount of overlap of the distributions of the predicted criterion scores for the two groups being compared. Finally, test unreliability decreases the standard deviation of the predicted criterion, $\sigma_{Y'} = \sigma_Y\sqrt{r_{XX}^2}$, by an amount equal to $\Delta\sigma_Y = ((1-\sqrt{1-r_{XX}^2})\sigma_Y)$.

An Example of Corrected Test Bias Detection Analyses

Carretta (1997) provided a practical example in a study of gender and ethnic group differences in the predictive utility of aptitude composites used to select U.S. Air Force pilot trainees. Uncorrected results showed group differences in predicted pilot training completion rates with overestimation for the minority group (women = .07 and Hispanics = .12) relative to the majority group (men and Whites). After correction for unreliability of the predictors, all differences were reduced to a trivial .0004 or less.

Validity Coefficient

Test unreliability reduces the validity coefficient for both groups by an amount equal to $\Delta r_{XY} = (\sqrt{1-r_{XX}^2}\, r_{XY})$. In addition, test unreliability increases the amount of overlap of the distributions of the predicted criterion scores for the two groups being compared. Finally, test unreliability decreases the standard deviation of the predicted criterion, $\sigma_{Y'} = \sigma_Y\sqrt{r_{XX}^2}$, by an amount equal to $\Delta\sigma_Y = (\sqrt{1-r_{XX}^2}\,\sigma_Y)$.

A particularly interesting situation occurs in the tests of predictive bias (Cole, 1973) using regression models (Lautenschlager & Mendoza, 1986). Usually the first test of models in the detection of bias is a comparison of a four-parameter regression model against a two-parameter model. The two models tested are

$$\hat{Y} = a_1 + b_1 S + b_2 X + b_3 XS \tag{9}$$

and

$$\hat{Y} = a_4 + b_4 X \tag{10},$$

where X is a test score, S is a categorical variable (often called a dummy variable) of 0 and 1 denoting group membership, and XS is the cross-product of X and S. Note that XS has a peculiar distribution, with zeros for the group coded 0 and test scores for the group coded 1. In the first model, a_1 and b_1 are intercepts and b_2 and b_3 are slopes. In the simpler model, a_4 is the intercept and b_4 is the slope. The first regression model can provide two lines; the second regression model can provide only one line. Frequently, the two groups considered have a mean score difference of 1σ. Consequently, the test has a different reliability for each group, and depending on the placement of the minimum cut score, the reliability may be made to differ further between the groups after selection. The effects of unreliability in the full model are

more difficult to specify than in the reduced model, and comparison of the models in the presence of measurement error may lead to inappropriate inferences in the population.

Factor Analysis

The role of reliability in factor analysis is well known and generally straightforward. The general model of factor analysis is that the variance of the observed variable is a linear combination of common factors and unique factors. The ratio of the variance associated with the common factor to the total variance of an observed variable is known as the communality of the observed variable (Fuller, 1987, pp. 60-61). Using Fuller's notation, communality can be written as

$$k_{11}^2 \equiv [\beta_{11}^2 \sigma_{xx} + \sigma_{ee11}]^{-1} \beta_{11}^2 \sigma_{xx} \equiv 1 - \sigma_{YY11}^{-1} \sigma_{ee11}. \tag{11}$$

This quantity k_{11}^2, the communality, is an estimate of the reliability of the variable. The communality of a variable provides a lower bound estimate of its reliability (Baggaley, 1964). It is a lower bound estimate because it does not include the reliable variance measured by specific factors. The unique variance or uniqueness of a variable is (1 – communality). The unique factors are composed of specific variance and error variance. Symbolically, these relationships can be expressed as

$$X = h + u \tag{12}$$

or

$$X = h + s + e, \tag{13}$$

where X is an observed variable, h is the commonality, u is the uniqueness, s is the specific, and e is the error.

If the variable is associated with the factor, as the reliability of the observed variable increases and the error decreases, the loadings of the variable on the factors can be expected to increase. For example, if there are three variables that have true loadings that are equal but are measured with differing reliability, the observed loadings will differ as a function of the reliability, with the more reliably measured variables receiving higher loadings.[3] Interpretations of these observed loadings will lead to erroneous conclusions about the factorial causation of the variables because the differences are due to differing reliabilities and not differing relationships to the factor. To correct factor loadings for unreliability, the loadings for the observed variables can be divided by the reliability of the observed variables. These corrected loadings give more appropriate estimates of the true relationships of the factors to the observed variables.

Ree and Carretta (1998) reported a study that showed the correlation between the unrotated first-factor loadings of multiple aptitude battery[4] scores and average validity of those scores. The correlation was .76. The factor loadings were then corrected for unreliability, and the correlation became .98.

Analysis of Variance (ANOVA) and Analysis of Covariance (ANCOVA)

ANOVA and ANCOVA are examples of the linear model as is regression analysis. The effects of measurement error are similar; however, the independent variables in ANOVA are usually uncorrelated owing to random assignment of participants. As stated above, the effects of unreliability on uncorrelated independent variables are simpler.

