5-2010

Population Cross-Validity Estimation and Adjustment for Direct Range Restriction: A Monte Carlo Investigation of Procedural Sequences to Achieve Optimal Cross-Validity

David Matthew Goins
Western Kentucky University, david.goins236@wku.edu

POPULATION CROSS-VALIDITY ESTIMATION AND ADJUSTMENT FOR DIRECT RANGE RESTRICTION: A MONTE CARLO INVESTIGATION OF PROCEDURAL SEQUENCES TO ACHIEVE OPTIMAL CROSS-VALIDITY

A Thesis
Presented to
The Faculty of the Department of Psychology
Western Kentucky University
Bowling Green, Kentucky

In Partial Fulfillment
Of the Requirements for the Degree
Master of Arts

By
David Matthew Goins

May 2010

**POPULATION CROSS-VALIDITY ESTIMATION AND ADJUSTMENT FOR DIRECT RANGE RESTRICTION:
A MONTE CARLO INVESTIGATION OF PROCEDURAL SEQUENCES TO ACHIEVE OPTIMAL CROSS-VALIDITY**

Date Recommended __April 28, 2010_____

__Reagan Brown_____
Director of Thesis

__Elizabeth Shoenfelt _____

__Anthony Paquin_____

_____
Dean, Graduate Studies and Research Date

Table of Contents

POPULATION CROSS-VALIDITY ESTIMATION AND ADJUSTMENT FOR DIRECT RANGE RESTRICTION:
A MONTE CARLO INVESTIGATION OF PROCEDURAL SEQUENCES TO ACHIEVE OPTIMAL CROSS-VALIDITY

David Matthew Goins May 2010 Pages 39

Directed by: Reagan Brown, Elizabeth Shoenfelt, and Tony Paquin

Department of Psychology Western Kentucky University

The current study employs Monte Carlo analyses to evaluate the effectiveness of various statistical procedures for determining specific values of interest within a population of 1,000,000 cases. Specifically, the proper procedures for addressing the opposing effects of direct range restriction and validity overestimation were assessed through a comparison of multiple correlation coefficients derived using various sequences of procedures in randomly drawn samples. A comparison of the average bias associated with these methods indicated that correction for range restriction prior to the application of a validity overestimation adjustment formula yielded the best estimate of population parameters over a number of conditions. Additionally, similar methods were employed to assess the effectiveness of the standard $\Delta R^2 F$-test for determining, based on characteristics of the derivation sample, the comparative superiority of either optimally or unit weighted composites in future samples; this procedure was largely ineffective under the conditions employed in the current study.

POPULATION CROSS VALIDITY ESTIMATION AND ADJUSTMENT FOR DIRECT RANGE RESTRICTION:
A MONTE CARLO INVESTIGATION OF PROCEDURAL SEQUENCES TO ACHIEVE OPTIMAL CROSS-VALIDITY

At some fundamental level, all scientific inquiry, regardless of specific discipline, requires measurement. Lacking reliable methods for quantifying phenomena, scientists become indistinguishable from philosophers, and their hypotheses, regardless of the quality of the underlying logic or theory, are meritoriously indistinguishable from any other arguable notion ever conceived.

The study of human behavior is no exception. Although the instruments used in the field of psychology are rarely as intuitive or reliable as thermometers or rulers, the development and improvement of the procedures used to quantify human attributes has been of primary concern to the discipline since its inception. As these measures have become more sophisticated and established, the frequency with which they are employed in non laboratory settings has likewise increased. On a related note, due to the joint effects of legislation and escalating global competition, modern organizations have become increasingly dependent on systematic, data-driven procedures for use in administrative decision-making. This organizational demand has most frequently and effectively been met by industrial/organizational psychologists.

By applying the analytic techniques developed for the measurement of human behavior to the workplace, industrial/organizational psychologists have developed a number of instruments for the prediction of performance in various organizational settings. Furthermore, the use of advanced statistical procedures has resulted in the ability

to quantify the degree to which predictor performance is associated with future workplace outcomes. Though the aforementioned statistical procedures may be among the most powerful tools that industrial/organizational psychologists possess, they are less than perfect. Furthermore, some debate remains regarding the appropriate application and interpretation of these methods for the validation of selection processes. As such, the procedures used to estimate the validity of these instruments are the primary focus of this review.

Multivariate Prediction

In relation to the current discussion, it is important to note that multiple authors have identified the increasingly complex nature of positions in modern organizations (Guion, 1998; SIOP, 2003). Given this observation, it stands to reason that selection procedures must necessarily become congruently more sophisticated in order to accurately predict performance in the modern workplace. Often this standard is met through the utilization of multidimensional selection processes. To illustrate using a fairly intuitive example, one can assume that, due to a heavy reliance on the relatively complex mechanical systems utilized in contemporary firefighting operations, modern firefighters must be adept in terms of mechanical comprehension. However, because the tasks associated with this job necessarily entail a significant physical component as well, procedures designed to predict performance at this and similar positions must account for individual differences in both attributes. Due to the dissimilarity of these competencies, however, they must be conceptualized and measured as unique constructs. Therefore, selection for this position necessarily entails the combination of distinct predictor variables. Finally, before advancing the discussion, it is important to note that although

many additional attributes are likely relevant to the position described above, due to the didactic nature of the example, consideration of only these two proficiencies is sufficient for the purposes of this review.

Having noted the increasing complexity of the modern workforce and having provided an example to illustrate the necessity of multivariate prediction of performance, it is pertinent to outline the accepted procedures commonly associated with this type of selection scenario and discuss both the advantages and limitations associated with each.

Although predictor variables have been combined with a variety of prediction models, the most common are those associated with an additive linear or compensatory model (Guion, 1998). More explicitly stated, predictor variables are typically assumed to hold individual linear relationships to the criterion. As such, strength in a given proficiency may compensate for weakness in another. To illustrate using the previous example, under a compensatory prediction model, a prospective firefighter possessing weak physical ability but superlative mechanical comprehension will have the same composite score as a second candidate who demonstrates more uniformly average skills. Though selection scenarios exist under which this model is inappropriate, its prevalence in applied settings and relevance to the current proposal justifies its singular consideration here.

Under a compensatory model, a number of practical concerns become apparent. For example, given a selection scenario for which a number of relevant tests are available, in what way should these tests be combined in order to achieve maximally efficient selection decisions in practice? Furthermore, to what extent does an individual test contribute to the overall utility of selection procedures? Finally, in keeping with an

interest in efficiency, at what stage does the inclusion of additional tests reach a point of diminishing returns in terms of increased predictive accuracy? The procedure or statistical tool commonly employed to address these issues in practice is Ordinary Least Squares (OLS) regression analysis. OLS regression operates by first generating a composite of the predictor variables using optimal weights. More explicitly, this type of statistical procedure identifies the weighting scheme which will yield the least possible error of prediction within the sample from which it was derived. In addition to this optimally weighted composite, OLS regression analysis generates an index of the predictive validity of the combined tests in the form of the squared multiple correlation coefficient, R^2. Given these two outcomes, practitioners can quantify the relationship between the criterion and a battery of selection procedures and, as a result, meet the evidentiary standard required for the use of scientifically-based administrative decisions in practice.

