Criterion-Related Validity in Multiple-Hurdle Designs: Estimation and Bias

JORGE L. MENDOZA
DAVID E. BARD
MICHAEL D. MUMFORD
SIEW C. ANG
University of Oklahoma

Employee selection often involves a series of sequential tests (or hurdles). However, validation strategies under this complex design are not found in the literature. Missing is a discussion of the statistical properties important in establishing criterion-related validity in multiple-hurdle designs. The authors address this gap in the literature by suggesting a general statistical model for range restriction corrections. Because the multiple-hurdle design includes as special cases predictive and concurrent designs, the corrections apply also to these designs. The general correction model is based on algorithms from the missing data literature. Two missing data procedures are examined: the estimation-maximization procedure and the Bayesian multiple imputation (MI) procedure. These procedures are large-sample equivalent and often yield similar results. The MI procedure, however, has the added advantage of providing easily obtainable standard errors. A hypothetical example of a multiple-hurdle design is used to illustrate the procedures.

Keywords: *selection; range restriction; multiple-hurdle design; missing data; corrections*

Many validation strategies can be used to provide evidence for the meaningfulness of a test (Messick, 1998). Nonetheless, the criterion-related validation strategy remains one of the two most commonly applied test validation strategies; the other is content validation. The goal of any criterion-related validation effort is quite straightforward. Essentially, an attempt is made in a criterion-related validation effort to show that the test is related to one or more performance outcomes of interest (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

Authors' Note: We are grateful to the reviewers for comments on earlier versions of this article. Correspondence regarding this article should be directed to Jorge L. Mendoza, Psychology Department, University of Oklahoma, Norman, OK 73069; e-mail: jmendoza@ou.edu.

Organizational Research Methods, Vol. 7 No. 4, October 2004 418-441
DOI: 10.1177/1094428104268752

The main goal of any criterion-related validation effort is to provide an estimate of the population, or true, relationship between the test under consideration and the outcome of interest. However, a number of design considerations are known to influence the accuracy of the estimates of the test-criterion relations obtained in these studies. For example, the accuracy of these estimates may be influenced by unreliability in the test, or criterion, measures (Bobko, 1995; Mendoza & Mumford, 1987; Sackett, Laczo, & Arvey, 2002) as well as the operation of contaminating variables distorting the test-criterion relationship within the sample at hand (James, Demaree, Muliak, & Mumford, 1988). However, one of the most pervasive problems influencing the accuracy of estimates obtained in this validation design is range restriction. The reason range restriction exerts a pervasive biasing effect on criterion-related validation studies arises from the simple fact that any prior, or concurrent, selection decision involving an attribute related to a test generally results in a reduction of the observed test-criterion relationship vis-à-vis the population test-criterion relationship (Ghiselli, 1966; Lord & Novick, 1968; R. T. Linn, 1968). Range restriction is moreover endemic in most selection situations in which prior selection and existing tests act to induce restriction.

This problem is not new: Pearson (1903) and, later, Lawley (1943) proposed a set of equations for correcting observed correlations for range restriction (direct and indirect). Modifications and extensions of these equations have appeared in a number of studies (see Greener & Osburn, 1979; Gross & McGanney, 1987; Ree, Carretta, Earles, & Albert, 1994; Sackett & Yang, 2000). The correction for direct range restriction assumes that the regression line characterizing the test-criterion relationship is constant in both the unrestricted and the restricted populations. It also assumes that the error variances are equal in both the unrestricted and restricted populations. These assumptions hold when the predictor and criterion are jointly normally distributed (before selection). Implicitly, the range restriction correction also assumes that a valid estimate of the unrestricted variance for the predictor is obtainable (Hoffman, 1995). Sackett and Ostgard (1994) and, later, Ones and Viswesvaran (2003) have studied this issue and concluded that in many validation situations, the predictor variance obtained in an applicant pool, although often affected by self-selection, is a fairly accurate estimate of the unrestricted variance. Corrections for range restriction are commonly applied in meta-analyses in which an attempt is made to identify the true test-criterion relationship as well as in many validation designs in which range restriction is present.

Pearson (1903) and Lawley (1943), in their development of the correction formulas, did not address estimation or hypothesis testing, and the specific distributional properties of their corrections are still unknown. They developed the corrections formulas at the population level not dealing with estimation. Using computer simulations, however, researchers have shed light on the distributional properties of their corrections (e.g., see Greener & Osburn, 1979, 1980). However, others (e.g., Gross, 1990; Mendoza, 1993) have shown that at the sample level, these corrections are maximum likelihood estimators of the population parameters. The main problem with the Pearson-Lawley approach, however, is that as given, it does apply to complex but realistic validation designs. An example of this is a validation design in which multiple selection tests are used sequentially: a multiple-hurdle design. The multiple-hurdle design has been described by Sackett and Yang (2000) as “simultaneous or sequential selection, all variables measured, and unrestricted variances are not known for one or more selection variables” (p. 116).

According to Sackett and Yang (2000), corrections for this design have not appeared in the literature. This omission is of some importance because many selection systems are implemented in a sequential mode. Our interest in the present effort is to offer a general approach, with a solid statistical framework, to correct observed test-criterion relationships for range restriction in designs involving sequential selection. The interested reader is referred to Sackett and Yang (2000) for an excellent review and taxonomy of current correction procedures.

The corrections given here for the multiple-hurdle validation design comprehend the Pearson-Lawley corrections and are based on missing data procedures. These procedures were developed by Rubin (1978) and others (e.g., see Dempster, Laird, & Rubin, 1977; Little & Rubin, 1987) and are based on multiple imputation and maximum likelihood estimation. Because at the sample level, the Pearson-Lawley corrections are maximum likelihood estimators, they are subsumed under the more general corrections given here. By observing that range restriction is a special case of missing data, we develop a comprehensive set of corrections and principles that apply to most selection designs. In addition, we observe that a multiple-hurdle validation is equivalent to a monotonic missing data structure. Showing that the monotonic data structure is equivalent to a multiple-hurdle validation design allows us to borrow simple estimation procedures from the missing data literature. In complex selection situations such as the multiple-hurdle validation design, these missing data procedures are easier to implement and more versatile than the Pearson-Lawley corrections, yielding estimates and their standard errors that can be used to form confidence intervals and test hypotheses.

Parameter estimation under range restriction is feasible only when we meet either the missing at random (MAR) or missing completely at random (MCAR) assumption (Rubin, 1976). Under MCAR, the more restrictive of the two assumptions, the missing data are truly missing at random and as such are not related to either observed or any other data. For example, if we were to hire at random from a pool of applicants, the criterion data for those not selected would be MCAR. Clearly, the MCAR assumption is not likely to be met in most selection situations. The assumption that is likely to be met in most selection situations is the MAR assumption. For the moment, it suffices to say that MAR assumes that the criterion data are missing because of decisions made with data (test results) under our control (observed). The criterion values are missing because the test scores were too low and the applicants were not selected. The basic idea is that because we can establish a relationship between the missing and the observed, we can use the observed data to “fill in” for the missing data, thus enabling us to estimate the parameters of the referent population. Although random is part of the definition, the MAR assumption does not mean that the observed data are a random sample from the referent population. Instead, the MAR assumption implies that the observed data contain sufficient information to estimate the regression parameters, which in turn can be used to fill in for the missing data. We believe that in most multiple-hurdle (and simpler) validation designs, the MAR assumption is likely to be met. In those rare instances in which the MAR (or the MCAR) assumption is not met, the missing data are missing not at random (MNAR), and we must model the missing data mechanism to estimate the parameters. These procedures are complex and will not be discussed here.

Although at the sample level, correction formulas yield comparable results to missing data procedures, their value lies in being able to show explicitly the nature of the statistical bias. The formulas are especially useful in helping us understand the biasing effects present in multiple-hurdle designs. They also provide a general algebraic solution to the difficult problem of correcting for range restriction under conditions in which multiple hurdles have been employed for selection. Before turning to their development, we consider the nature of the bias arising in regression and correlation estimates in three (concurrent, predictive, and multiple-hurdle) criterion-related validation designs.

Biases in Regressions and Correlations Due to Selection

In a concurrent validity design, a (new) test is validated by administering it to the incumbents, then correlating the scores with a measure of performance. The important point of the design is that these incumbents had been previously selected by another test (the old test) and that the predictor and criterion information are collected at about the same time. The selection process, as it should, has biased the sample by bringing into the organization those more qualified. The incumbents are not a random sample of the general population. Consequently, statistical procedures that assume a random sample are not appropriate.

