# A Step-Down Hierarchical Multiple Regression Analysis for Examining Hypotheses About Test Bias in Prediction

**Gary J. Lautenschlager**
**University of Georgia**

**Jorge L. Mendoza**
**Texas A&M University**

The problem of determining test bias in prediction using regression models is reexamined. Past approaches have made use of separate regression analyses in each subgroup, moderated multiple regression analysis using subgroup coding, and hierarchical multiple regression strategies. Although it is agreed that hierarchical multiple regression analysis is preferable to either of the former methods, the approach presented here differs with respect to the hypothesis testing procedure to be employed in such an analysis. This paper describes the difficulties in testing hypotheses about the existence of bias in prediction using step-up methods of analysis. Some shortcomings of previously recommended approaches for testing these hypotheses are discussed. Finally, a step-down hierarchical multiple regression procedure is recommended. Analysis of real data illustrates the potential usefulness of the step-down procedure.

The concept of predictive bias in testing, as used here, refers to equivalence of predictions for a given test score regardless of subgroup membership. In effect, this is the Cleary (1968) definition of test bias as equivalence of regressions. An important assumption made when using any of the regression models is that the criterion measure represents an acceptable measure of performance. Therefore, the criterion is assumed to be fair for both groups, and any bias in prediction must be due to problems in the predictor(s).

*APPLIED PSYCHOLOGICAL MEASUREMENT*
*Vol. 10, No. 2, June 1986, pp. 133–139*

*0146-6216/86/020133-07$1.60*

The issue of whether or not a predictor test is biased for a given situation is actually the test of an hypothesis about predictive bias. Though this statement may seem obvious (and perhaps unnecessarily redundant), it is made to highlight the fact that a form of hypothesis testing occurs when the issue of predictive bias is examined. Therefore, it is important to keep in mind the nature of the hypotheses that are being tested. As will be shown, the method chosen for testing hypotheses about predictive bias can potentially influence the results obtained.

In order to examine the issue at all, the usefulness of a test for predicting a criterion must be assumed or demonstrated. It is differences in prediction that represent sources of test bias. To determine whether a test is biased, the null hypothesis (i.e. that the test is unbiased) is first tested. In the most general case, the alternative hypothesis tested is that the test is biased, i.e. there are differences in prediction. This last statement is crucial. Note that it does not address possible differences between predictions with respect to separate slopes and/or intercepts. Tests of slopes and intercepts should be addressed after it has been determined that there is good reason to suspect the test may be biased.

## Previous Approaches

Considerable debate has revolved around the issue of differential validity, and has focused largely

on the examination of zero-order correlations of a predictor test with a criterion variable. Linn (1978) questioned the efficacy of examining the validity coefficients and pointed out the importance of examining differences in prediction. This has led to the development of a number of different approaches that examine whether or not there is differential prediction. Some of these approaches are discussed below. All of them assume that the variance estimates for the subgroups can be pooled. If this is not the case, the reader is referred to Ragosa (1980).

All of these approaches can be expressed as variants of multiple regression. Three predictors are used to determine the criterion score: the test score, a coded subgrouping variable, and a variable representing the cross-product of the test score with the subgrouping variable. The first variable is used to determine if the predictor test itself is useful for prediction. The second variable addresses the question of whether there are intercept differences, and the third variable focuses on possible slope differences. The significance of the regression weights for each of the predictors is examined.

As Bartlett, Bobko, Mosier, and Hannan (1978) pointed out, the coding scheme adopted for the subgrouping factor can affect the correlations of the predictor (test) and the subgrouping variable with their cross-product. In addition, any changes in the origin of the test scores, such as using deviation scores, can also influence these correlations. However, the tests of significance for the subgrouping variable and the interaction are not affected by these transformations (see Cohen, 1978), although the tests for the predictor and the intercept are affected. As a rule, then, coding or centering need not be matters of concern when testing for predictive bias.