ANOVA

Let us consider a one-way ANOVA with three levels of the independent variable with μ_1, μ_2, and μ_3. Remembering that ANOVA is a linear model and that the parameter estimates can be found by means of regression, we note that $\mu_1 = \alpha + \beta_1$, $\mu_2 = \alpha + \beta_2$, and $\mu_3 = \alpha$, where α is the regression additive constant (intercept) and β_1 and β_2 are the multiplicative partial regression coefficients for the two categorical variables needed to represent the three levels of the independent variable. Furthermore, note that $\alpha = \mu_3$, $\beta_1 = \mu_1 - \mu_2$, and $\beta_2 = \mu_2 - \mu_3$. Suppose $\mu_1 = +1$, $\mu_2 = 0$, and $\mu_3 = -1$ and that the reliabilities $r_{XX1} = r_{XX2} = r_{XX3} = .50$. The true differences are 1 or 2 points, but there is a loss of statistical power. In addition, the effect size (e.g., $(\mu_1 - \mu_2)/\sigma$) may be substantially underestimated because σ is inflated by error variance. Much the same may be found in an *N*-way ANOVA. Consider a two-way ANOVA with the independent variables of gender and political party affiliation. There are two gender (male/female) and three political affiliation (Democrat, Independent, and Republican) levels, respectively. Using the same logic as before, the group means may be represented as follows:

Gender	Political Affiliation	Group Means
Female	Democrat	$\mu_{f1} = \alpha + \beta_1 + \beta_3$
	Independent	$\mu_{f2} = \alpha + \beta_2 + \beta_3$
	Republican	$\mu_{f3} = \alpha + \beta_3$
Male	Democrat	$\mu_{m1} = \alpha + \beta_1$
	Independent	$\mu_{m2} = \alpha + \beta_2$
	Republican	$\mu_{m3} = \alpha$

Again, both statistical power and effect size may be reduced. If the two independent variables above are correlated, as they well may be given the impossibility of randomly assigning gender and party affiliation, the same biases will be found as in a multiple regression with correlated predictors. With less than perfectly reliable variables, the results can be very misleading.

An instructive example is provided in the work of Guttman (2000) in a study of 16- to 40-year-old females. Independent variables for the analysis of variance were based on meeting the criteria in the *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed., revised; American Psychiatric Association, 1987) for the clinical conditions of anorexia nervosa and borderline personality disorder. The control participants were admitted to the study on three less than perfectly reliable self-report clinical instruments. Of particular interest were the dependent variables assessed by a 28-item measure of an individual's cognitive and emotional capacity for empathy. This instrument yielded four scales whose median reliability was reported by Guttman to be about .70. The *DSM-III-R* assessments have less than perfect

reliability, and .6 is a reasonable approximation of the reliability of the categorization into the anorexic and borderline personality groups. Given the magnitudes of less than perfect reliability in both the independent and dependent variables, it is likely that all effect sizes were underestimated and many significant differences were undetected.

ANCOVA

ANCOVA is a linear model with categorical variables and at least one continuous variable as a covariate. For example, suppose we were interested in examining the effects of political party affiliation, a three-level categorical variable (Democrat, Independent, and Republican) and annual income measured in dollars earned (a continuous variable) on amount of support for the president's proposed budget (a continuous variable). The three levels of party affiliation are represented by two categorical independent variables (X_1 and X_2). Income level (X_3) is the continuous covariate independent variable. The following linear model represents the relationship of party affiliation and income to the dependent variable, support for the president's budget (Y):

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3. \tag{14}$$

Considering less than perfect reliability of the independent variables, the equation can be rewritten as

$$Y = \alpha' + r_{11}\beta_1 X_1 + r_{22}\beta_2 X_2 + r_{33}\beta_3 X_3, \tag{15}$$

where r_{11}, r_{22}, and r_{33} are reliabilities and α' denotes the additive regression coefficient affected by the unreliability of the independent variables.

The effects of less than perfect reliability will be found and the higher the correlation between the covariate X_3 and the independent variables X_1 and X_2, the more bias will be noted in the analysis and the greater the loss of statistical power. The same biases will be found as would be found in a multiple regression with correlated predictors.

Ameliorating or Correcting the Problem

In each of the cases we reviewed, it has been shown that using less than perfectly reliable variables creates bias in the parameter estimates. This reduces statistical power and provides the opportunity for misinterpretation of findings and misstatement of fundamental relationships.

There are several ways to ameliorate or correct this problem. The first approach is to use variables that yield highly reliable scores for your sample (see Ree, Carretta, & Steindl, 2001). Revising unreliable test items or observational techniques, adding test items or observations, revising vague or confusing instructions, and clarifying ambiguous scoring and coding procedures can accomplish this. This alleviates most of the problems but does not entirely remove the bias due to unreliability. A second approach is to correct the observed variables for the effects of unreliability and conduct the analyses on the corrected values. This can be accomplished with reliability estimates from the participant sample in the study. Finally, the use of latent variable analyses, such as confirmatory factor analyses or structural equation

modeling, which eliminate or substantially reduce the unreliability of the variables, is a third worthwhile approach.