There are two basic statistical questions that must be asked in a concurrent validity study: What is the correlation between the criterion (generally some measure of job performance) and the (new) test? And is the correlation between the new test and the criterion larger than the correlation between the old test and criterion? These two questions are fundamentally at the population level. Consequently, we must be able to accurately estimate the two population correlations. But these correlations are difficult to estimate: They often suffer from range restriction and unreliability. The reliability of the criterion and predictor are important in that they attenuate the correlation between predictor and criterion. Reliability also affects the regression of the criterion on the predictor. In this article, however, we focus on the effects of range restriction on estimation.

In contrast to the concurrent design, in a predictive validation design, the new and old tests (in some applications, only the new test is given) are given to the applicants, but only one test is used in the decision to hire. A predictive study aims to establish how accurately test scores can predict criterion scores that are obtained at a later time. As in the concurrent design, the two basic questions are as follows: What is the correlation between the criterion and the (new) test? Or, as stated in the 1999 standards, “How accurately do test scores predict criterion performance?” And is the correlation between the new test and the criterion larger than the correlation between the old test and criterion? The predictive situation is different from the concurrent situation; rather than having the test scores on the incumbents, we have test scores on the applicants.

In a multiple-hurdle situation, an applicant progresses from one stage to another only after passing a specific test (or hurdle). The number of hurdles varies, but usually a system does not include more than three or four hurdles. For example, the recently enacted Transportation Security Administration used a multiple-hurdle design to select airport screeners (Kolmstetter, 2003). The first hurdle in this selection system involved an online application that was used to screen for minimum qualifications. Eligible applicants then were invited to take a computerized test battery (with a 48% pass rate). Those who passed the test proceeded to the next hurdle. The next hurdle entailed a structured interview, a physical abilities test, and a medical evaluation (with an 86% pass rate). Those who passed received employment offers with employment contingent on passing a security background check.

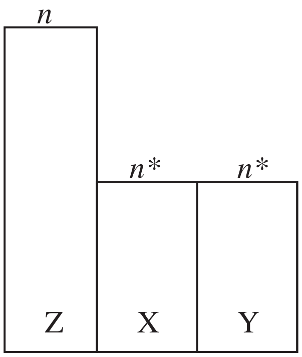
The multiple-hurdle approach is often useful when job training is long, complex, or expensive, or when the organization is very selective or has a large applicant pool. Although these selection situations are well known and are often mentioned in measurement and industrial psychology texts, the statistical issues of the multiple-hurdle validation designs are not well understood. It should be pointed out that a concurrent (or predictive) design containing a new test and an old test is a simple version of a multiple-hurdle design. Next, we discuss the statistical properties of multiple-hurdle designs.

To develop the statistical argument, we introduce three variables, say **Z**, **X**, and **Y**, as Ghiselli (1964) did in his three-variable case. Without loss of generality, we conceptualize each variable either as a scalar or as a vector containing several measures. We will let **Z** be a vector with p_z elements, **X** a vector with p_x elements, and **Y** a vector with p_y elements. In addition, because of the sequential nature of the multiple-hurdle design, we denote the number of original observations as n, the number of observations selected with **Z** as n^*, and the number of observations selected with **X** as n^{**}.

In a concurrent validation design (a single-hurdle situation), we select with **Z** and observe both **X** and **Y** some time after the individuals have been selected. (The **Z** in this design could be the "old" test used to select the incumbents, **X** the "new" test that we wish to validate, and **Y** the measure of performance.) Figure 1 illustrates this concurrent validation situation. It is known that in this situation, we encounter both direct range restriction and indirect range restriction. We encounter direct range restriction when we look at the relations between **Z** and **X** and between **Z** and **Y** and indirect range restriction when we look at the relation between **X** and **Y**. Under direct restriction, the regression weights are unbiased, but the correlations are biased. Under indirect range restriction, both the correlation coefficients and regression weights (and the standard errors) are biased (Sackett & Yang, 2000).

Figure 1 illustrates which regressions and correlations are biased in a concurrent validity study. We can see from Figure 1 that all the correlations are biased including the multiple R for the regression of **Y** on (**Z**, **X**). On the other hand, all of the regression weights are unbiased except for the regression of **Y** on **X**. For the unbiased regressions, the standard errors are also correct and can be used to test hypotheses about the population regression weights. We will show later why the regression and correlations are biased in some situations. To answer the first question posed, we must obtain an unbiased estimate of the correlation (between **Y** and **X**) and regression of **Y** on **X**. Both the correlation and regression are biased under this design. They both must be adjusted for range restriction to answer the first question appropriately.

Figure 2 illustrates a two-hurdle situation. In this situation, we first select with **Z**, then we select with **X**. The criterion **Y**, again, is available only for those who were selected. Usually, **Y** takes the form of some sort of performance appraisal obtained 6 to 12 months after selection. As Figure 2 illustrates, only the regression of **Y** on (**Z**, **X**) and the regression of **X** on **Z** (based on n^* cases) are unbiased. The other regressions are biased and must be corrected for range restriction. All correlations, as in the concurrent design, are biased and must be corrected for range restriction. These observations generalize to selection designs involving more than two hurdles.



Model	Regression[a]	R^2	Sample Size
Y= z x	unbiased	biased	$n*$
Y= z	unbiased	biased	$n*$
Y= x	biased	biased	$n*$
X= z	unbiased	biased	$n*$

Figure 1: Regressions and Correlations in a Concurrent Validation Design: Selection With **Z**

a. Unbiased (or biased) refers to regression model specified on left.

Figure 3 illustrates a predictive validation design. In this situation, we administer **Z** and **X** to the applicants, then select with **Z** ("old" test). The criterion **Y** is obtained some time after selection, say, 6 to 12 months after selection. In this design, the correlation and regression of **X** on **Z**, based on the *n* observations, are unbiased and need not be corrected. Also, the regression of **Y** on (**Z**, **X**) and that of **Y** on **Z** are unbiased. However, the regression of **Y** on **X** must be corrected for indirect range restriction. When we look at the correlations, we see that the correlations between **Z** and **Y**, and between **X** and **Y**, are biased and must be corrected for range restriction. Not surprisingly, in each of the three designs, the correlation and regression of **Y** on **X** are biased and must be corrected.[1]

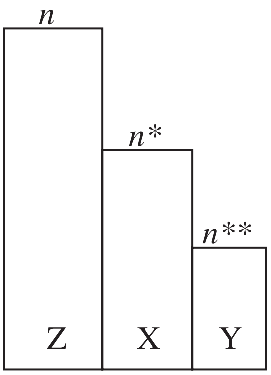
Statistical Argument

Single-Hurdle Selection: Concurrent Validation

To obtain the regressions and correlations in the unrestricted population, we must first find the unrestricted variance-covariance matrix Σ (of **Z**, **X** and **Y**). The statistical task in the single-hurdle design illustrated in Figure 1 involves obtaining Σ from the restricted variance-covariance matrix Σ* of the selected (marginal) population:

$$\Sigma^* = \begin{pmatrix} \Sigma^*_{zz} & \Sigma^*_{zx} & \Sigma^*_{zy} \\ & \Sigma^*_{xx} & \Sigma^*_{xy} \\ & & \Sigma^*_{yy} \end{pmatrix} \xrightarrow{to} \begin{pmatrix} \Sigma_{zz} & \Sigma_{zx} & \Sigma_{zy} \\ & \Sigma_{xx} & \Sigma_{xy} \\ & & \Sigma_{yy} \end{pmatrix} = \Sigma.$$

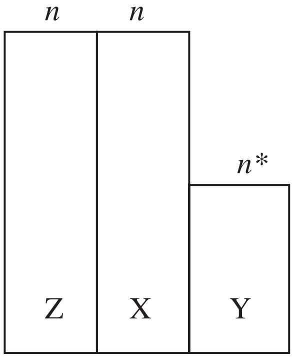
The submatrices in Σ* are all observable. The submatrices in Σ, however, are not and must be obtained from the observed matrices. The only observable submatrix in Σ is Σ_{zz}, which is based on the "applicants."