A refinement of the multiple regression strategy was given by Cohen and Cohen (1975) and Bartlett et al. (1978) in their hierarchical procedures. These two sources present similar approaches for using hierarchical multiple regression (HMR) analysis to examine test bias. Regardless of the difference in the two approaches, they both share the same shortcoming from an hypothesis testing standpoint. Both of these procedures will be briefly outlined here and will jointly be referred to as step-up HMR analysis. Following this discussion, some points of concern are raised regarding the way hypotheses about bias in prediction are conceptualized in the step-up hierarchical procedure.

The step-up HMR models start with a single predictor in a multiple regression equation and then test for the increment in $R^2$ as additional variables are added. A dummy-coded variable is used to represent the predictor variable of subgroup membership, and a cross-product of the dummy-coded variable with the predictor test is a third predictor variable used in the analysis. The two step-up HMR procedures differ with respect to which predictor is entered first.

Cohen and Cohen (1975) enter the subgroup variable first, in effect to control for differences between group means on the dependent variable. The predictor test is entered second, and the cross-product term is entered last. Bartlett et al. (1978) suggest entering the predictor test first to determine whether there is a significant overall relation between the predictor and the dependent variable. (Bartlett et al. recognize the potential weakness of this test for overall validity.) At the second step the subgroup variable is entered, and in the third step the cross-product is entered. At each step the $R^2$ is tested for significance to determine if the variable entered improves prediction.

A problem associated with both step-up HMR procedures is that at each step all higher order effects *not* included in the model are pooled into the sum of squared error term (SSE), potentially decreasing the power of the sequential testing procedure. For example, a term that is not included that has an effect will contribute more to SSE than is offset by its accompanying degree of freedom. Cohen and Cohen (1975, p. 303) acknowledged this problem in discussing their choice of an error term at stages in the step-up process, and pointed out that this problem could be ameliorated by using instead what they called the sum of squares pure error.

An HMR strategy which examines increments in prediction can indeed be a useful approach for examining test bias in prediction. However, such an approach should be conducted in a manner that

allows more power in testing for the existence of bias. A procedure of this type would begin with a test of the hypothesis of test bias against the null hypothesis that the test shows no bias. When the null hypothesis is rejected, further testing is conducted to examine the nature of the bias. In this way, tests that follow are always conditional on the test of bias; and the bias test is most powerful, since it only involves a single hypothesis which is tested with the smallest possible error term. This step-down procedure is in agreement with Cramer's (1972) recommendations for multiple regression tests.

Table 1
Hierarchical Multiple Regression Models

| | |
|---|---|
| 1. | $Y = b_{10} + b_{11} X + e$ |
| 2. | $Y = b_{20} + b_{21} X + b_{22} S + b_{23} XS + e$ |
| 3. | $Y = b_{30} + b_{31} X + b_{32} XS + e$ |
| 4. | $Y = b_{40} + b_{41} X + b_{42} S + e$ |

Note. Y is a fair measure of performance, X is a predictor test score, S is a dummy coded variable indicating subgroup membership, XS is a cross-product term obtained by multiplying X times S, and e is a residual.

### Step-Down Procedure

The step-down procedure outlined here tests the hypothesis of bias in prediction assuming the null hypothesis that a common regression line provides the best fit. The alternative is that a full model, including slope and intercept differences between subgroups, is required. This method is also based on a dummy-coded variable for group membership and a cross-product variable.