However, having stated both the theoretical and applied utility of this statistical tool, it is necessary to acknowledge that a fundamental shortcoming associated with the use of OLS regression for the validation of multivariate selection procedures has long been noted (Mosier, 1951). To elaborate, because predictor composites generated through OLS regression analysis are mathematically tailored to yield the best possible prediction of the criterion given the available data, the validity of the regression equation is specific to said data. When applied to future samples, the weighting scheme initially developed to produce the most accurate possible predictions within the derivation sample will always yield a weaker observed relationship between the predictors and the criterion. This phenomenon is especially problematic when considered within the context of an applied

validation effort. To illustrate, the best available estimate of the validity of a set of predictors to a given criterion when predictor variables are optimally weighted is the squared multiple correlation coefficient which, due to the aforementioned factors, is spuriously inflated. Practitioners are posed with a serious problem when required to determine the predictive utility of their composite in future samples or to the population as a whole. More simply stated, due to the validity reduction (also known as shrinkage) associated with the application of an optimally weighted set of predictors to future samples, the index of the validity provided in the initial validation study must necessarily be an overestimation of the validity of the selection procedure in future applications. If it can be assumed that the merit of any scientific investigation is, to some degree, proportional to the practical utility of its results, the fact that OLS regression analysis systematically generates a nongeneralizable outcome is an argument against its utilization for validation purposes. Because this phenomenon has long been identified, a number of accepted procedures have been proposed to address this issue (e.g., Mitchel & Klimoski, 1986; Mosier, 1951; Raju, Bilgic, Edwards, & Fleer 1999).

Correction for Validity Shrinkage

It is important at this stage to note that a discussion of the relevant terminology is necessary before the established procedures can be adequately reviewed. To begin, the population squared multiple correlation is the hypothetical multiple correlation between predictors and the criterion within a specified population. This value requires knowledge of the population regression weights and is unattainable in practice (Raju, Edwards, & Fleer, 1997). Of greater practical significance is the population cross-validated multiple correlation, which is calculated by applying regression weights derived from a sample to

the population from which it was drawn. This second value provides an index of the efficacy with which regression weights validated in one sample will predict criterion data when applied to future samples. The methods discussed in the following sections are designed to estimate the latter of the two values.

The earliest methods proposed for estimating the population cross-validated multiple correlation are the most intuitive. Specifically, these procedures, commonly referred to as empirical cross-validation techniques, are characterized by the application of regression weights derived from one sample to another sample from the same population. What has been referred to as the classic empirical cross-validation design (Mosier, 1951) involves the application of a regression equation derived through a previous sample to a relevant second sample randomly drawn from the same population. Using this method, the estimate of the population cross-validated multiple correlation is the correlation between predicted and actual criterion scores. In a similar empirical design, Mosier advocated a random division of a single sample into two half-samples. Regression weights were then to be derived from each half and applied to the opposite half. This method, termed double cross-validation, yields two separate estimates of the population cross-validated multiple correlation. As an extension of the previous design, Claudy (1978) proposed a method of population cross validity estimation designed to analyze all of the possible correlations associated with Mosier's double cross validation. Although this method was later shown to have significant bias under various sample conditions (e.g., Drasgow, Dorans, & Tucker, 1979), it was important in that it marked an early attempt to estimate population cross validity through the application of a formula to a multiple correlation coefficient derived from a single sample.

Following this early effort, a number of formula-based procedures designed to estimate the population cross-validated multiple correlation have been proposed (e.g., Drasgow et al., 1979; Murphy, 1984; Raju et al., 1999). In turn, these developments have instigated a prolonged debate regarding the comparative efficacy of said processes. Although the specific operations vary, these formula-based estimates are similar in that they all represent some function of the sample multiple correlation, sample size, and the number of predictors included in the regression equation (Murphy).

The impetus driving both the development of and subsequent debate surrounding the use of these formulas stems, in part, from the noteworthy inefficiency associated with empirical methods. Specifically, because empirical cross-validation relies on a second sample of data for comparison, the process necessarily involves the formulation of regression weights using a sample size less than the total number of available cases. Although fortuitous circumstances in which sufficient data are available for the generation of two sufficiently large samples are possible, the data loss associated with generation of the validation sample may not constitute a prudent investment. Additionally, because the stability of the regression weights is a function of the size of the sample from which they are derived, the generation of a second sample that does not contribute to the stability of said weights constitutes a significant opportunity cost that must necessarily reduce the predictive validity of the regression equation (Murphy, 1984). Stated differently, had the data employed to determine the cross validated multiple correlation been used to derive the regression weights, the cross-validity of the predictors would have been greater (i.e., reduced validity shrinkage). Based on this observation, some authors have suggested that empirical cross-validation procedures are less efficient

than their formula-based counterparts (Mitchell & Klimoski, 1986; Murphy, 1984). It is important to note that although the use of formula-based corrections has been generally endorsed for most multivariate validation scenarios, they are inappropriate when variable selection procedures are employed to identify and remove predictors (Mitchell & Klimoski). To elaborate, Raju et al. (1999) empirically determined that formula based estimates fail to perform sufficiently when variable selection procedures are employed due to the greater validity shrinkage that occurs under these circumstances. To elaborate, because variable selection procedures use sample data to determine whether predictor variables are included in the regression model, these procedures increase the reliance of predictor weights on sample data and induce a second source of error for which these formulae are unable to account.

In order to provide some guidance for future validation efforts, Raju et al. (1999) conducted a comparative investigation of the predictive validity of formula-based and empirical estimates of population cross-validity, as well as an equally weighted combination of predictors under a number of sampling conditions. Specifically, predictor and criterion data were drawn from 84,801 Air Force enlistees. The data were then divided into 501 samples of seven sizes ranging from 25 to 200. These samples were used to calculate estimates of population cross validity using eleven cross validity estimation procedures. Calculation of the population squared multiple correlation revealed low to moderate intercorrelation between predictor variables ranging from .27 to .61. Comparison of methods revealed that population cross validity formulae can be used in lieu of traditional empirical cross-validity procedures without significant reduction in predictive effectiveness (Raju et al.). Furthermore, Burket's (1964) squared population

cross-validity estimation formula proved to be the most effective OLS procedure across all sampling conditions. Burket's equation is as follows:

$$P_{cv} = \frac{nR^2 - k}{R(n-k)} \quad (1)$$

Where: P_{cv} is the estimated population cross-validated multiple correlation

n is the sample size

R^2 is the unadjusted squared multiple correlation coefficient

k is the total number of predictors

Finally weighting the predictors equally rather than applying them to sample regression weights proved the most effective method in terms of cross-validity when sample sizes were less than 150. However, the authors noted that this result may not hold true in populations with differing levels of predictor intercorrelation.

Although Raju et al. (1999) answered some questions, other questions remain unanswered regarding cross validation adjustments. More explicitly, although the data prescribe the use of Burket's (1964) formula-based correction equation over other formula-based and empirical cross-validation procedure, one could also conclude that optimal weighting of predictor variables is inappropriate under many validation scenarios due to its relative inferiority to unit weighting. However, in spite of the finding that greater cross validity is attainable through unit weighting when compared to optimal weighting procedures in sample sizes less than 150 (Raju et al.), the universal applicability of this rule is illogical for two reasons. For one, the argument that all predictor variables are equally significant to the prediction of performance at any job is simply untenable. Additionally, although the differential predictive impact of certain variables may often be obscured by discrepancies of sample and population

characteristics, it is unlikely that this phenomenon is universally present in samples of less than 150. To support the previous assertion, it is first important to state that there is nothing magical about *n* sizes greater than 150. Although previous research comparing the performance of optimal and unit weighted composites indicated that the smallest sample at which optimal weighting performs as well as unit weighting was 150, the ability to infer differential predictive contribution under these circumstances is due to a proportional reduction of the probability of the influence of random (sampling) error as *n* increases. Therefore, it is possible, in certain scenarios, for smaller samples to yield a generalizable optimal weighting scheme. This outcome becomes increasingly likely when true differences exist between the relative contributions of various predictor variables. For example, if the population zero order correlations with *Y* are .25 and .6 respectively, the probability of even smaller samples yielding an optimally weighted composite that will cross validate more effectively than a unit weighted alternative is greater than if said population zero order correlations were more uniform (e.g., .30 and .35). In summary, although Raju et al.'s results indicate that assigning equal weights to predictor variables is often more effective than using OLS regression to form optimally weighted composites, the universal application of this rule seems inappropriate given the aforementioned considerations.

