The Predictive Validity of the Components of the Process of Selection for Higher Education in Israel: A Correction for Sample-Selection Bias Using Heckman's Method

A Hebrew version of this report was issued
as NITE report no. 236 (Sept. 1997)

Tamar Kennet-Cohen and Shmuel Bronner

October 1998

Summary

Studies of predictive validity are usually conducted on selected samples. In such circumstances, the question arises as to whether the sample statistics (correlation coefficients and regression coefficients) estimate correctly the population parameters prior to selection. The assumption prevalent in the professional psychometric literature is that regression coefficients do not change as a result of the selection process. Correction of the validity coefficients for the effect of the selection ("correction for restriction of range") is based on this assumption. This report explores the assumption that the regression coefficients remain unchanged by the selection process.

First, evidence is presented which demonstrates that this assumption does in fact hold in the context of selection processes for institutions of higher education. Studies carried out in the United States and in Israel provide empirical proof that correlation coefficients which were obtained in selected samples and corrected for restriction of range under the assumption that the regression equations do not change as a result of the selection, accurately reflect correlations calculated for an unselected population.

Second, we describe the set of circumstances in which the selection process may lead to obtaining biased estimates of the regression parameters. Analysis of a model describing the effect of selection (Heckman, 1979) supports the claim that the process characterizing selection for institutions of higher education tends, if anything, to lead to a *downward bias* in the estimation of the regression coefficients. This assumption is supported by the work of researchers familiar with the variables typically involved in selection processes for higher education and the relationships among these variables.

Given the a priori assumption that the selection process leads to a *downward bias* in the regression coefficients (and, therefore, also to an *underestimate* of the correlation coefficients, obtained by the restriction of range correction formulas), it is desirable to have a statistical model which permits empirical investigation of the existence and direction of bias resulting from selection, and suggests a method for correcting it. Such a model, proposed by Heckman (1979), was applied by Ayalon and Yogev (1997) in an investigation of the predictive validity of the components of the selection process for Tel Aviv University.

The main argument of the current report is that Ayalon and Yogev have applied Heckman's model erroneously, which led to their obtaining incorrect estimates for the parameters which

they wished to estimate. This study presents and supports this argument through a theoretical explication of Heckman's model. The argument is further supported by an analysis of simulated data, the results of which are reported herein. These results testify to a systematic departure of the estimates produced by Ayalon and Yogev's application of Heckman's model from the true population parameters. The study shows that if Heckman's model is applied correctly, then analyses of the simulated data provide accurate estimates of the population parameters.

Both applications of Heckman's model were used in a study of the predictive validity of the major components of the selection process (the total score on the Psychometric Entrance Test - PET, and the average of grades on the high school matriculation certificate - HSM) for applicants to Tel Aviv University for the academic year 1994-95. When the correct application was used, both predictors were found (on average, across all departments) to have positive predictive validity and a positive marginal contribution in predicting first-year grade average.

Handling Selection Bias: An Analytic Investigation

In predictive validity studies, researchers attempt to estimate the correlation coefficients between the predictors and the criterion (validity coefficients) and/or the regression coefficients for predicting the criterion from the predictor variables. Such studies are usually conducted on samples which have undergone prior selection (selected samples). In this situation, it is necessary to consider whether the statistics obtained in the sample (validity coefficients and regression coefficients) correctly estimate the population parameters. In the psychometric literature (e.g., Donlon, 1984), it is assumed that the regression coefficients remain unaffected by the selection process (this assumption is fulfilled, for instance, when selection takes place solely on the basis of the predictor of interest). However, validity coefficients do change as a result of the selection process: the coefficients obtained in a selected sample underestimate the population validity coefficients, due to restriction in the range of students' grades, which are used as the predictors. The formulas which have been developed in the professional literature to correct for restriction of range of the predictors (e.g., Guilford, 1967; Lord & Novick, 1968) are based on the assumption, stated above, that the sample regression coefficients are unbiased estimators of the population regression coefficients.