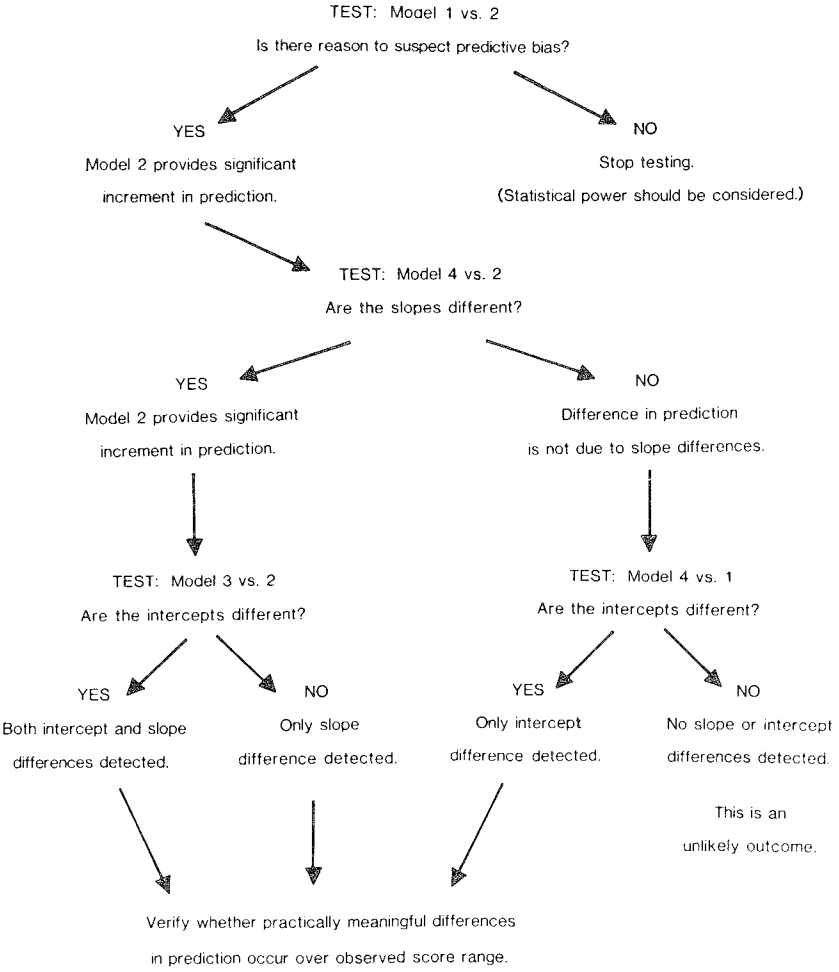
The models that will be used to test hypotheses about bias in prediction are presented in Table 1. The procedure begins by testing the hypothesis that a common regression line alone is sufficient to account for the relation of the predictor test with the criterion. This is the test of the omnibus hypothesis of prediction bias. The null hypothesis in this case is given in Model 1, where only the predictor test, X, is used. The increment in $R^2$ gained by using Model 2 rather than Model 1 is tested for significance. It is important to note that the SSE for Model 2 is the smallest SSE possible (Cohen's pure error; Cohen & Cohen, 1975) for the simple one-predictor, two-subgroup case. This statistical test examines whether there is a significant reduction from Model 1 SSE and is a test of whether two regression lines are identical (Neter & Wasserman, 1974). (In terms of Model 2, this is the simultaneous test that both $b_{22}$ and $b_{23}$ equal zero.) If there is a significant reduction in SSE using Model 2, prediction bias should be inferred. To determine the nature of the bias, further tests for slope and/or intercept differences must be performed.

It is at this stage that the role of the SSE term involved in the denominator of statistical tests can lead to different outcomes regarding tests of slope and intercept differences. The occurrence of a significant reduction in SSE in the previous step implies that there are either slope and/or intercept differences. To determine whether such differences exist, a step-down procedure tests separate models that eliminate each of these differences sequentially against Model 2, the model with the smallest SSE possible. The specific sequence of tests that are used will depend on the outcome of previous tests.

Figure 1 presents a flowchart that shows the sequence of tests using the step-down HMR approach to examine hypotheses about predictive bias. The test for differences in slopes should follow the omnibus test of predictive bias. It involves testing Model 4 against Model 2 to determine whether there is a significant reduction in SSE by using Model 2. A significant increment in $R^2$ based on a comparison of these models would imply different slopes, indicating a potential need for separate regression equations.