Model	Regression[a]	R^2	Sample Size
Y= z x	unbiased	biased	n^{**}
Y= z	biased	biased	n^{**}
Y= x	biased	biased	n^{**}
X= z	biased	biased	n^{**}
X= z	unbiased	biased	n^{*}

Figure 2: Regressions and Correlations in a Two-Hurdle Validation Design: Selections With **Z** Then **X**

a. Unbiased (or biased) refers to regression model specified on left.



Model	Regression[a]	R^2	Sample Size
Y= z x	unbiased	biased	n^{*}
Y= z	unbiased	biased	n^{*}
Y= x	biased	biased	n^{*}
X= z	unbiased	unbiased	n
X= z	unbiased	biased	n^{*}

Figure 3: Regressions and Correlations in a Predictive Validation Design: Selection With **Z**

a. Unbiased (or biased) refers to regression model specified on left.

Assuming that **Z**, **X**, and **Y** are multivariate jointly normally distributed (before selection) and applying Lawley's (1943) results,[2] we find that the relationship between the restricted and unrestricted matrices is

$$\begin{aligned} \Sigma_{zx} &= \Theta^{*}\Sigma^{*}_{zx} \\ \Sigma_{zy} &= \Theta^{*}\Sigma^{*}_{zy}, \end{aligned} \tag{1}$$

where

$$\Theta^* = \Sigma_{zz}\Sigma_{zz}^{*-1}.$$

Because the covariance between **X** and **Y** has been affected by indirect restriction, untangling the relation between restricted and unrestricted parameters is a bit more complicated. Following Ree et al. (1994), we can show that the relation is

$$\begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ & \Sigma_{yy} \end{pmatrix} = \begin{pmatrix} \Sigma_{xx}^* & \Sigma_{xy}^* \\ & \Sigma_{yy}^* \end{pmatrix} + \begin{pmatrix} \Sigma_{xz}^*\Phi^*\Sigma_{zx}^* & \Sigma_{xz}^*\Phi^*\Sigma_{zy}^* \\ & \Sigma_{yz}^*\Phi^*\Sigma_{zy}^* \end{pmatrix}, \tag{2}$$

where

$$\Phi^* = \Sigma_{zz}^{*-1}\Sigma_{zz}\Sigma_{zz}^{*-1} - \Sigma_{zz}^{*-1}.$$

Noting that the regression of **X** on **Z** in the selected population is $B_{x.z}^* = \Sigma_{zz}^{*-1}\Sigma_{zx}^*$ and that the regression of **Y** on **Z** is $B_{y.z}^* = \Sigma_{zz}^{*-1}\Sigma_{zy}^*$, we can rewrite Equations 1 and 2 in terms of these regressions as follows:

$$\begin{aligned} \Sigma_{zx} &= \Sigma_{zz}^* B_{x.z}^* \\ \Sigma_{zy} &= \Sigma_{zz}^* B_{y.z}^* \end{aligned} \tag{3}$$

$$\begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ & \Sigma_{yy} \end{pmatrix} = \begin{pmatrix} \Sigma_{xx}^* & \Sigma_{xy}^* \\ & \Sigma_{yy}^* \end{pmatrix} + \begin{pmatrix} B_{x.z}^{*\prime}(\Sigma_{zz} - \Sigma_{zz}^*)B_{x.z}^* & B_{x.z}^{*\prime}(\Sigma_{zz} - \Sigma_{zz}^*)B_{y.z}^* \\ & B_{y.z}^{*\prime}(\Sigma_{zz} - \Sigma_{zz}^*)B_{y.z}^* \end{pmatrix}.$$

Both of these regressions are computed on the $n*$ cases. From Equation 3, we can see that the variance-covariance matrix of the unrestricted population is a function of Σ_{zz} and the restricted parameters Σ_{zz}^*, $B_{x.z}^*$, and $B_{y.z}^*$.

Next, we turn our attention to the relation between the regressions in the unrestricted and restricted populations. Recall that the regression of **X** on **Z** is $B_{x.z} = \Sigma_{zz}^{-1}\Sigma_{zx}$, and that the regression of **X** on **Z** on the restricted population is $B_{x.z}^* = \Sigma_{zz}^{*-1}\Sigma_{zx}^*$. From Equations 1 and 2, we see that these two regressions are equal:

$$B_{x.z}^* = \Sigma_{zz}^{*-1}\Sigma_{zx}^* = (\Sigma_{zz}^{-1}\Theta^*)(\Theta^{*-1}\Sigma_{zx}) = \Sigma_{zz}^{-1}\Sigma_{zx} = B_{x.z}.$$

Similarly, we can show that $B_{y.z} = B_{y.z}^*$. The regressions are not affected by direct range restriction and can be computed on the restricted population without having to be corrected. Under multivariate normality, it is not hard to show using a similar argument that direct range restriction does not affect the conditional variances, that is, $\Sigma_{xx.z} = \Sigma_{xx.z}^*$, $\Sigma_{yy.z} = \Sigma_{yy.z}^*$ and $\Sigma_{xy.z} = \Sigma_{xy.z}^*$. Incidentally, Ghiselli (1964, p. 364), in deriving his formula to correct the correlation between **X** and **Y** for direct and indirect range restriction, assumed that $\Sigma_{xy.z} = \Sigma_{xy.z}^*$. This assumption, as we pointed out, will be met under multivariate normality. His correction equations are obtainable under our set of equations.

We now turn our focus to the regression of **Y** on **X**. In this situation, indirect range restriction, rather than direct, affects the covariance between **X** and **Y**. Under indirect selection, the regressions are not equal. The reason for the inequality is not difficult to show. Consider the regression of **Y** on **X** on the unrestricted $B_{y.x} = \Sigma_{xx}^{-1}\Sigma_{xy}$ and restricted $B_{y.x}^{*} = \Sigma_{xx}^{*-1}\Sigma_{xy}^{*}$ populations. Again, using Equation 2, we rewrite the unrestricted regression in terms of the restricted parameters and find that

$$B_{y.x} = (\Sigma_{xx}^{*} + \Sigma_{xz}^{*}\Phi^{*}\Sigma_{zx}^{*})^{-1}(\Sigma_{xy}^{*} + \Sigma_{xz}^{*}\Phi^{*}\Sigma_{zy}^{*}). \tag{4}$$

We can see from Equation 4 that the restricted and unrestricted regressions are not equal unless the covariance between **X** and **Z** is zero. In other words, the unrestricted and restricted regressions are equal only when **Z** is uncorrelated with **X**.

Two-Hurdle Selection

Next, we discuss the two-hurdle situation given in Figure 2. In the two-hurdle situation, we assume that n^* individuals are selected with **Z** from a pool of n individuals. Then, n^{**} are selected with **X** from the n^* previously selected ($n > n^* > n^{**}$). This selection process yields what Little and Rubin (1987) have called a monotonic missing data structure. Again, we are interested in obtaining the unrestricted variance-covariance matrix Σ from the variance-covariance matrices Σ* and Σ** of the selected (marginal) populations:

$$\begin{pmatrix} \Sigma_{zz}^{*} & \Sigma_{zx}^{*} & \Sigma_{zy}^{**} \\ & \Sigma_{xx}^{*} & \Sigma_{xy}^{**} \\ & & \Sigma_{yy}^{**} \end{pmatrix} \xrightarrow{to} \begin{pmatrix} \Sigma_{zz} & \Sigma_{zx} & \Sigma_{zy} \\ & \Sigma_{xx} & \Sigma_{xy} \\ & & \Sigma_{yy} \end{pmatrix} = \Sigma.$$

The matrices with one asterisk (*) are observed after selection on **Z** and are based on n^* cases. The matrices with two asterisks (**) are observed after selection on **X** and are based on n^{**} cases. Consequently, the matrices Σ_{zz}, Σ_{zz}^{*}, Σ_{zz}^{**}, Σ_{zx}^{*}, Σ_{zx}^{**}, Σ_{xx}^{*}, Σ_{xx}^{**}, Σ_{zy}^{**}, Σ_{xy}^{*}, Σ_{yy}^{**} are the ones observable. The matrices that are not observable, and must be solved for, are the matrices Σ_{zx}, Σ_{xx}, Σ_{zy}, Σ_{xy}, Σ_{yy}. We first solve for the matrices involving selection with **Z**. From Equations 1 and 2, we get

$$\begin{aligned} \Sigma_{zx} &= \Theta^{*}\Sigma_{zx}^{*} \\ \Sigma_{xx} &= \Sigma_{xx}^{*} + \Sigma_{xz}^{*}\Theta^{*}\Sigma_{zx}^{*} = \Sigma_{xx}^{*} + B_{x.z}^{*\prime}(\Sigma_{zz} - \Sigma_{zz}^{*})B_{x.z}^{*} \end{aligned} \quad . \tag{5}$$