The extent to which this assumption reflects the actual reality may be *empirically* investigated by comparing validity coefficients obtained in a selected sample and corrected for range restriction, with validity coefficients obtained in the population. This type of comparison has been reported (Donlon, 1984) regarding the validity of high school record and of SAT scores for prediction of achievement in college. The average correlation between SAT scores and first-year grade average, in over two thousand studies conducted at 685 colleges, was 0.42. In the same studies, an average correlation of 0.48 was obtained between high school record and first-year grade average. Correction of the correlations for restriction of range (under the assumption that selection is done explicitly on the one predictor that is being corrected) resulted in correlation coefficients of 0.52 and 0.51 for SAT scores and high school record respectively. In order to test the degree of precision of the estimates obtained by the application of the correction formula, a separate analysis was performed on a sub-group of colleges where restriction of range did not occur. In this analysis, the average correlations with first-year grade average were 0.56 for SAT scores and 0.55 for high school record. These

values are similar (and even high in comparison) to the average correlations which were obtained as a result of applying the correction formula. Validity studies conducted in Israel – for example, at the faculties of agriculture (Ben-Shakhar & Beller, 1981) and humanities (Ben-Shakhar & Beller, 1983; Kiderman, 1987) at the Hebrew University– have also demonstrated empirically that correlations obtained in selected samples and corrected for restriction of range reflect accurately the correlations in the unselected population.

The empirical evidence presented above strengthens the assumption that the nature of the selection process is such that it does not alter the regression line of the criterion on the predictors (as mentioned above, explicit selection based on the predictors is a special case of this type of selection process). This assumption is a necessary condition for the application of the traditional approach of correction for restriction of range. However, as a rule, such empirical evidence is not available; as mentioned above, criterion scores are typically available only for individuals who have already undergone some sort of selection. Therefore, the question raised by Ayalon and Yogev (1994) regarding the validity of this assumption is largely justified. The possibility that the selection process leads to biased estimates of the regression coefficients – a phenomenon known as "sample selection bias" – is worthy of consideration. Nevertheless, we disagree with the statement made by Ayalon and Yogev (1994, p. 112) that "it is impossible to predict the direction of bias resulting from prior selection". As will be discussed next, knowledge that has been gained regarding the selection process for higher education leads to the expectation that the selection results in *downward bias* of measures of the relationship between the predictors and the criterion.

Bias resulting from prior selection may exist when the predictor one wishes to validate is subject to incidental selection, i.e., explicit selection is also based on other variables which are *unknown* or *unmeasured*. We note that a situation of incidental selection with respect to the predictor of interest does not in and of itself preclude correct estimation of regression and correlation coefficients for that predictor; when all the variables involved in the selection are known, unbiased estimates of the regression and correlation coefficients in the population may be obtained, through the use of the appropriate formulas (Gulliksen, 1950, pp. 158-172). For example, Donlon (1984) reports that if one assumes that there is explicit selection *both* on the SAT and on high school record, then the corrected validities would be 0.56 for SAT scores and 0.57 for high school record (versus 0.52 and 0.51 respectively, for the

single-predictor corrections). Obviously, these values are very similar to the average correlations which were obtained in the unselected population (0.56 and 0.55 respectively). The situation becomes more problematic, as mentioned above, when selection is based on an unknown or unmeasured variable (y_s) that is a weighted composite of variables, some of which can be observed (generally the predictors whose validation is desired) and others that are unobservable. The *a priori assumption*, based on the literature which deals with the validation of predictors used in selection decisions for higher education, is that this type of selection process causes *downward bias* in the regression coefficients of the observed predictors.

For example:

> "...the typical effect of selection is to flatten the slope and produce a concomitant increase in the intercept [in the regression of *y*, the criterion, on *x*, the predictor]. It is possible, as Levin (1972) demonstrated for the correlation between *x* and *y*, for the selection to increase the slope. But this result occurs only for combinations of correlations that are unlikely to be encountered in practice" (Linn, 1983a, p. 6).

and:

> "In practical selection situations, the regression of *y* on *x* will generally have a flatter slope and a higher intercept in the selectable population than in the unselected population" (Linn, 1983b, p. 30).

Thus, the following expectation is supported by cumulative evidence from the literature about the nature of the variables typically involved in the selection process for higher education and about the relationships among them: In a selected sample, a weaker relationship (in terms of the slope of the regression line of the criterion on the predictor) than actually exists in the population will be found. Therefore, the (single predictor) restriction of range correction formula, which assumes that the slope of the regression line in a selected sample is identical to the slope in the population, will provide an *underestimate* of the correlation in the population:

> "Corrections for range restriction that treat the predictor as the sole explicit selection variable are too small. Because of this undercorrection, the resulting estimates still provide a conservative indication of the predictive value of the predictor" (Linn, Harnisch & Dunbar, 1981, p. 66).