The test for differences in intercepts for the subgroups will be dependent upon the outcome of the test for slope differences. If there are slope differences, the test of intercepts will involve testing Model 3 against Model 2 to determine if there is a significant increase in SSE when the subgrouping variable is eliminated. If this test shows a significant change in $R^2$, then separate regression

**Figure 1**
Flowchart Depicting Steps in the Use of
the Step-Down HMR Procedure for Examining Prediction Bias




equations may be required. Given that slope differences exist, if the hypothesis of intercept differences is tested by using Models 1 and 4, in the sequence suggested by Bartlett et al. (1978), there will be a greater risk of Type II error. Differences in slope should be taken into account when testing for differences in intercepts. Taking the possibility of slope differences into account reduces the SSE and makes for a more powerful and appropriate test. (From a practical standpoint, the test for intercept differences, after finding slope differences, may not be as important as an examination of the meaningfulness of differences in prediction, as discussed below.) If slope differences were not detected initially, then intercept differences should be tested by testing Model 4 against Model 1.

The individual tests of intercept and slope in the step-down procedure are identical to the tests that would be obtained using a procedure for comparing regression equations on different groups (see Gulliksen & Wilks, 1950, or Kerlinger & Pedhazur, 1973, for a description of the procedure). These

tests are the same as the partial tests alluded to earlier, and can be easily obtained by running a regression equation on the three variables (grouping, predictor, and cross-product). The *b* weights of the grouping and cross-product terms give the tests of intercept and slope, respectively. The tests are found in many computer programs, e.g. in SAS (SAS Institute, 1979) under type IV sum of squares in the GLM procedure.

The final step in the process of examining predictive bias should be a consideration of the practical meaningfulness of any evidence of prediction bias. It would be desirable to determine whether the differences in regression occur within a range of the predictor score distribution that will result in differential predictions; it may well be that differences in predictions are negligible at particular score values, or over the entire observed range of predictor and criterion scores of both subgroups. Confidence regions can be established at particular scores to determine whether predicted scores differ. The Johnson-Neyman technique or related methods can be used to determine regions of significant differences over the score ranges (cf. Pedhazur, 1982; Ragosa, 1980, 1981; Schmidt & Hunter, 1982). The use of such procedures takes into account differences in intercepts that likely coexist with differences in slopes.

Statistical power should also be a concern. Bias may exist, but may go undetected due to small sample size(s). Conversely, any difference in regression equations could be detected as statistically significant bias in prediction, given large enough sample sizes. A useful guide under both circumstances would involve using $R^2$s (proportions of variance accounted for) to aid in determining the practical usefulness of results.

Power may also be an issue with respect to the "unlikely outcome" of the testing procedure outlined in Figure 1, in which the first test suggests that bias is present, but neither the test of slope differences nor the test of intercept differences is found significant. This incoherence in multiple regression model tests has been addressed by Cramer (1972).

When meaningful predictive bias is detected it would be useful to examine the individual items for bias. It may be possible to remove sources of bias by eliminating certain items from the test. Purification of the test may result in a useful, unbiased test. Various procedures exist for examining item bias and they have been discussed extensively elsewhere. Some methods for detecting item bias are based on classical test theory (e.g., Berk, 1982); others are based on modern test theory (e.g., Lord, 1980). It should be noted that research examining the usefulness of item bias indices from either theoretical base has not been without its own set of problems (Shepard, Camilli, & Williams, 1984).

## Comparison of Step-Up Versus Step-Down Approach

To illustrate the advantage of using the step-down procedure, data from a military training school were examined for possible prediction bias. The results of following the step-up procedure suggested by Bartlett et al. (1978) are presented at the top of Table 2. Note that the procedure would stop after finding no significant difference in intercepts, and so the test of slope differences would not even be recommended by those authors. The results from following the step-down procedure advocated in this paper are presented in the bottom part of Table 2. Note that in this case the intercept difference is significant, as is the slope difference.