Because the multiple-hurdle situation involves direct selection with **Z** and **X**, we use our previous results involving equality of slopes and conditional variances to solve for the remaining matrices. We solve for the matrices Σ_{zy}, Σ_{xy} by noting that the slopes in the unrestricted and restricted populations are equal, $B_{y.zx} = B_{y.zx}^{**}$ This equality follows from the normality assumption. Next, consider the regression of **Y** on (**Z**, **X**), and partition the predictor vector so it corresponds to the elements in **Z** and **X** such that

$$B_{y.zx} = \Sigma^{-1}_{zx,zx}\Sigma_{zx,y} = \begin{pmatrix} \Sigma_{zz} & \Sigma_{zx} \\ \Sigma_{xz} & \Sigma_{xx} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{zy} \\ \Sigma_{xy} \end{pmatrix} = \begin{pmatrix} B_z \\ B_x \end{pmatrix}. \tag{6}$$

Rewriting Equation 6, we can see that

$$\begin{pmatrix} \Sigma_{zz} & \Sigma_{zx} \\ \Sigma_{xz} & \Sigma_{xx} \end{pmatrix} \begin{pmatrix} B_z \\ B_x \end{pmatrix} = \begin{pmatrix} \Sigma_{zy} \\ \Sigma_{xy} \end{pmatrix}.$$

Now, note that the regressions are equal, $B_{y.zx} = B^{**}_{y.zx}$, and that

$$B^{**}_{y.zx} = \Sigma^{**-1}_{zx,zx}\Sigma^{**}_{zx,y} = \begin{pmatrix} \Sigma^{**}_{zz} & \Sigma^{**}_{zx} \\ \Sigma^{**}_{xz} & \Sigma^{**}_{xx} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma^{**}_{zy} \\ \Sigma^{**}_{xy} \end{pmatrix} = \begin{pmatrix} \mathrm{B}^{**}_z \\ \mathrm{B}^{**}_x \end{pmatrix}.$$

Because $B^{**}_{y.zx}$ is based on the n^{**} cases, it is directly observable. Substituting into Equation 6 with the observed regression, we find that

$$\Sigma_{zy} = \Sigma_{zz}B^{**}_z + \Sigma_{zx}B^{**}_x$$

and

$$\Sigma_{xy} = \Sigma_{xz}B^{**}_z + \Sigma_{xx}B^{**}_x . \tag{7}$$

We next solve for the last remaining unsolved matrix Σ_{yy}. Because the conditional variances are not affected by direct selection, $\Sigma^{**}_{yy.zx} = \Sigma_{yy.zx}$. From the definition of the conditional variances, it follows that

$$\Sigma_{yy} - B'_{y.zx}\Sigma_{zx,zx}B_{y.zx} = \Sigma^{**}_{yy} - B^{**\prime}_{y.zx}\Sigma^{**}_{zx,zx}B^{**}_{y.zx} .$$

Solving for the variance of **Y** in the unrestricted space, we obtain

$$\Sigma_{yy} = \Sigma^{**}_{yy} - B^{**\prime}_{y.zx}\Sigma^{**}_{zx,zx}B^{**}_{y.zx} + B'_{y.zx}\Sigma_{zx,zx}B_{y.zx} .$$

But the slopes are equal under direct selection, so

$$\Sigma_{yy} = \Sigma^{**}_{yy} + B^{**\prime}_{y.zx}(\Sigma_{zx.zx} - \Sigma^{**}_{zx,zx})B^{**}_{y.zx} . \tag{8}$$

Note that in the multiple-hurdle situation, the equations must be solved sequentially. First, we solve Equation 5. Then, using the results from Equation 5, we solve Equation 7. After obtaining the results from Equation 7, we solve Equation 8. Putting all of the results together, we get Σ, the unrestricted population.

We focus next on the relationship between the unrestricted and restricted regressions of **Y** on **X**. The unrestricted regression of **Y** on **X** is by definition

$$B_{y.x} = \Sigma_{xx}^{-1} \Sigma_{xy} \ .$$

If we rewrite the regression using Equations 5 and 7, we can see that in the multiple-hurdle situation, the regression of **Y** on **X** depends on the relation between **X** and **Z**, the **Z** selection ratio, and the restricted variance of **X**,

$$\begin{aligned} B_{y.x} &= (\Sigma_{xx}^{*} + \Sigma_{xz}^{*} \Phi^{*} \Sigma_{zx}^{*})^{-1} (\Sigma_{xz} B_{z}^{**} + \Sigma_{xx} B_{x}^{**}) \\ &= (\Sigma_{xx}^{*} + \Sigma_{xz}^{*} \Phi^{*} \Sigma_{zx}^{*})^{-1} (\Sigma_{xz} \Sigma_{xx}) \begin{pmatrix} B_{z}^{**} \\ B_{x}^{**} \end{pmatrix} \\ &= B_{z.x} B_{z}^{**} + B_{x}^{**} . \end{aligned} \tag{9}$$

Furthermore, notice that when the covariance between **X** and **Z** is zero in the unrestricted population, the covariance between **X** and **Z** is also zero in the restricted population. (If **Z** is uncorrelated with **X**, then it follows from Equation 5 that $\Sigma_{zx} = \Theta^{*} \Sigma_{zx}^{*} = 0$ and $\Sigma_{zx}^{*} = 0$). When the covariance between **X** and **Z** is zero, it follows from Equation 9 that unrestricted and restricted regressions are equal,

$$B_{y.x} = \Sigma_{xx}^{-1} (0 + \Sigma_{xx} \Sigma_{xx}^{**-1} \Sigma_{xy}^{**}) = \Sigma_{xx}^{**-1} \Sigma_{xy}^{**} = B_{y.x}^{**} \ .$$

In situations in which the covariance is not zero, the unrestricted and restricted regressions are not equal.

Three-Hurdle Situation and Beyond

The three-hurdle selection situation involves just a minor generalization of the two-hurdle situation. Consider the situation in which we first select with **Z**, then **X**, and then **W**. The **Y** is observed after the last hurdle. Using our previous result, we can see that the first set of equations is

$$\begin{aligned} \Sigma_{zx} &= \Theta^{(1)} \Sigma_{zx}^{(1)} \\ \Sigma_{xx} &= \Sigma_{xx}^{(1)} + B_{x.z}^{(1)\prime} (\Sigma_{zz} - \Sigma_{zz}^{(1)}) B_{x.z}^{(1)} . \end{aligned} \tag{10}$$

To simplify the notation, we are using a number between the parentheses to take place of the asterisks. Our second set of equations yields

$$\begin{aligned} \Sigma_{zw} &= \Sigma_{zz} B_{z}^{(2)} + \Sigma_{zx} B_{x}^{(2)} \\ \Sigma_{xw} &= \Sigma_{xz} B_{z}^{(2)} + \Sigma_{xx} B_{x}^{(2)} \\ \Sigma_{ww} &= \Sigma_{ww}^{(2)} + B_{w.zs}^{(2)} (\Sigma_{zx,zx} - \Sigma_{zx,zx}^{(2)}) B_{w.zx}^{(2)} , \end{aligned} \tag{11}$$

where

$$B^{(2)}_{w.zx} = (\Sigma^{(2)}_{zx,zx})^{-1}\,\Sigma^{(2)}_{zx,w} = \begin{pmatrix} \Sigma^{(2)}_{zz} & \Sigma^{(2)}_{zx} \\ \Sigma^{(2)}_{xz} & \Sigma^{(2)}_{xx} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma^{(2)}_{zw} \\ \Sigma^{(2)}_{xw} \end{pmatrix} = \begin{pmatrix} B^{(2)}_{z} \\ B^{(2)}_{x} \end{pmatrix}.$$