Range Restriction

As with validity shrinkage due to optimal weighting of predictor variables, non random sampling for the validation study represents a practical complication associated with validity research that systematically biases the observed correlation in the experimental setting. To illustrate how this typically occurs using a hypothetical example,

if a researcher attempts to provide criterion-related validity evidence for a cognitive ability test already in use by an organization, job performance information will only be available for those individuals who achieved the minimum cut score and were subsequently hired by the firm. Because the principal objective was to validate the instrument for use in the general applicant population, the portion of the sample for which data is available is not representative of the population for which the test will be operationalized. Furthermore, restriction of range in the validation sample is not inherent to the validity of the selection procedure (i.e., the test is not intended nor will it universally be employed in scenarios entailing comparable range restriction) and, thus, its effect spuriously reduces the generalizability of the validity coefficient to future testing situations (Schmidt, Hunter, & Urry, 1976). Although the most intuitive solution would be to simply obtain criterion data from the entire sample, this method is so vastly impractical in most selection scenarios that it merits no further consideration here. To more accurately estimate the operational validity of a selection procedure when criterion data for the validation sample is incomplete, Thorndike (1949) proposed the following correction.

$$R_{xy} = \frac{r_{xy}\left(\frac{S_X}{s_x}\right)}{\sqrt{1 + r_{xy}^{\,2}\left(\frac{S_X^{\,2}}{s_x^{\,2}} - 1\right)}} \qquad (2)$$

Where: R_{xy} is the unrestricted correlation between the predictor and criterion variables

S_X is the unrestricted predictor standard deviation

s_x is the restricted predictor standard deviation

r_{xy} is the attenuated correlation between X and Y

r_{yy} is an estimate of the reliability of scores on Y

Regarding the above correction, it is significant to note that although it is often employed in conjunction with multivariate prediction procedures (i.e., when selection decisions are made using a composite that represents multiple predictors), no empirical effort has, as of yet, been conducted to investigate the specific effect of a direct range restriction correction on a validity coefficient derived via multiple regression. The lack of research in this area is especially noteworthy given that the positive effect of range restriction corrections works in the opposite direction of the reductive outcome of Burket's (1964) and other formula-based estimates of population cross-validity.

Regarding the future of test validation for selection and personnel decisions, considerable development is necessary in order to increase the accuracy of the estimate of the validation coefficient, and by extension, the scientific integrity of industrial/organizational psychology. Although the sophistication and accuracy of the processes currently applied are testament to the progress made thus far in the field, the significance of the administrative implications of test validation within the context of personnel selection are such that any level of avoidable imprecision or error is unacceptable. In an effort to partially satisfy the implications of the previous statement, the empirical investigation of the following procedures is necessary.

First, although much has been written about the overestimation of validity coefficients obtained during multivariate prediction using OLS regression, little is known about the interaction of this source of bias when it occurs in conjunction with the underestimation of validity due to range restriction or criterion unreliability. In order to provide adequate guidance for practitioners facing validation scenarios in which both corrections are appropriate, differences among the possible combinations should be

examined in order to determine whether an optimal method exists. Specifically, such an examination conducted within controlled conditions should determine whether a particular combination of procedures is optimal and determine the efficiency of this optimal procedure. Additionally, all procedures and combinations must be assessed under varying degrees of predictor intercorrelation, selection ratio, and sample size in order to provide maximal generalizability across applied selection scenarios.

Optimal vs. Unit Weighting

In addition to an empirical test of the efficacy of opposing correction equations, it is also important to consider the problem posed by overfitting the regression equation to the derivation sample. Specifically, because sample characteristics often fail to adequately mirror population characteristics, the utility of a system of prediction that incorporates such inconsistencies is questionable. When considered in conjunction with Raju et al.'s (1999) empirical work, many well-informed practitioners may conclude that equal weighting represents the superlative choice of procedure under most multivariate prediction scenarios. However, given the logical arguments against the universal superiority of equal weighting listed previously, it stands to reason that this conclusion may be misapplied under many circumstances. Additionally, because the flaw inherent in the use of OLS regression is simply an artifact of sampling error, it is possible that the influence of sampling error can be modeled with a significance test. If one were to test the null hypothesis that all predictor variables contribute equally well to the prediction of the criterion, then a standard could be set to justify the use of OLS regression for multivariate prediction of performance. More simply stated, if using sample data to determine the probability of the chance occurrence of deviations from an equally

weighted scheme were possible, then the decision to employ OLS weighting would no longer depend on professional judgment or the results of previous simulations which may not be applicable to the current situation, but rather, the objective result of a significance test.

Although the logic of the previous assertion is sound, the efficacy of such a system must be tested empirically in order to provide tangible direction for applied scenarios. Specifically, I propose to employ the standard $\Delta R^2 F$-test typically used to determine the statistical significance of the change in multiple correlation coefficients following the removal or addition of predictor variables to a regression equation (Pedhazur, 1997). By treating the equally and optimally weighted composites as predictor variables and regressing the criterion first on the equally weighted composite alone and then both composites, this test will determine the statistical significance of the increase in validity when the optimally weighted composite is included in the regression equation. Essentially, the existence of a statistically significant finding would indicate that the superiority of the optimally weighted scheme is unlikely to have occurred by chance. Conversely, a nonsignificant ΔR^2would indicate that the inflated validity of the optimally weighted composite was likely due to the random effects of sampling error. Given this deduction, it appears that only conditions (i.e., large variance between individual predictor bivariate correlations with the criterion) for which optimal weighting outperforms unit weighting (as indicated by the aforementioned significance test) will the cross-validated optimally weighted composites be greater than the cross validated unit weighted composites.

One problem with the use of a hierarchical regression procedure to assess the value of an optimally weighted composite over a unit weighted composite, however, relates to the fact that the two composites will correlate very strongly when all predictors have nearly equal predictive power. The use of two highly correlated independent variables in the same regression equation results in what is known as colinearity. If colinearity is too high, then the regression analysis may fail to execute as desired. In summary, the procedure described above for testing the value of an optimally weighted composite may fail to execute under certain circumstances. Fortunately, $R_{Y,\,unit,\,optimal}^2 = R_{Y,optimal}^2$ (where unit is the unit weighted composite and optimal is the optimally weighted composite) because the unit weighted composite can never increase R^2 beyond the optimally weighted composite in the derivation sample. In other words, because the optimally weighted composite is the best possible set of weights for relating the independent variables to the dependent variable in that sample, no set of weights can exceed the optimal weights in that sample. As such, the residuals of the dependent variable when regressed on the optimally weighted composite will be uncorrelated with the unit weights. These residuals will also be uncorrelated with any other combination of the independent variables (again, in the derivation sample). In summary, the test for the value of optimal weights is given by $\Delta R^2 = R_{Y,\,unit,\,optimal}^2 - R_{Y,\,unit}^2$, with the usual F-test for the change in R^2 to determine significance.

In summary, the current study addressed two distinct yet related goals. First, the combination of the direct range restriction correction and Burket's (1964) adjustment formula will be empirically assessed to determine the optimal procedure as well as the efficiency of this procedure.

Second, although Raju's (1999) previous work suggested that optimally weighted composites typically do not outperform their unit weighted counterparts in many selection scenarios, for the reasons listed previously, it is illogical to conclude that unit weighting is not universally preferable. To test whether the ΔR^2 *F*-test outlined previously could be employed to identify situations in which optimally weighted composites would outperform unit weighted composites when cross validated, the following hypothesis was tested. Specifically, it is predicted that when the ΔR^2 test is significant the optimal weighting scheme would result in significantly greater cross-validated multiple correlations only for data sampled from a population in which the bivariate validity coefficients were dissimilar.