## Conclusions

The results of the two approaches are contradictory; the step-up procedure gives no evidence of bias, while the results of the step-down procedure imply that the test is biased. However, the step-up procedure carried through to a test of slope differences would have detected those differences. The step-down procedure makes use of such differences when testing the hypothesis of intercept differences and is therefore a more appropriate test. Furthermore, it is a more powerful test of the omnibus bias hypothesis, and it has a single specifiable alpha level.

Table 2
Comparison of Step-Up and Step-Down HMR Procedures

| Procedure and Test | $R^2$ | $R^2$ Change | F |
|---|---|---|---|
| Step-Up Procedure | | | |
| Test for overall relations: | | | |
| Model 1. (Ability only) | .0753 | .0753 | 45.27** |
| Test for intercepts: | | | |
| Model 4. vs. 1. | .0765 | .0012 | .72 |
| Test for slopes: | | | |
| Model 4. vs. 2. | .0881 | .0116 | 7.25** |
| Step-Down Procedure | | | |
| Test of omnibus hypothesis of predictive bias: | | | |
| Model 1. vs. 2. | .0881 | .0128 | 4.00* |
| Test of slopes: | | | |
| Model 4. vs. 2. | .0881 | .0117 | 7.25** |
| Test of intercepts: | | | |
| Model 3. vs. 2. | .0881 | .0121 | 7.56** |

Note: The $R^2$s were based on a sample of 558 cases.

* $p < .05$.
** $p < .01$.

## References

Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannan, R. (1978). Testing for fairness with a moderated regression strategy: an alternative to differential analysis. *Personnel Psychology, 31,* 233–241.

Berk, R. A. (Ed.) (1982). *Handbook of methods for detecting test bias*. Baltimore: Johns Hopkins University Press.

Cleary, T. A. (1968). Test bias: prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement, 5,* 115–124.

Cohen, J. (1978). Partial products are interactions; partial powers are curve components. *Psychological Bulletin, 85,* 858–866.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale NJ: Lawrence Erlbaum.

Cramer, E. M. (1972). Significance tests and tests of models in multiple regression. *The American Statistician, 26,* 26–30.

Gulliksen, H., & Wilks, S. S. (1950). Regression tests for several samples. *Psychometrika, 15,* 91–114.

Kerlinger, F. P., & Pedhazur, E. (1973). *Multiple regression and behavioral research*. New York: Holt, Rinehart & Winston.

Linn, R. L. (1978). Single-group validity, differential validity and differential prediction. *Journal of Applied Psychology, 63,* 507–512.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.

Neter, J., & Wasserman, W. (1974). *Applied linear statistical models*. Homewood IL: Richard Irwin.

Pedhazur, E. (1982). *Multiple regression in behavioral research* (2nd ed.). New York: Holt, Rinehart & Winston.

Ragosa, D. (1980). Comparing nonparallel regression lines. *Psychological Bulletin, 88,* 307–321.

Ragosa, D. (1981). On the relationship between the Johnson-Neyman region of significance and statistical tests of parallel within-group regressions. *Educational and Psychological Measurement, 41,* 73–84.

SAS Institute (1979). *The SAS User's Guide* (1979 edition). Raleigh NC: Author.

Schmidt, F. L., & Hunter, J. E. (1982). Two pitfalls in assessing fairness of selection tests using the regression model. *Personnel Psychology, 35,* 601–607.

Shepard, L. A., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9,* 93–128.

## Acknowledgments

*The authors thank Stephanie Kewley and Jane Curtis for raising some issues that resulted in the ideas expressed here, and for providing the data used in the example. The authors also thank three anonymous reviewers for their helpful comments. All remaining inaccuracies are the responsibility of the authors.*

## Author's Address

Send requests for reprints or further information to Gary J. Lautenschlager, Department of Psychology, University of Georgia, Athens GA 30602, U.S.A.