Similarly, we obtain the last set of equations,

$$\begin{aligned} \Sigma_{zy} &= \Sigma_{zz}B^{(3)}_{z} + \Sigma_{zx}B^{(3)}_{x} + \Sigma_{zw}B^{(3)}_{w} \\ \Sigma_{xy} &= \Sigma_{xz}B^{(3)}_{z} + \Sigma_{xx}B^{(3)}_{x} + \Sigma_{xw}B^{(3)}_{w} \\ \Sigma_{wy} &= \Sigma_{wz}B^{(3)}_{z} + \Sigma_{wx}B^{(3)}_{x} + \Sigma_{ww}B^{(3)}_{w} \\ \Sigma_{yy} &= \Sigma^{3}_{yy} + B^{(3)\prime}_{y.zxw}(\Sigma_{zxw,zxw} - \Sigma^{(3)}_{zxw,zxw})B^{(3)}_{y.zxw}, \end{aligned} \tag{12}$$

where

$$B^{(3)}_{y.zxw} = (\Sigma^{(3)}_{zxw,zxw})^{-1}\,\Sigma^{(3)}_{zxw,y} = \begin{pmatrix} \Sigma^{(3)}_{zz} & \Sigma^{(3)}_{zx} & \Sigma^{(3)}_{zw} \\ \Sigma^{(3)}_{xz} & \Sigma^{(3)}_{xx} & \Sigma^{(3)}_{xw} \\ \Sigma^{(3)}_{wz} & \Sigma^{(3)}_{wx} & \Sigma^{(3)}_{ww} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma^{(3)}_{zy} \\ \Sigma^{(3)}_{xy} \\ \Sigma^{(3)}_{wy} \end{pmatrix} = \begin{pmatrix} \mathrm{B}^{(3)}_{z} \\ \mathrm{B}^{(3)}_{x} \\ \mathrm{B}^{(3)}_{w} \end{pmatrix}.$$

Again, the equations must be solved sequentially. We solve Equation 10 first. Then, using the results from Equation 10, we solve Equation 11. After obtaining the results, we solve Equation 12. Generalizations to more complicated designs are possible by following the logic described. Before discussing parameter estimation, we consider Figure 3, a variant of Figure 1.

The Modified Single-Hurdle Selection: Predictive Validation

Figure 3 describes a selection situation in which **Z** and **X** are collected but only **Z** is used for selection. The criterion **Y** is observed only on the individuals selected. This design is often used to establish predictive validity. In this design, the elements of the variance-covariance matrix between **Z** and **X** are directly available, $\Sigma_{zx,zx}$. So our task is much simpler. The only matrices that are not observable are Σ_{zy}, Σ_{xy}, and Σ_{yy}. These matrices are obtained as follows:

$$\begin{aligned} \Sigma_{zy} &= \Theta^{*}\Sigma^{*}_{2y} \\ \Sigma_{xy} &= \Sigma_{xz}B^{*}_{z} + \Sigma_{xx}B^{*}_{x}\text{ , and} \\ \Sigma_{yy} &= \Sigma^{*}_{yy} + B^{*\prime}_{y.zx}(\Sigma_{zx,zx} - \Sigma^{*}_{zx,zx})B^{*}_{y.zx}. \end{aligned} \tag{13}$$

Following the previous argument, we can show that the only biased regression in Figure 3 is the regression of **Y** on **X**. As in the concurrent validity situation, the regression of $B^{*}_{y.x}$ is biased. In terms of bias for the regression of **Y** on **X**, the predictive and concurrent designs are the same. One advantage of this design, however, is that we can directly obtain an unbiased estimate of the correlation between **Z** and **X**.

Estimation

The statistical equations presented above helped us with the understanding of biases in validation designs. However, the equations are only partially helpful when it comes to estimation because the equations must be modified every time that we add or delete a hurdle. More general procedures based on computer iterations exist that do not require such modification. We discuss two of these procedures: the estimation maximization (EM) computer algorithm of Dempster et al. (1977) and the multiple imputation procedure (MI) of Rubin (1978).

The equations (formulas) given in the previous section and the EM procedure yield maximum likelihood (ML) estimators. The use of the equations for estimation can be justified by a factorization of the likelihood under a monotonic missing data structure (e.g., see Little & Rubin, 1987; Mendoza, 1993). Because both EM and the formulas (under a monotonic data structure) yield ML estimators, they give very similar results. The EM, however, is a computer algorithm procedure that yields ML estimators under a variety of missing data structures and is not limited to the monotonic missing data structure (see McLachlan & Krishnan, 1997). The EM algorithm works iteratively by estimating the conditional distribution of the missing values, conditional on a current estimate of the parameter and the observed values. The EM and MI procedures are similar in that both basically rely on "filling in" for the missing values to obtain the appropriate estimators of the unrestricted parameters. The MI procedure, however, approaches the filling-in problem from a Bayesian perspective (Rubin, 1987).

Filling in for the missing data requires, however, that we meet the MAR assumption (Rubin, 1976). Under the MAR assumption, the missing data mechanism can be ignored in estimating parameters; the estimation procedure does not need to explicitly account for the distribution of the missing data. The MAR assumption is likely to be tenable in many selection situations because missingness is under the control of the investigator. Although the MAR assumption may not be a problem, the sample size could be. In a multiple-hurdle with two or more hurdles, the initial sample size must be large or the selection ratios must not be too small to avoid having a small sample at the criterion level. Next, we discuss the MAR as it applies to selection.

Consider the data matrix **W** for a simple validation design containing observed and missing values (not observable), $\mathbf{W} = (\mathbf{X}_{\mathbf{obs}}\mathbf{Y}_{\mathbf{obs}}\mathbf{Y}_{\mathbf{mis}})$. The predictor data are observed, and the criterion data are available only for those who were selected. The MAR assumption requires that the probability of missing be independent of the missing data, $\mathbf{Y}_{\mathbf{mis}}$, but depend on the observed data, $\mathbf{X}_{\mathbf{obs}}$. In other words, the fact that an observation on **Y** is missing does not depend on the value of **Y** but on the value of **X**. The basic idea here is that **X** is sufficient to predict whether **Y** is missing. This is, of course, what happens in many selection situations: We use a test score to decide whether to select an individual. If they are selected, the criterion is observed; if they are not selected, the criterion is missing. Note that the MAR assumption implies that the conditional distribution of **Y** given **X** is the same whether **Y** is missing or observed (Rubin, 1976). According to Schafer and Schenker (2000), the observed data provide no information to support or contradict the MAR assumption. The evidence to support the assumption must be found from sources outside the observed data. If the MAR assumption is met and the observed variables are related to the criterion, issues of sample size aside, the EM, MI, and the formulas should provide accurate corrections in many selection situations. The reader is referred to Little and Schenker (1995), Schafer and Graham (2002), or Allison (2001) for an excellent discussion of the MAR assumption and other topics dealing with missing data.

Next, consider the multiple-hurdle design with variables **Z**, **X**, and **Y** and assume that **Z** and then **X** are used in the selection of individuals. Here, **Z** is sufficient to predict whether **X** is missing, but **Z** and **X** are needed to predict whether **Y** is missing. Notice that **X** by itself is not sufficient for the presence or absence of **Y**. The independence-from-missing-data assumption embedded in MAR requires that we have data on all of the variables that are related to selection. For example, suppose that unbeknownst to us another variable besides **Z** and **X** was used occasionally to make hiring decisions. Because selection here is a function of **Z**, **X**, and the undisclosed variable, MAR would not be met if you take into consideration only **Z** and **X**. Any correction not taking into account the undisclosed variable would be biased, with the degree of bias depending on the variable's effect on selection and its correlation with **Z** and **X**. Of course, if the variable were highly correlated with **Z** and **X**, then the bias would be small. Similarly, if self-selection bias were to keep high performers from applying, the missingness would be a function of the unobserved performance, and MAR would not be met. If MAR is not met, the missing data are said to be MNAR, and other procedures must be used. This situation is also referred to as a nonignorable.

MAR is not an all-or-none condition, and different degrees of MAR can exist for a variable or variables in a data set. Collins, Schafer, and Kam (2001) showed that the ML and MI estimates are robust to mild violations of the MAR assumption. They also showed that inclusive designs (those that collect auxiliary variables, variables that are likely to be related to the missingness) are more robust than restrictive designs (those that do not collect these variables) to MAR violations.