Method

Overview

Data consisting of 1,000,000 cases were generated with SAS version 9.2. All variables were normally distributed and set with specific correlations (defined below) to one another. These 1,000,000 cases constituted the hypothetical population. In order to adequately assess the accuracy of multiple validity estimation procedures both separately and in conjunction, each of the relevant conditions were repeated 500 times. The variables in the study included a single criterion as well as four predictor variables which were specific to the various conditions listed bellow.

Conditions

Again, in the interest of enhancing the generalizability of results, a number of conditions designed to emulate a variety of selection scenarios were employed. First, two levels of predictor intercorrelation were considered. Specifically, predictor variables demonstrated either moderate or low intercorrelation (.30 between all four predictors versus .10 correlations between all predictors). The rational for selecting these values was that although completely uncorrelated predictors are rare, predictor variables with correlations stronger than .30 are so redundant that their would likely be considered inefficient in applied settings. Therefore, in keeping with these assumptions, the levels of intercorrelation were selected to represent a realistic range of values. In addition to this conditional manipulation, cases were be selected within individual samples in order to induce direct range restriction at two distinct selection ratios (SR = .10 and SR = .33). Again, these values were selected in the interest of providing a realistic range of conditions. Samples were drawn at two sizes n = 150 and n = 200, which emulated

samples employed in antecedent research. In total, this process resulted in eight test conditions and required a total of 4,000 separate samples employing some 700,000 individual cases. Finally, although irrelevant to the evaluation of opposing correction sequences, it is important to note for informational purposes that across all eight of the conditions specific to this portion of the study, bivariate correlations with the criterion were as follows: r_{x1y} = .35, r_{x2y} = .35, r_{x3y} = .40, r_{x4y} =.40.

Next, regarding the assessment of the ΔR^2 *F*-test for identifying samples in which optimal weighting is appropriate, only two conditions were employed. First, in the interest of generating specific scenarios under which optimal weighting was appropriate, the bivariate correlation between the criterion and four of the predictors were highly dissimilar (r_{x1y} =.25, r_{x2y} = .25, r_{x3y} = .40, r_{x4y} = .40). In contrast, in order to generate an equal number of test cases in which both weighting schemes would yield identical results in the population (i.e., optimal weighting is not superior to equal weighting), the zero order correlations between a distinct second set of four predictors and the criterion were uniform (r_{x1y} = .35, r_{x2y} = .35, r_{x3y} = .35, r_{x4y} =.35). All predictor variables were correlated at .10. The sample size for all ΔR^2conditions was 150. There was no range restriction.

Procedure

As has been stated previously, the goals of this project involved the empirical assessment of various correlation adjustment procedures when direct range restriction occurs in conjunction with regression overfitting. Additionally, the study was intended to empirically assess the efficacy of the squared ΔR^2 *F*-test for determining the appropriateness of optimal weighting during multivariate prediction of performance. As

the intended outcome of this study was two-fold, an account of the procedures employed is likewise best presented sequentially.

Regarding the assessment of the correlation adjustment equations, it is necessary to consider the study in light of the applied scenarios which were emulated. Specifically, because the intention was the determination of the correct course of action under conditions of direct range restriction, the experimental samples underwent procedures that were similar to those which typically occur in practical circumstances. To elaborate, in an applied setting, an organization intent on making hiring decisions based on two untested predictors (and for which no regression equation yet exists) would likely form an equally weighted composite of the experimental predictors and select applicants top-down based on their combined scores. Therefore, within each subsample, unit weighted composites of the predictors were generated (using the sample mean and standard deviation to standardize predictor variables) and cases were selected top-down at the designated selection ratio. Having effectively simulated direct range restriction within our samples, optimally weighted composites were then generated using the selected cases (i.e., those not removed to induce range restriction) and estimates of the population cross-validated multiple correlation were generated using the various procedures of interest. Specifically, two estimates of the population squared cross-validity were generated using alternate sequences of these two adjustments. More explicitly stated, under one condition coefficients were adjusted first for direct range restriction followed by Burket's (1964) adjustment; whereas in the other, Burket's formula adjustment was applied first followed by the standard adjustment for range restriction.

In addition to the assessment of these sequences, two baseline conditions were generated to provide an idea of the relative effectiveness of dual corrections when compared with either procedure alone. Although these conditions were termed *baseline*, it is important to note that they are not necessarily representative of a specified standard of cross-validity performance. Rather, it may be more accurate to conceptualize these estimates as being free of the influence of any interaction of the two adjustments. To illustrate, optimally weighted baseline conditions were generated such that the sample was unaffected by range restriction, and the resulting multiple correlation was adjusted using only Burket's (1964) method. Likewise, range restricted baselines were subjected to direct range restriction but predictors were equally weighted, thus avoiding validity overestimation due to optimal weighting. As such, the multiple correlation coefficients in these conditions were adjusted using only the standard correction for direct range restriction.

Under both predictor sets (i.e., those with highly divergent zero-order correlations with the criterion vs. those with more uniform zero-order correlations with the criterion), unit weighted and optimally weighted composites were generated. To generate the unit weighted composite, predictor variables were standardized using the sample mean and standard deviation. The optimally weighted composites were produced using OLS procedures. The R^2 for the unit and optimally weighted composites were used to determine the squared correlation between the dependent variable and the optimally weighted composite, controlling for the unit weighted composite (i.e., $R^2_{optimal} - R^2_{unit}$) and the significance of this correlation was examined via the F-test for the change in R^2.

Assessment of Outcomes

To evaluate the effectiveness of these procedures (two baseline conditions and two dual correction conditions with varied order) across all of the aforementioned selection conditions (predictor intercorrelation, sample size, and selection ratio), the estimated validities were compared to the population squared multiple correlation in terms of bias and squared difference. To clarify, bias is the deviation of the sample validity from the population squared multiple correlation (bias = population cross validated R^2 – sample R^2). The squared difference is the squared bias (squared bias = [population cross validated R^2 – sample R^2]^2). Essentially bias demonstrates the average error of the estimate whereas squared bias is an index of the variability of said estimates. Finally, it is helpful to reiterate that the population squared multiple correlation was determinable in this instance due to the ability to access the relevant values of the population (1,000,000 cases).

In order to evaluate the utility of the ΔR^2 for determining the appropriate weighting scheme in applied scenarios, a second sample was randomly drawn to assess the accuracy of predictions made in the original sample. This method was nothing more than the classic cross validation study in which a prediction equation is formed on the basis of one sample and tested on a second sample. Specifically, the weights for both optimal and unit weighted composites generated in the derivation sample were applied to the cross-validation sample with the expectation that the average cross-validated R^2 for the optimally weighted composites would be greater than those derived through the unit weighting procedure when the unit weighted composite yielded a significant result in the original (i.e., derivation) sample.

Results

As is shown in Appendix A, correcting for direct range restriction prior to application of Burket's (1964) formula resulted in lower mean bias and mean squared bias across all conditions with one exception (unlike all other conditions, bias was slightly greater in the seventh condition when coefficients were adjusted for range restriction first). Application of Burket's formula prior to correction for direct range restriction resulted in 36% greater mean squared bias across conditions. It is also important, albeit unsurprising, to note that, both sequences yielded greater bias under more stringent selection ratios (see Appendix A) and when the sample size was smaller (see Appendix B). Furthermore, when range restriction was corrected first, greater bias occurred when predictor variables were more strongly intercorrelated to one another (see Appendix C). However, this final difference did not occur when the opposite sequence of corrections was applied.

As can be seen in Appendix D, correcting for range restriction first resulted in underestimation of the population cross-validated multiple correlation in all conditions whereas applying Burket's (1964) formula first resulted in overestimation of this value across all conditions. Finally it is important to note that, due to a programming error, baseline conditions were only generated under two of the conditions (i.e., Conditions 1 and 8). When dual corrections were compared to baseline estimates in these conditions, Burket's formula alone yielded the least mean squared bias. Inversely, the sequence in which Burket's formula was applied first resulted in the most bias compared to both baselines and the opposing sequence. Interestingly, when range restriction was corrected prior to the application of Burket's formula, the resulting mean squared bias was nearly

identical to that which was observed under the baseline condition in which samples were affected by and corrected for range restriction only (see Appendix E).