In addition to supporting our argument with the relevant measurement literature, it is possible, in terms of the concepts of the sample selection bias model, to demonstrate the rationale which leads to the anticipation that prior selection typically decreases the strength of the observed relationship between the predictor and the criterion, rather than increase it. Biased estimates of the coefficients of the prediction equation occur due to a correlation (in the population) between the disturbances of two equations: the *prediction equation* and the *selection equation*. The *prediction equation* is an equation which estimates the relationship between the predictors – an admissions test and/or high school record, and the criterion – i.e., first-year grade average; Berk (1983) terms this equation the "substantive equation." The *selection equation* (Berk, 1983) estimates the relation between observed selection variables – an admissions test and/or high school record, if these were used in the selection process, and additional variables if any are known – and the true, unobservable, selection variable (y_s) which is the actual basis of the selection. If a particular variable is relevant to the criterion variables of both equations, yet is omitted from both, the disturbances in the two equations may be correlated. When the effect of this variable on the two criterion variables is in the same direction, the disturbances will be positively correlated. In this situation, the estimate of the slope of the regression line of the prediction equation in a selected sample will be smaller than the slope in the population. The opposite situation will occur when the omitted variable is related to the criterion variables of the two equations in opposite directions.
We argue that it is more reasonable to expect that such an omitted variable will affect the likelihood of being included in the sample of students and success at university in the same direction and not in opposite ones. Our reasoning is as follows. If the omitted variable is related to the institutional selection decision, and, working under the assumption that the selection process aims at maximizing the expected criterion scores, we may assume that selection of candidates based on this variable will match the direction of the relationship between this variable and success in university studies. For self-selection as well, it is more reasonable to expect a positive correlation between the disturbances in the two equations. This can be seen, for example, among students having the same score(s) on the predictor(s): Those who have lower criterion scores are more likely to drop out of the sample.

In sum, the literature cited above, coupled with an attempt to demonstrate the mechanism that leads to selection bias, supports the expectation that prior selection *weakens* the observed relationship, rather than the opposite. This assumption does not rule out and even compels

empirical investigation of the phenomenon. However, results which differ from those assumed here should be regarded as surprising.

Whatever the a priori assumptions (and even in the absence of any), Ayalon and Yogev's initiative in applying Heckman's (1979) model for correction for selection bias in their investigation of the predictive validity of the selection process at Tel Aviv University is laudable. However, their findings are distorted by one major defect in their application of the model; the following discussion will be limited to this particular defect.

One assumption of Heckman's model is that the two equations – the selection equation and the prediction equation – are properly specified in the population. In other words, any variables omitted from the equations are assumed to be uncorrelated with predictors that are included in the equations. Specifically, with regard to the prediction equation, specification error exists when a predictor (for example, HSM), which is correlated with an included predictor (PET), is omitted from the equation. Such an omission leads to a biased estimate of the regression coefficient of the included predictor – e.g., the regression coefficient of PET in the simple regression will differ from its regression coefficient in the "correct" regression (a multiple regression which also includes HSM). It is true that in nonexperimental research, particularly in the social sciences, it is difficult if not impossible to formulate models which are completely free of specification error (Pedhazur, 1982). We even frequently estimate models while being aware that they contain specification error (as when we investigate the regression coefficient of PET or that of HSM as single predictors in a simple regression, or when we study their simple correlations with the criterion). However, the problem which arises in the current context is that Heckman's model is sensitive to the presence of specification error, and given certain combinations of predictors in the two equations, it "corrects" for this error. (It should be emphasized that this correction occurs, although it is not the intention of the researcher who applies the model; the researcher may be interested in estimating a model which does include such error.) As a result:

> "...there really is no way of knowing whether the correction is responding to a real sample selection bias or a pseudo-bias produced by preselection specification errors in the substantive and selection equations" (Berk, 1983, p. 396).