Procedures that are appropriate for nonignorable (MAR is not met) situations (the missing data are a function of missing information) have been given by Heckman (1976), Amemiya (1984), Little (1995), and, in the context of range restriction, Gross and McGanney (1987). However, these procedures are sometimes difficult to implement and often are very sensitive to assumptions (Allison, 2001). Because it is our position that with a bit of planning and careful data collection, most selection situations can assume MAR, we focus on procedures that are appropriate under MAR. However, before leaving this subject, we make a point about attrition.

If data are missing because of selection and attrition, attrition must be taken into consideration. To meet the MAR assumption, we must have data on the selection variables and the variables related to attrition. If attrition is related to performance (and the performance measure is taken before attrition takes place and again later after attrition takes place), then one could incorporate attrition as an additional hurdle in the design. In this situation, we are likely to meet MAR because we have information that is related to attrition. When attrition is not related to the observed data, then the MAR assumption is not likely to be met unless attrition is truly a random phenomenon. We can ignore attrition only when it is truly random; that is, attrition is unrelated to both the observed and unobserved data. In this situation, the missing data (due to attrition) are MCAR. This is the only situation in which we can ignore attrition (see Little & Rubin, 1987; Switzer, Roth, & Switzer, 1998, for a discussion of MCAR). In validation designs in which attrition can be a factor, the researcher will be well served by collecting auxiliary variables that are likely to be related to attrition.

Multiple Imputation Procedures

As we discussed earlier, the MAR condition is important because under MAR, the observed data likelihood provides the appropriate likelihood function for estimating the parameters of the (unrestricted) population. Consequently, inferences using ML or MI procedures are valid when the missing data are MAR. Although it is generally easy to obtain ML estimates using the EM algorithm (or any other ML algorithm), the standard errors are generally not easy to obtain. Although the stochastic EM algorithm (Meng & Rubin, 1991), the bootstrap (McLachlan & Krishnan, 1997), and others (Schafer & Schenker, 2000) have appeared in the literature for obtaining standard errors of ML estimates, they are not always available and easy to implement. At this point, the MI procedure appears to be the more flexible and easier to implement in general situations in which we desire standard errors, confidence intervals, and tests of hypotheses.

Rubin (1978) formally introduced the concept of multiple imputations. In multiple imputations, each missing value in the data set is replaced by a set of *m* (> 1) values drawn from the posterior predictive distribution of $\mathbf{Y}_{\mathbf{mis}}$ (see Rubin, 1987; Schafer & Olsen, 1998). The MI procedure yields *m* imputed (complete) data sets. The imputed data in the MI procedures, by the way, reflect both sampling uncertainty and model uncertainty. Because the data sets have no missing observations, they can be analyzed with standard statistical procedures. After analyzing each of the data sets, the results are combined in one overall analysis using the combination rules given by Rubin (1987). Rubin showed that the combined estimates follow approximately the *t* distribution. In addition, Li, Raghunathan, and Rubin (1991) have given an *F* approximation to be used in multiparameter inferences, allowing for the test of the null hypothesis that all of the parameters are zero.

Many multiple imputation procedures have been proposed in the literature. Basically, these procedures can be described as being Bayesian or non-Bayesian. The Bayesian procedures fill in the missing values from the conditional distribution

$$\mathbf{P}(\mathbf{Y}_{\mathbf{mis}} \mid \mathbf{Y}_{\mathbf{obs}}, \psi).$$

However, in Bayesian terms, the parameter ψ is a random variable with a posterior distribution, and $\mathbf{Y}_{\mathbf{obs}}$ refers to observed data in the entire data set. Thus, before we can fill in any missing values, we must first formulate the posterior distribution of ψ. The fill-in process involves sampling first from the posterior distribution of ψ to obtain the realization ψ_o, then to sample from

$$\mathbf{P}(\mathbf{Y}_{\mathbf{mis}} \mid \mathbf{Y}_{\mathbf{obs}}, \psi_o)$$

to fill in the missing values. The process is repeated *m* times to obtain *m* values for each missing observation.

The non-Bayesian alternatives do not rely on the predictive posterior distribution to model the uncertainty of the missing data. Rao and Shao (1992) gave a procedure to estimate the mean and variance of variables with missing data (under MAR conditions) using a modified jackknife procedure. In addition, Efron (1994) also has provided a number of nonparametric procedures to handle missing data. This procedure can be applied to parametric and nonparametric situations, but it is computationally demanding. We will not discuss the non-Bayesian procedures further.

The MI procedure must incorporate the appropriate variability among the *m* sets of imputations to be “proper” (Rubin, 1987). Imputation procedures that are not proper yield bias estimators (see Rubin, 1996, for a technical discussion of proper imputation procedures). However, Schafer (1997) has demonstrated that the MI procedures are robust to modest departures from the imputation model. The MI procedure is also robust to the normality assumption (Graham & Schafer, 1999). Nevertheless, care should be taken in selecting the appropriate imputation model and ensuring that the assumptions are met. Also, care should be taken to include in the imputation model any association that may prove important to subsequent analyses.

Recently, MI implementation has been made easier and more effective by the work of Schafer and Olsen (1998). They presented a data augmentation (DA) technique based on the work of Tanner and Wong (1987). DA is an MI procedure that uses a Monte Carlo approach to simulate the Bayesian posterior distribution. The procedure alternates between the random imputation of missing data under assumed values of the parameters and draws from the Bayesian posterior distribution of the parameters. Once DA has converged, it can be used to obtain the *m* multiple imputations needed. Schafer and Olsen (1998) have presented three computer programs in S-Plus NORM, CAT, and MIX for the implementation of this DA procedure. NORM is also available as a stand-alone for Windows. The computer programs are easy to use and can be obtained from Schafer’s Web site. NORM is used for multivariate normal data, MIX is used for mixed models, and CAT is used for loglinear models. Similar procedures are now available in SAS. SAS 8.2 and later versions contain two missing data analytic procedures (Proc MI and Proc MIANALYZE). Unlike NORM, Proc MI can set maximum and minimum values for the imputed data. In addition, Proc MI allows for rounding off the missing values (for a more complete list of available software, see Allison, 2001; Horton & Lipsitz, 2001; Schafer & Graham, 2002).

NORM can be used to generate *m* imputed (complete) data sets from a data set with missing data. The *m* data sets are then each analyzed using standard statistical procedures. The results are then averaged (combined) over the *m* data sets to estimate the parameter and construct a test of the hypothesis. The number of data sets *m* in most situations does not have to be large. Schafer (1997) recommends three to five data sets per estimator. So if we were interested in estimating two parameters, we would set $m =$ 6-10. Once the *m* data sets are generated, NORM can be used again to combine the results. To check the stability of the results, one can replicate the entire imputation process with a larger *m* and see whether there are qualitative differences between the two runs. Because in most situations computer time is not an issue and because the larger the *m* the more stable the results are likely to be, the researcher should not be too conservative with the size of *m*.

An Example

We illustrate the procedures next using simulated data. The advantage of using simulated data is that we know the population parameters. Thus, the results are easy to evaluate. For analyzing the data, we wrote two SAS macros (Macro A and Macro B) that create files to be used in NORM and in SAS. Both the MI and EM procedures are available through NORM version 2.03. Schafer's NORM program was used to estimate the population variance-covariance matrix; with it, we obtained an ML (EM) and an MI estimate of this matrix.

Macro A was used to perform the regression analyses, and Macro B was used to perform the correlation analyses. Both macros are available from the authors. The correlation macro has a procedure created within SAS PROC IML to estimate the large-sample variance-covariance matrix of the correlations. The procedure was based on the work of Olkin and Finn (1995).

A random sample of 300 observations was obtained from a multivariate normal distribution with variables **Z**, **X**, and **Y**. The population means were 100, the variances were 30, and the covariances were each 15 (and each correlation is .5). After the random sample size of 300 was obtained, the data were rank ordered along the **Z** variable, and the top 200 observations were selected. Next, we rank ordered the remaining data set along the **X** variable, and the top 100 observations were selected. The final sample had 300 **Z** scores, 200 **X** scores, and 100 **Y** scores, simulating a multiple-hurdle design (Figure 2 depicts the shape of this sample). Because Figures 1 and 3 are special cases of Figure 2, the example followed Figure 2. The analysis of either of the other two is very similar.