With regard to the assessment of the $\Delta R^2 F$-test, the results were somewhat convoluted. No statistically significant mean differences were found for cross-validated multiple correlations between significant and nonsignificant ΔR^2results ($\alpha = .01$). Additionally, this outcome remained constant when the stringency of the test was reduced ($\alpha = .05$) and increased ($\alpha = .001$) for exploratory purposes. Furthermore, when unit-weighted cross-validated multiple correlations were subtracted from the optimally weighted cross-validated multiple correlations in the condition tailored to elicit greater cross validity for optimal weighting schemes (i.e., the condition in which predictor zero-order correlations were nonuniform), the resulting difference was negatively correlated with the results of the ΔR^2 test, $r = -.149$, $\alpha = .01$. Essentially, this result indicated that, contrary to the stated hypothesis, the magnitude of the F statistic was associated with scenarios in which the unit weighting scheme resulted in greater cross validated multiple correlations.

Table 1
Mean Bias and Mean Squared Bias Across Conditions

Condition	Mean Bias	Mean Squared Bias
Condition 1		
Optimally Weighted Baseline	.00316	.00381
Range Restricted Baseline	.00417	.01205
Range Restriction Correction, Cross Validity	-.02943	.01219
Cross Validity, Range Restriction Correction	.05069	.01996
Condition 2		
Range Restriction Correction, Cross Validity	-.01092	.00783
Cross Validity, Range Restriction Correction	.03266	.01108
Condition 3		
Range Restriction Correction, Cross Validity	-.05044	.01476
Cross Validity, Range Restriction Correction	.05441	.01934
Condition 4		
Range Restriction Correction, Cross Validity	-.02451	.01003
Cross Validity, Range Restriction Correction	.03489	.01278
Condition 5		
Range Restriction Correction, Cross Validity	-.01951	.01085
Cross Validity, Range Restriction Correction	.04208	.01685
Condition 6		
Range Restriction Correction, Cross Validity	-.00594	.00602
Cross Validity, Range Restriction Correction	.02728	.00822
Condition 7		
Range Restriction Correction, Cross Validity	-.05071	.01364
Cross Validity, Range Restriction Correction	.03124	.01571
Condition 8		
Optimally Weighted Baseline	.00631	.00302
Range Restricted Baseline	-.00149	.00805
Range Restriction Correction, Cross Validity	-.02146	.00834
Cross Validity, Range Restriction Correction	.02451	.01021

Note. For conditions 1 - 4, n = 150. For conditions 5-8, n = 200. For Conditions 2, 4, 6, and 8, SR = .33. For Conditions 1, 3, 5, and 7, SR = .1. For Conditions 1, 2, 5, and 6 predictor intercorrelation was .10. For Conditions 3, 4, 7, and 8 predictor intercorrelation was .30. Baseline conditions were corrected using only the appropriate calculation (e.g., coefficients derived in the optimally weighted baseline were corrected using Burket's, 1964, equation only)

Discussion

Given the superiority of the procedural sequence in which coefficients were first adjusted using the standard correction for direct range restriction followed by Burket's (1964) adjustment formula in terms of both mean squared bias across all conditions and mean bias in seven of the eight sets of conditions, it is clear that this method is the proper procedure when multiple correlation coefficients are to be adjusted to more closely approximate the population cross-validated multiple correlation. Furthermore, because this method consistently underestimates this value whereas its counterpart was shown to overestimate it across all test conditions, range restriction followed by Burket's adjustment is clearly the superior method for research settings in which conservative adjustments are preferable.

Additionally, it appears that both procedural sequences resulted in relatively similar and predictable increases both under the smaller of the two sample sizes and when the selection ratio was more stringent. Although future analyses considering a greater range of sample sizes and selection ratios may identify differential effects of these factors on the resulting bias associated with either sequence of corrections, the current results indicate that the increased bias associated with these conditions is similar for both sequences. Furthermore, because a correction for direct range restriction followed by Burket's (1964) adjustment yielded the least bias across all levels of samples size and selection ratio, it would appear that this sequence is the better choice regardless of the stringency of these conditions in applied scenarios. Similarly, although the current results demonstrate a slightly greater increase in bias associated with stronger predictor

intercorrelation when range restriction is corrected first, future research could further investigate the differential effects of predictor intercorrelation on the procedural sequences by considering more levels of this variable. If an interaction between procedural sequence and predictor intercorrelation were to be identified, there may also be a level of the latter variable at which the opposite sequence of corrections should be employed. However, it is important to note that although consideration of more extreme levels of intercorrelation may reveal this type of interaction, the range considered in this study (i.e., $r_{xx} = .10$ and $r_{xx} = .30$) represents the range typically found in applied scenarios. Therefore, future research investigating the efficacy of these correction sequences under more extreme levels of predictor intercorrelation may not hold much applied value.

With respect to the baseline conditions, it is somewhat unsurprising that the least bias occurred when multiple correlation coefficients were subjected to only the inflating effects of optimal weighting (i.e., the baseline condition of cross-validity only) and were subsequently corrected using Burket's (1964) formula. Similarly unsurprising was the finding that the greatest bias occurred when coefficients were subject to both sources of bias (i.e., optimal weighting and range restriction) and the least effective procedural sequence was applied (i.e., Burket's formula followed by the standard correction for direct range restriction). Of much greater interest was the finding that the better of the two sequences of corrections (i.e., correction for range restriction followed by application of Burket's formula) resulted in bias nearly identical to the baseline condition in which the only source of bias was range restriction corrected using the standard method (i.e., the range restriction baseline condition). This finding suggests that the additional bias

associated with optimal weighting is negligible when corrected in the appropriate manner. Although a more comprehensive study comparing baseline and dual correction procedures across a greater range of conditions may provide more information as to the comparative utility of these procedures, the current findings suggest that when validity coefficients are subject to both range restriction and overfitting due to optimal weighting, one can effectively account for these opposing influences by first correcting multiple correlation coefficients for range restriction and then applying Burket's formula. Furthermore, given these results, it would appear that multiple correlation coefficients subject to both influences but are corrected using the appropriate sequence of corrections will be nearly as accurate as those that are subject only to range restriction and corrected using the standard procedure.

Regarding the failure of the ΔR^2 test to identify samples in which an optimal weighting scheme would cross-validate more effectively than the unit weighting scheme derived from the same sample, a few considerations are relevant. First, under the condition in which all of the bivariate correlations were identical, any deviation of the sample from the population resulted in an optimal weighting scheme that was unlikely to cross-validate as effectively as its unit weighted counterpart. Stated differently, because all of the predictors in this population were equally useful to the prediction of the criterion, equal weighting of sample predictors yielded a set of weights that was likely to cross-validate well in future samples derived from this population (i.e., equally weighting of predictors in a sample emulated the relationship of predictors to the criterion in the population). As such, any deviation of the sample from the population resulted in optimal weighting schemes in which predictor weights were not equal and were, therefore, less

likely to cross-validate as effectively in other samples derived from this population. Had the ΔR^2 test worked perfectly, virtually none of these samples would have yielded a significant result because the desired outcome of the ΔR^2 procedure was the identification of samples for which the superiority of an optimal weighting scheme was sufficiently robust to conclude that it reflected the population values and would function better than an equally weighted composite when applied to future samples. Again, because this type of a relationship was not present in this population and therefore could only occur in individual samples as a result of sampling error, any statistically significant ΔR^2 results simply represented Type I errors. As such, it was interesting that at $\alpha = .01$, the test yielded a significant (i.e., erroneous) result in roughly 10% of the samples from this condition. Although, the observed frequency of Type I error was greater than that which is expected at $\alpha = .01$, this finding alone was not sufficient to discard the test. Had this been the only issue, the stringency of the *p*-value necessary to assume that optimal weighting schemes would be preferable in future samples could have been adjusted to yield a preferable rate of this outcome.