Cohen and Cohen (1983, p. 411) reported that Dunivant (1981) conducted simulation studies to evaluate the last two approaches and concluded that both "have merit and yield reasonable results." Unreliability poses a threat to our knowledge and practice, whether in theoretical studies or in practical application. Baugh (2002) expressed it well, stating, "As the winds of change continue to shape responsible research practice, it is hoped that researchers will give more thoughtful consideration to the influence that measurement error variance exerts" (p. 261).

Notes

1. We follow the convention of using Greek letters for population parameters and Roman letters for sample statistical estimates of population parameters. Equation 1 is the single exception to this rule as many are familiar with the equation when written as presented.

2. Due to sampling error, the observed correlation could take on numerous higher values. We present the maximum expected observed correlation.

3. We noted a similar finding for regression coefficients in the section on multiple regression. In factor analyses of test items or questionnaire items, it may be difficult to estimate the reliability of items.

4. The unrotated first factor of a multiple aptitude battery is a measure of general cognitive ability (*g*).

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., rev). Washington, DC: Author.

Baggaley, A. R. (1964). *Intermediate correlational techniques*. New York: Wiley.

Baugh, F. (2002). Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably. *Educational and Psychological Measurement, 62*, 254-263.

Carretta, T. R. (1997). Group differences on U.S. Air Force pilot selection tests. *International Journal of Selection and Assessment, 5*, 115-127.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Cole, N. S. (1973). Bias in selection. *Journal of Educational Measurement, 10*, 237-255.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Harcourt Brace Jovanovich.

Dunivant, N. (1981). *The effects of measurement error on statistical models for analyzing change* (Final Report, Grant NIE-G-78-0071). Washington, DC: National Institute of Education, U.S. Department of Education.

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

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Gulliksen, H. (1987). *Theory of mental tests*. Mahwah, NJ: Lawrence Erlbaum.

Guttman, H. A. (2000). Empathy in families of women with borderline personality disorder, anorexia nervosa, and a control group. *Family Process, 39*, 345-358.

Hunter, J. E., & Schmidt, F. L. (1976). A critical analysis of the statistical and ethical implications of various definitions of "test bias." *Psychological Bulletin, 83*, 1053-1071.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis*. Newbury Park, CA: Sage.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis* (2nd ed.). Thousand Oaks, CA: Sage.

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

Johnson, C., & Zeidner, J. (1991). *The economic benefits of predicting job performance: Volume 2. Classification efficiency*. New York: Praeger.

Lautenschlager, G. J., & Mendoza, J. (1986). A step-down hierarchical multiple regression analysis for estimating hypotheses about test bias in prediction. *Applied Psychological Measurement, 10*, 133-159.

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

Ree, M. J., & Carretta, T. R. (1998). General cognitive ability and occupational performance. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial organizational psychology, 1998* (pp. 159-184). Chichester, UK: Wiley.

Ree, M. J., Carretta, T. R., & Steindl, J. R. (2001). Cognitive ability. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Vol. Eds.), *International handbook of work and organizational psychology* (Vol. 1, pp. 219-232). London: Sage.

Ree, M. J., & Earles, J. A. (1993). *g* is to psychology what carbon is to chemistry: A reply to Sternberg and Wagner, McClelland, and Calfee. *Current Directions in Psychological Science, 2*, 11-12.

Russell, T. L., & Peterson, N. G. (2002). The experimental battery: Basic attribute scores for predicting performance in a population of jobs. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification* (pp. 269-306). Mahwah, NJ: Lawrence Erlbaum.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology, 15*, 201-293.

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

Thompson, B. (2003). Understanding reliability and coefficient alpha, really. In B. Thompson (Ed.), *Score reliability* (pp. 3-30). Thousand Oaks, CA: Sage.

Thorndike, R. L. (1949). *Personnel selection*. New York: Wiley.

Ward, J. H., Jr., & Jennings, E. (1973). *Introduction to linear models*. Englewood Cliffs, NJ: Prentice Hall.

Malcolm James Ree received his Ph.D. in psychometrics and statistics from the University of Pennsylvania in 1976. Currently, he is a professor of leadership studies at Our Lady of the Lake University in San Antonio, Texas. Previously, he was a senior scientist in the Manpower and Personnel Research Division at the Air Force Research Laboratory. His research interests include multivariate methods, problems in estimation, individual differences, and the history of statistics.

Thomas R. Carretta received a Ph.D. in psychology in 1983 from the University of Pittsburgh. Currently, he is an engineering psychologist in the System Control Interface Branch of the Human Effectiveness Directorate of the Air Force Research Laboratory (AFRL) at Wright-Patterson Air Force Base, Ohio, and conducts research regarding human factors issues in crew-system interface development. Prior to his current position, he spent more than 12 years in the Manpower and Personnel Research Division of the AFRL in San Antonio, Texas, working on aircrew selection and classification issues including test development and validation and the role of general and specific abilities in skill acquisition. His professional interests include personnel measurement, selection, classification, and individual and group differences. He has published more than 50 journal articles and book chapters on topics related to these research interests.