Table 1 gives the EM, MI, and formula-corrected estimates of the population variance-covariance matrix as well as estimates obtained from the listwise-deletion analysis ($n^{**} = 100$). The population values are also given for comparisons. Two sets of estimates were computed for the MI procedure, one with 10 and one with 20 imputations, but only the results for the 20 imputations are shown in Table 1. Because NORM estimated the fraction of missing data to be high, .57, we increased the number of imputations from 10 to 20 and reran the analysis. By increasing the number of imputations to 20, one can increase the efficiency of the MI estimation (see Schafer & Olsen, 1998). As a note of caution, NORM may yield different estimates of the fraction of missing information depending on the number of imputations, so it is important that the initial number of imputations not be too small. Otherwise, NORM seems to be very precise and useful.

We see from Table 1 that the listwise-deletion (uncorrected) analysis underestimated the population variances and covariances. On the other hand, the EM and the formula estimates were much closer to the population parameters but not as accurate as the MI estimates when the number of imputations was 20. Next, the corrected variance-covariance matrices were submitted to SAS Proc Reg to obtain the regressions and multiple correlations. These are given in Tables 2 and 3.

Table 2 gives the regression coefficients for five regression analyses (listwise, EM, both MIs, and formulas) when (a) **Y** was regressed on **Z**, (b) **Y** was regressed on **X**, and (c) **Y** was regressed on (**Z**, **X**). Conditions a and b are biased. We can see from Table 2 that in the unbiased condition, as expected, the EM, MI, and formula estimates were not very different from the listwise regression analysis. In the bias conditions, however, the EM, MI, and formula estimates performed much better than did the listwise deletion. Again, the MI with $m = 20$ yielded the most accurate estimates.

Not given in Table 2 are the MI standard errors for the regression of **Y** on (**Z**, **X**). Using Macro A, we found the MI standard errors, when $m = 20$, to be .15 for both the **Z**

Table 1
Estimates of the Variance-Covariance Matrix Under Population, Listwise Deletion, Estimation Maximization (EM), Multiple Imputation (MI), and Formulas

Method		Z	X	Y
Population	**Z**	30.00	—	—
	X	15.00	30.00	—
	Y	15.00	15.00	30.00
Listwise deletion	**Z**	15.57	—	—
	X	4.80	10.30	—
	Y	4.76	4.26	26.50
EM	**Z**	29.89	—	—
	X	16.41	27.50	—
	Y	11.41	12.11	30.12
MI[a]	**Z**	31.72	—	—
	X	16.58	29.88	—
	Y	16.05	16.62	32.35
Formulas	**Z**	29.98	—	—
	X	16.47	27.72	—
	Y	11.47	12.21	30.92

a. Number of imputations = 20.

and **X** regression coefficients. When we tested the null hypothesis that the two population regression coefficients were zero (multiple correlation is zero), we reject it with $F(2, 53) = 8.45$, $p < .01$.

Table 3 is analogous to Table 2 but gives the squared correlations instead of the regression coefficients. We see from Table 3 that the listwise-deletion analysis underestimated the population multiple R^2. The EM and MI also underestimated the R^2, but by a lesser amount. The underestimation is present also in the squared correlation between **Y** and **Z** and between **Y** and **X**. Both the EM and MI (with 20 imputations) gave better estimates than did the listwise-deletion analysis.

Next, to demonstrate the use of the procedures when they involve correlations, we tested the null hypothesis that all of the correlations (ρ_{zx}, ρ_{zy}, ρ_{xy}) were zero in population. Because this test involves a simultaneous inference, we obtained the variance-covariance matrix of the correlations for each of the imputed data sets. We obtained this matrix by implementing the procedure given by Olkin and Finn (1995) in Macro B, and the test was executed in NORM using the multiparameter inference option. As expected, the null hypothesis of no correlations was rejected, yielding $F(3, 86) = 17.15$, $p < .01$. The corrected correlation and the standard errors were $r_{zx} = .57$ (.09), $r_{zy} = .40$ (.11), and $r_{xy} = .45$ (.10), not too far from their population value of .50. (Clearly, these correlations correspond to those given in Table 3. However, the advantage of this analysis is that we also get the standard errors.)

Next, to test for H_o: $\rho_{zy} - \rho_{xy} = 0$ (in many designs, this would be the hypothesis that the new test is better than the old test), we created a file (with Macro B) containing the variance of the difference between the correlations and submitted it to NORM. As expected (recall that both correlations are .5 in the population), the null hypothesis was not rejected. The t ratio was –.50 with $p = .62$.

Table 2
Estimates of the Regression Coefficients Under Population, Listwise Deletion, Estimation Maximization (EM), Multiple Imputation (MI), and Formulas

	Unbiased		*Biased*	
Method	**Z**[a]	**X**[a]	**Z**[b]	**X**[c]
Population	.33	.33	.50	.50
Listwise deletion[d]	.21	.32	.31	.41
EM	.21	.32	.38	.44
MI [e]	.16	.24	.28	.33
MI [f]	.22	.35	.42	.48
Formulas	.21	.32	.38	.44

a. **Y** on **Z** and **X**.
b. **Y** on **Z** only.
c. **Y** on **X** only.
d. *N* = 100.
e. Number of imputations = 10.
f. Number of imputations = 20.

Table 3
Biased Squared Correlations Under Population, Listwise Deletion, Estimation Maximization (EM), Multiple Imputation (MI), and Formulas

Method	*Multiple* R^2 [a]	r^2 [b]	r^2 [c]
Population	.33	.25	.25
Listwise deletion[d]	.09	.06	.07
EM	.21	.15	.18
MI [e]	.14	.09	.12
MI [f]	.25	.17	.21
Formulas	.20	.14	.17

a. **Y** on **Z** and **X**.
b. **Y** on **Z**.
c. **Y** on **X**.
d. *N* = 100.
e. Number of imputations = 10.
f. Number of imputations = 20.

It should be pointed out that any of these correlations also could be corrected for attenuation if an estimate of reliability is available. Also, if the standard error of this reliability coefficient is available (see Fan & Thompson, 2001) a confidence interval on the doubly corrected correlation can be found following Mendoza, Stafford, and Stauffer (2000). For example, suppose that the alpha reliability of **X** is estimated outside of the validation study to be .90. Further assume that by applying the procedure illustrated in Fan and Thompson (2001), you find the lower bound and upper bound of the reliability for a 95% confidence interval to be .70 and .97, respectively. Then, applying the Bonferroni technique given in Mendoza et al. (2000), the 90% lower bound for the doubly corrected correlation is

$$\rho l = \frac{.449 - 1.96(.101)}{\sqrt{.70}} = .30,$$

and the upper bound is

$$\rho u = \frac{.449 + 1.96(.101)}{\sqrt{.97}} = .657\,.$$

Discussion

Before turning to the broader conclusions flowing from the present study, certain limitations should be noted. To begin, our example focused on only one multiple-hurdle design; other validation designs are clearly possible. However, the multiple-hurdle design used is rather representative, and we hope it provides a relatively reasonable test of the proposed procedures. Also, it should be recognized that the estimates produced by the multiple imputation procedure were examined under only two conditions in which 10 and 20 imputed data sets were examined. With the use of 20 imputations, the MI procedure produced estimates that were similar to the equation estimates and provided an accurate representation of population parameters. Nonetheless, it is also true that use of a larger number of imputations, up to some asymptotic level, would likely have provided more stable results.

With these caveats in mind, we believe the results obtained in the present effort have some noteworthy implications. Perhaps the most important implication pertains to the assumptions we make in framing the problem of range restriction. Earlier, we argued that range restriction can be viewed as a specific instance of the more general problem of estimating full-sample variance-covariance matrices under conditions in which data are missing. Using this proposition, we have suggested a general approach for correcting for range restriction that encompasses many of the traditional correction procedures as special cases. Moreover, application of this approach yielded corrections for correlation coefficients and regression weights, along with sampling errors, for designs that had proven intractable using more traditional approaches.

The EM and MI procedures, as well as the formulas, all produced far more accurate estimates of population regression weights and correlations than did listwise-deletion analyses. Of course, relative to uncorrected estimates, listwise-deletion analyses evidenced the expected underestimation of test-criterion relations. In this regard, however, it should be noted that although the EM and MI procedures, as well as the formulas, did not display the same level of gross underestimation as listwise-deletion analyses did, they were somewhat conservative estimators of population parameters. Preliminary work by Chasteen and Mendoza (2003) in the context of a multiple-hurdle design also has shown that the EM and MI estimators are generally conservative as long as the sample size is not too small ($n^{**} = 30$) or all of the correlations are zero. Although these conditions are not very likely in practice, R. L. Linn, Harnisch, and Dunbar's (1981) recommendation to routinely report both observed and corrected correlations makes sense. More research is needed to fully understand the strengths and weaknesses of these procedures in the context of validation designs.