In the condition in which predictor zero-order correlations to the criterion were divergent in the population (i.e., $r_{x1y} = .25$, $r_{x2y} = .25$, $r_{x3y} = .40$, $r_{x4y} = .40$), the ΔR^2 test again failed to identify samples for which the cross-validated optimally weighted multiple correlation was greater than the cross-validated unit weighted multiple correlation. Due to the fact that predictors were not equally related to the criterion in this population condition, a number of samples produced stronger optimally weighted cross-validities, meaning that the optimal weighting scheme cross-validated more effectively than the unit weighting scheme when applied to a second sample. Again, however, a

number of scenarios occurred in which the magnitude of the superiority of the optimal weighting scheme from the derivation sample was sufficient to yield a significant ΔR^2 but failed to outperform unit weighting when applied to the cross-validation sample. Furthermore the ΔR^2 procedure did not consistently distinguish these samples from those in which optimal weighting did outperform unit weighting when applied to a second, cross-validation sample. It is worth noting that the mean cross-validated optimally weighted multiple correlation was only slightly greater than the mean cross-validated unit weighted multiple correlation. Stated differently, even though the relation of the predictor variables to the criterion varied in the population, optimally weighted composites derived in samples of this population cross validated only marginally better on average than did unit weighted composites generated in the same way. That said, had the predictor zero-order correlations been even more divergent (e.g., r_{x1y} =.10, r_{x2y} = .10, r_{x3y} = .55, r_{x4y} = .55), the population may have yielded more samples in which the superiority of optimal weighting was more easily discernable. However, it is relevant to note in relation to the previous assertion that the predictor zero-order correlations used in this study represented as extreme a range as was possible without becoming unrealistic. Although it may be mathematically possible to produce a scenario in which optimal weighting was vastly superior to unit weighting, the results of such a study would likely hold little applied utility.

An explanation for the failure of the ΔR^2 test may be the sample size. It is possible that sample sizes of 150 were simply too small to accurately and consistently model the population weights in both the derivation and cross-validation samples. Had samples of 200 or 250 been considered, it is likely that the superiority of optimally weighted cross-

validities may have been more easily discernable. As such, it is possible that future studies with larger sample sizes may identify conditions in which this test may usefully be employed.

Limitations

As was alluded to in previous sections, the exploratory limitations associated with the current study were due, in part, to the absence of additional levels for each condition. A more comprehensive analysis considering a wider range of sample sizes, selection ratios, and levels of predictor intercorrelation may provide a more comprehensive understanding of the these procedures. Additionally, due to the aforementioned programming error, baseline estimates were calculated in only two of the eight sets of conditions. Had this not occurred, conclusions with regard to these conditions may have been further substantiated.

Furthermore, although the evaluation of various correction sequences in terms of their ability to approximate population values yielded a clear and unbiased means of assessment, it may also have been beneficial, in terms of applied value, to have assessed these correction sequences using empirical cross-validation. To elaborate, although the ability to estimate the strength of relationships in the population is of key importance, due to sampling error, performance in the previous domain is not identical to performance with respect to predicting the strength of relationships in future samples. Again, although the ability to predict within the population is obviously linked to prediction in future samples, examination of the latter may be informative.

Direction for Future Research

In addition to the modifications listed previously, it is important to note that a number of other correction formulae have been used to account for various statistical artifacts (i.e., criterion unreliability and indirect range restriction). Furthermore, these procedures are similar to the correction for direct range restriction in that they are all upward adjustments, meaning that they are designed to increase validity coefficients that are spuriously reduced for a specified reason. As such, each of these types of adjustments, like the correction for direct range restriction, works in the opposite direction of Burket's (1964) Formula. Furthermore, because these operations are mathematically distinct from the correction for direct range restriction, the findings of the present study are not generalizable to scenarios in which these corrections are appropriate. As such, future efforts should similarly assess the performance of various sequences of these equations in conjunction with Burket's adjustment formula.

Finally, it is important to note in relation to the assessment of the ΔR^2 procedure, future research should assess the performance of this procedure in larger samples. Specifically, if the current procedure were repeated using samples of 200 or greater, this procedure may either be more conclusively discarded. Alternatively, a scenario in which this procedure may be usefully employed would be identified.

Conclusion

Based on the current findings, two general conclusions can be made. First, under selection scenarios employing multivariate procedures that are subject to direct range restriction, practitioners can obtain a more accurate estimate of the cross validated squared multiple correlation by first correcting validity coefficients for direct range restriction and then adjusting them using Burket's (1964) formula. Second, based on the

available evidence, estimates generated using this sequence should be only slightly less accurate than those that can be expected when validity coefficients are subjected to and corrected for the direct range restriction alone. To reiterate, however, when the conditions of the selection scenario require both types of correction, the findings of the current study suggests that range restriction should be corrected prior to the application of Burket's formula. Finally, regarding the assessment of the ΔR^2 test, it would appear that this procedure, as outlined previously, is not an effective means for predicting the appropriateness of various weighting schemes under the conditions included in this analysis.

References

Burket, G. R. (1964). A study of reduced rank models for multiple prediction. *Psychometrika Monograph Supplement,* No. 12.

Claudy, J. (1978). Multiple regression and validity estimation in one sample. *Applied Psychological Measurement*, *2*, 595-607.

Drasgow, F., Dorans, N, J., & Tucker, L. R. (1979). Estimators of the squared cross-validity coefficient: A Monte Carlo investigation. *Applied Psychological Measurement, 3,* 387-399.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions.* Mahwah, NJ: Lawrence Erlbaum Associates.

Mitchell, T. W., & Klimoski, R. J. (1986). Estimating the validity of cross-validity estimation. *Journal of Applied Psychology, 71,* 311-317.

Mosier, C. (1951). The need and means of cross validation. I. Problems and designs of cross-validation. *Educational and Psychological Measurement*, *11,* 5-11.

Murphy, K. (1984). Cost-benefit considerations in choosing among cross-validation methods. *Personnel Psychology*, *37,* 15-22.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Fort Worth, TX: Harcourt Brace.

Raju, N. S., Edwards, J. E., & Fleer, P. F. (1997). Methodology review: Estimation of population validity and cross-validity, and the use of equal weights. *Applied Psychological Measurement, 21,* 291-305.

Raju, N., Bilgic, R., Edwards, J., & Fleer, P. (1999). Accuracy of population validity and cross-validity estimation: An empirical comparison of formula-based, traditional empirical, and equal weights procedures. *Applied Psychological Measurement, 23,* 99-125.