Although the derivation of the equations is of interest in its own right, the importance of the equations lies in the fact that they provide a general approach to assess the impact of selection at the population level, pointing out that selection is not a problem that can be solved by simply increasing sample size. Although the equations can be used for estimation, the estimation of the parameters is best implemented using a multiple imputation procedure that provides estimates of the standard errors.

Application of this approach can improve the accuracy of corrected validity coefficients, especially if an extensive set of variables is collected in validation studies. The variables collected should be able to model both the sources of missing data and potential influences on performance. Although some cost is entailed in such extended data collection efforts, it is also the case that these additional variables provide a more comprehensive basis for appraising the validity of the resulting selection system and improve the accuracy of the estimation. Application of this strategy, moreover, will result in the integration of corrections and performance theories. This integration of correction estimates with substantive theory was recommended by James et al. (1988) and should serve to provide requisite evidence for the substantive meaningfulness of corrected estimates.

It appears from the results obtained that it is possible to find a general solution that allows for estimation, confidence intervals, and hypothesis testing of corrected coefficients in many range-restriction problems through the application of missing data procedures. These procedures have the added advantage that they are a reasonably robust to moderate departure from normality. Furthermore, procedures that do not require the normality assumption are increasingly becoming more available. These are the only procedures, given that appropriate measures are collected, that are capable of addressing the multitude of factors that generate range restriction in validation designs, including recruiting, self-selection bias, and attrition. Although it is not yet clear what measures (auxiliary variables) should be collected—more research must be done before we can formulate an answer—we have a set of procedures with a theoretical framework that gives us the flexibility needed for the estimation of complex range-restriction effects. Given the advantageous characteristics of the missing data approach, we believe that missing data models can provide a coherent framework for effectively addressing range-restriction problems in many criterion-related validity studies.

Notes

1. In a predictive validity design in which the new test is used for selection, the design reduces to a two-variable design. In this case, the correlation between the test and criterion is corrected for direct range restriction.

2. Normality is assumed for convenience. For the equations to hold, we need only (a) equal (linear) conditional expectations over the selected and referent populations and (b) constant conditional variances over the selected and referent populations.

References

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.

Amemiya, T. (1984). Tobit models: A survey. *Journal of Econometrics*, *24*, 3-61.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bobko, P. (1995). *Correlation and regression: Principles and applications for industrial/organizational psychology and management*. New York: McGraw-Hill.

Chasteen, C. S., & Mendoza, J. L. (2003, April). *Restriction in range issues in validation designs: Modern tools for old problems*. Paper presented at the national meetings of the American Educational Research Association, Chicago, IL.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods*, *6*, 330-351.

Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, *89*, 463-475.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *39*(series B), 1-38.

Fan, X., & Thompson, B. (2001). Confidence intervals about score reliability coefficients, please: An EPM guidelines editorial. *Educational and Psychological Measurement*, *61*, 517-531.

Ghiselli, E. E. (1964). *Theory of psychological measurement*. New York: McGraw-Hill.

Ghiselli, E. E. (1966). *The validity of occupational aptitude tests*. New York: John Wiley.

Graham, J. W., & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 1-29). Thousand Oaks, CA: Sage.

Greener, J. M., & Osburn, H. G. (1979). An empirical study of the accuracy of corrections for restriction in range due to explicit selection. *Applied Psychological Measurements*, *3*, 31-41.

Greener, J. M., & Osburn, H. G. (1980). Accuracy of corrections for restriction in range due to explicit in heteroscedastic and nonlinear distributions. *Educational and Psychological Measurement*, *40*, 337-346.

Gross, A. L. (1990). A maximum likelihood approach to test validation with missing data and censored dependent variables. *Psychometrika*, *55*, 533-549.

Gross, A. L., & McGanney, M. L. (1987). The restriction of range problem and nonignorable selection process. *Journal of Applied Psychology*, *72*, 604-610.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. *Annals of Economics and Social Measurement*, *5*, 475-492.

Hoffman, C. (1995).Applying range restriction corrections using published norms: Three case studies. *Personnel Psychology*, *48*, 913-923.

Horton, N. J., & Lipsitz, S. R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *American Statistician*, *55*, 244-254.

James, L. R., Demaree, R., Muliak, S. A., & Mumford, M. D. (1988). Validity generalization: A rejoinder to Schmidt, Hunter, and Raju. *Journal of Applied Psychology*, *73*, 443-452.

Kolmstetter, E. (2003). I-Os making an impact: TSA transportation security screener skill standards, selection system, and hiring process. *Industrial-Organizational Psychologist*, *40*(4), 39-46.

Lawley, D. N. (1943). A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society of Edinburgh*, *62*, 28-30.

Li, K. H., Raghunathan, T. E., & Rubin, D. B. (1991). Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, *86*, 1065-1073.

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

Linn, R. T. (1968). Range restriction problems in the use of self-selected groups for test validation. *Psychological Bulletin*, *69*, 69-73.

Little, R. J. A. (1995). Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association*, *90*, 1112-1121.

Little, R. J. A., & Rubin, D. B (1987). *Statistical analysis with missing data*. New York: John Wiley.

Little, R. J. A., & Schenker, N. (1995). Missing data. In Arminger, G., Clogg, C. C., & Sobel, M. E. (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 39-75). New York: Plenum.

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

McLachlan, G. J., & Krishnan, T. (1997). *The EM algorithm and extensions*. New York: John Wiley.

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

Mendoza, J. L., & Mumford, M. D. (1987). Corrections for attenuation and range restriction. *Journal of Educational Statistics*, *12*, 282-293.

Mendoza, J. L., Stafford, K. L., & Stauffer, J. M. (2000). Large-sample confidence intervals for validity and reliability coefficients. *Psychological Methods*, *5*, 356-369.

Meng, X. L., & Rubin, D. B. (1991). Using the EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, *86*, 899-909.

Messick, S. J. (1998). Alternative models of assessment, uniform standards of validity. In M. D. Hakel (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection* (pp. 59-74). Mahwah, NJ: Lawrence Erlbaum.

Olkin, I., & Finn, J. D. (1995). Correlation redux. *Psychological Bulletin*, *118*, 155-164.

Ones, D. S., & Viswesvaran, C. (2003). Job-specific applicant pools and national norms for personality scales implications for range-restriction corrections in validation research. *Journal of Applied Psychology*, *88*, 570-577.

Pearson, K. (1903). Mathematical contributions to the theory of evolution XI: On this influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society, London, Series A*, *200*, 1-66.

Rao, J. N. K., & Shao, J. (1992). Jackknife variance estimation with survey data under hot-deck imputation. *Biometrika*, *79*, 811-822.

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 selection formulas. *Journal of Applied Psychology*, *79*, 298-301.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581-592.

Rubin, D. B. (1978). Multiple imputations in sample surveys—A phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section American Statistical Association*, 20-34.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*, 473-489.

Sackett, P. R., Laczo, R. M., & Arvey, R. D. (2002). The effects of range restriction on estimates of criterion interrater reliability: Implications for validation research. *Personnel Psychology*, *55*, 807-825.

Sackett, P. R., & Ostgard, D. L. (1994). Job-specific applicant pools and national norms for cognitive ability tests: Implications for range restriction corrections in validation research. *Journal of Applied Psychology*, *79*, 680-684.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, *85*, 112-118.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147-177.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, *33*, 545-571.

Schafer, J. L., & Schenker, N. (2000). Inferences with imputed conditional means. *Journal of the American Statistical Association*, *95*, 144-154.

Switzer, F. S., Roth, P. L., & Switzer, D.M. (1998). Systematic data loss in HRM settings: A Monte Carlo analysis. *Journal of Management*, *24*, 763-777.

Tanner, M. A., & Wong, W.H. (1987). The calculation of the posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528-550.

Jorge L. Mendoza is a professor in the psychology department at the University of Oklahoma.

David E. Bard is a graduate student in the psychology department at the University of Oklahoma.

Michael D. Mumford is a professor in the psychology department at the University of Oklahoma.

Siew C. Ang is a graduate student in the psychology department at the University of Oklahoma.