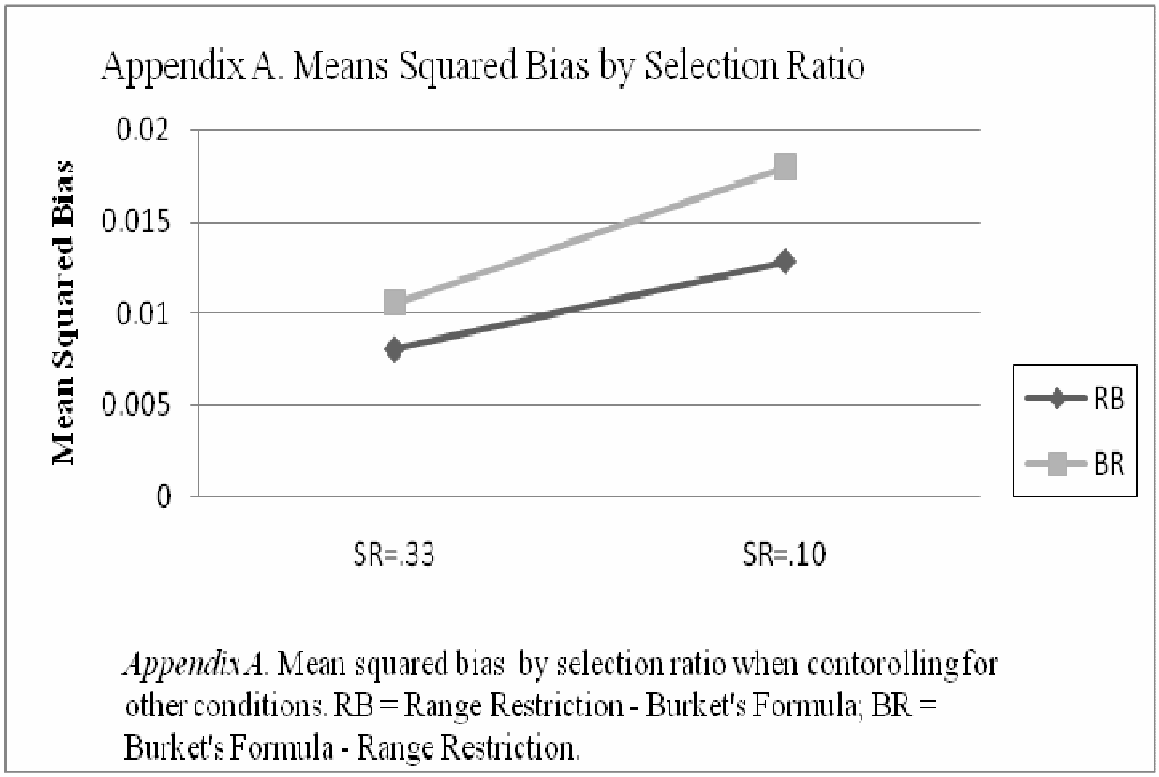
Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology, 61,* 473-485.

Society for Industrial Organizational Psychology, Inc. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Author.

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

Appendix A

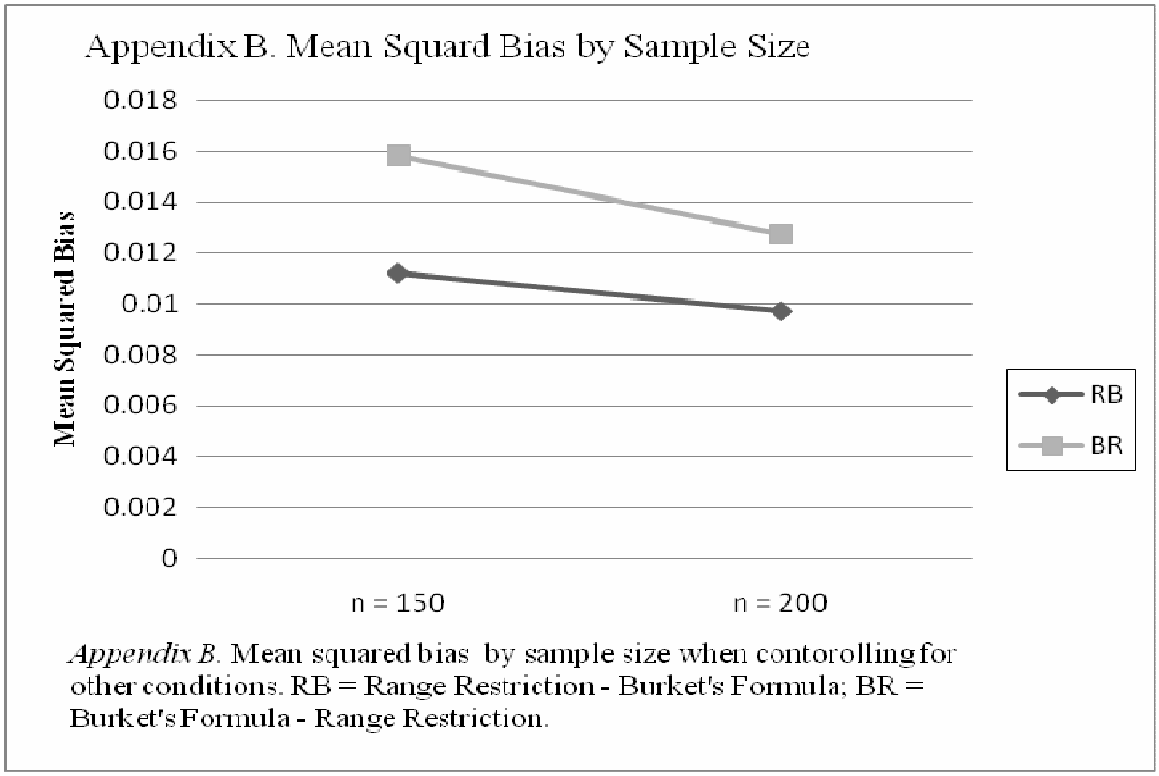
Mean Squared Bias by Selection Ratio



Appendix A. Mean squared bias by selection ratio when contorolling for other conditions. RB = Range Restriction - Burket's Formula; BR = Burket's Formula - Range Restriction.

Appendix B

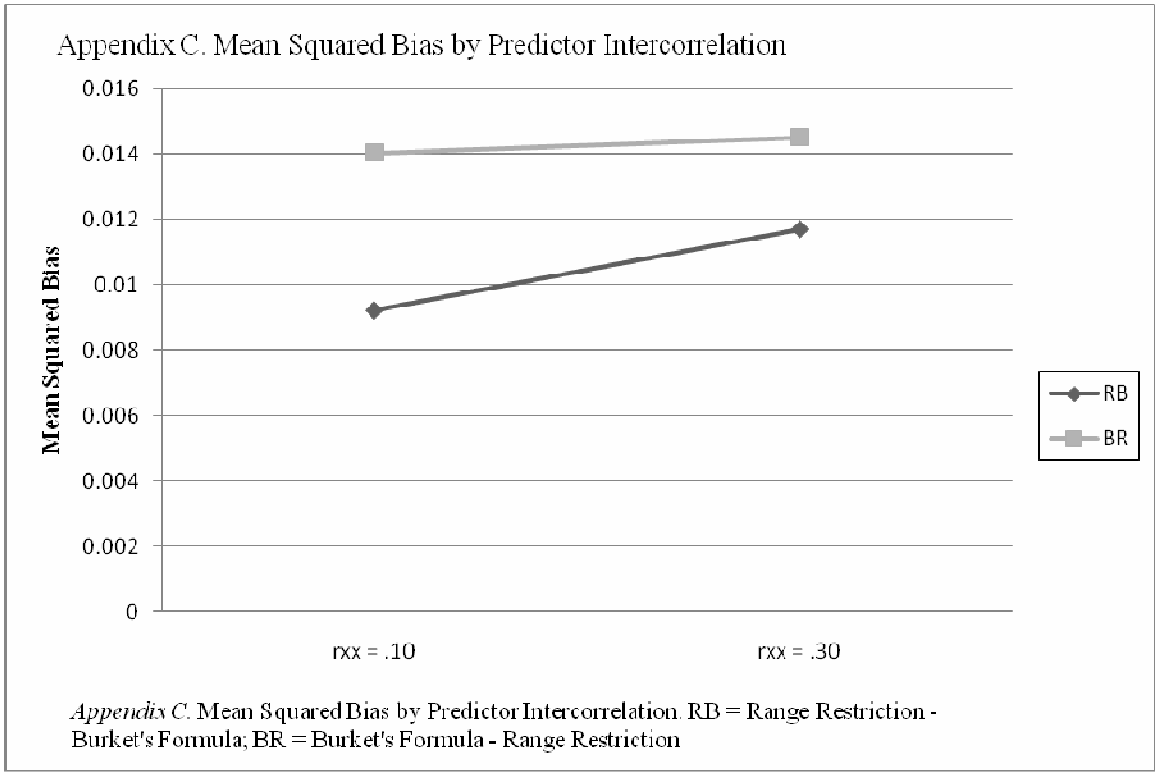
Mean Squared Bias by Sample Size



Appendix B. Mean squared bias by sample size when contorolling for other conditions. RB = Range Restriction - Burket's Formula; BR = Burket's Formula - Range Restriction.

Appendix C

Mean Squared Bias by Predictor Intercorrelation



Appendix C. Mean Squared Bias by Predictor Intercorrelation. RB = Range Restriction - Burket's Formula; BR = Burket's Formula - Range Restriction

Appendix D

Mean Bias Across Conditions

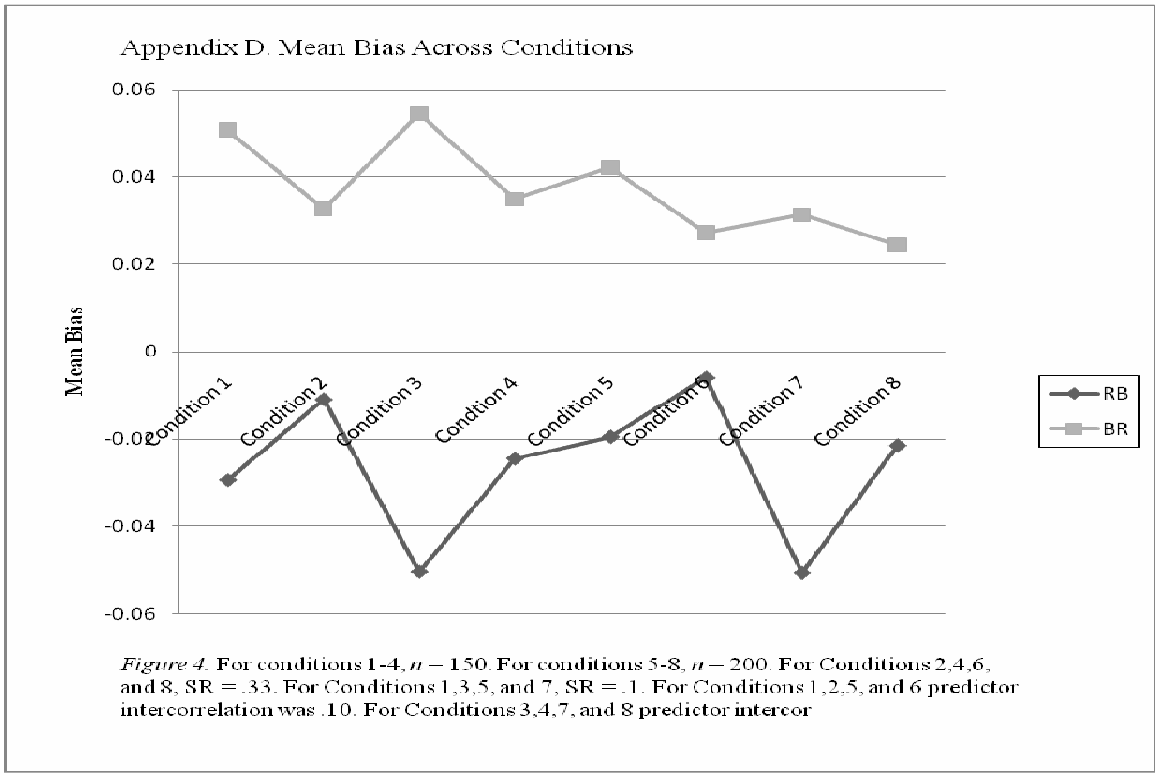
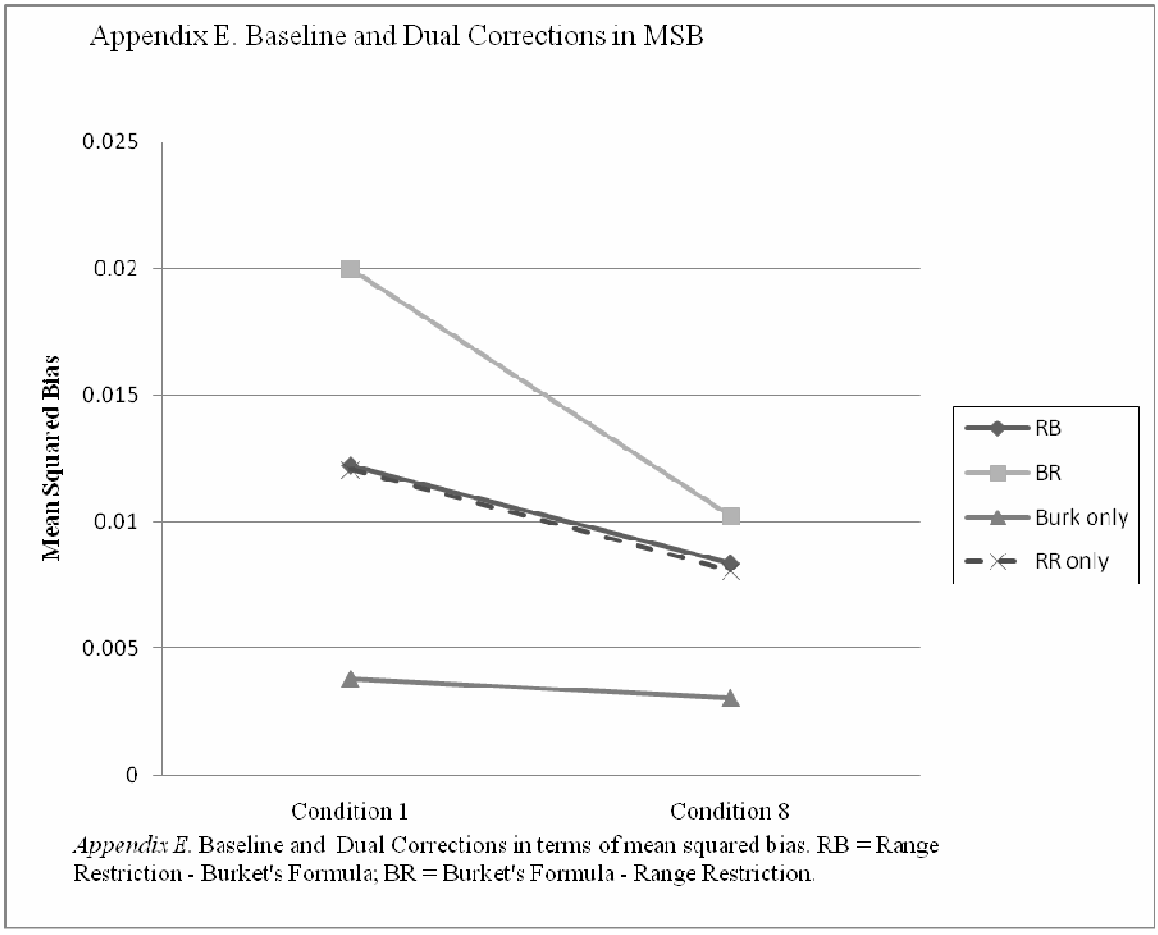


Figure 4. For conditions 1-4, $n = 150$. For conditions 5-8, $n = 200$. For Conditions 2,4,6, and 8, SR = .33. For Conditions 1,3,5, and 7, SR = .1. For Conditions 1,2,5, and 6 predictor intercorrelation was .10. For Conditions 3,4,7, and 8 predictor intercor

Appendix E

Baseline and Dual Corrections in MSB



Appendix E. Baseline and Dual Corrections in terms of mean squared bias. RB = Range Restriction - Burket's Formula; BR = Burket's Formula - Range Restriction.