This sensitivity of Heckman's model to specification error requires that researchers be cautious in defining the model equations, and particularly when choosing the combination of predictors for each of the two equations. In their application of Heckman's model, Ayalon and Yogev failed to demonstrate such prudence. They defined a prediction equation with one predictor: PET or HSM. It is abundantly clear to anyone involved in this field of study that such an equation contains specification error. In general, as mentioned above, it is not uncommon to define and estimate such an equation, provided, hopefully, that one is aware of the meaning and implication of the specification error. The problem arises when one attempts to apply Heckman's model to such a prediction equation while defining a selection equation which includes predictors which were omitted from the prediction equation and are correlated with predictors that were included in that prediction equation. This is precisely what Ayalon and Yogev did in their analysis. They defined a selection equation whose predictors were HSM and the PET. The variable λ (which is constructed from the predicted values of the selection equation) is, therefore, highly correlated with the combination of HSM and PET. Adding λ to the prediction equation will alter the coefficient of the predictor which was originally included as a single predictor. This reflects a correction not only for the possible selection bias, but also for the specification error, this last correction being unintended and uncontrolled. The result of this misapplication is that the values reported by Ayalon and Yogev are not the estimates of the simple population regression coefficients of HSM or of PET, but rather a distorted result of the combination of two effects: first, a partial correction for specification error and, second, a correction for selection bias. Not only is it impossible to distinguish between these two effects, but the first occurs in contrary to the intention of the formulator of the model.

A correct application of Heckman's model in the current context involves including both variables, HSM and PET, as predictors in both the selection equation and the prediction equation. Such an application is documented in the professional literature in the context of admissions for higher education: in a study of the predictive validity of Law School Admission Test (LSAT) and the undergraduate grade point average in predicting first-year grade average in law schools (Linn & Hastings, 1984). One criticism which has been made concerning this type of application, in which the combination of predictors is the same in both equations, is certainly justified. In this situation, there is multicollinearity between the original predictors of the prediction equation and λ; high multicollinearity results in the estimators of the regression coefficients having large variances. However, such a shortcoming

of a correct application of the model is less severe than a wholly incorrect application, which estimates (albeit precisely) a parameter other than the one the researcher intended to measure. Still, the suitability of Heckman's model to the current context deserves further consideration.

Handling Selection Bias: An Analysis of Simulated Data

However convincing the analytic basis for our argument concerning Ayalon and Yogev's misleading application, the most appropriate way to test our argument is through analysis of simulated data in which the population parameters (i.e., the parameters in an unselected population) are known. Comparison of the relevant population parameters to the estimates obtained by applying Heckman's model as Ayalon and Yogev did and to the estimates obtained by the application which is suggested here, will unequivocally answer the question: which is the proper application of Heckman's model to correct for selection bias?

The simulations involved controlling three dimensions of the population parameters, as follows:

(1) The predictive validity of PET. This was either positive (the population parameter according to information gathered by the National Institute for Testing and Evaluation), or negative (the population parameter which may, apparently, be inferred from some of Ayalon and Yogev's results).

(2) The relative weights of PET and HSM as used in the selection process. A greater weight was assigned to PET than to HSM (this is the typical situation, according to the findings of Ayalon and Yogev), or vice versa.

(3) The correlation between the disturbances of the selection equation and the prediction equation. This was either positive (the population parameter which we assume, as explained above) or negative (the population parameter which is inferred from the argument that the estimates of the regression coefficients in selected samples *overestimate* the population coefficients).

Our hypothesis was that the estimation produced by a model in which the prediction equation includes just one predictor (either PET or HSM) and the selection equation includes both predictors would yield biased results, in contrast to a model in which the two equations each include both predictors. This would be the case regardless of the population parameters.

The illustrations presented below are based on simulated data for populations of 10,000 candidates. The following variables were defined for each candidate:

(1) First predictor (x_1 = 'HSM').

(2) Second predictor (x_2 = 'PET').

(3) A criterion in the selection equation (y_s).

(4) A criterion in the prediction equation (y).

For each variable, a normal distribution was defined at the candidate level with a mean of 0 and a standard deviation of 1.

The factors which were controlled in the simulation correspond to the three population parameters in which we were interested:

(1) The correlation between x_2 and the criterion y is positive/negative (in either case, the correlation is smaller, in absolute values, than the correlation of x_1 with y).

(2) x_2 has a greater/smaller weight than x_1 in the selection process.

(3) The correlation between the disturbances of the prediction equation and the selection equation ($r_{y,y_s|x,x_s}$) is positive/negative (x and x_s are the predictors in the prediction equation and in the selection equation respectively).

These three factors produce eight conditions of the population.

In every condition, a positive correlation exists between each of the predictors and y_s (the higher the score on the predictor, the greater the likelihood of being included in the sample).

In every condition, the predictors are positively correlated (r = 0.35).

The selected sample included the 50% of candidates who were highest on y_s.

Table 1 presents descriptive statistics for the candidate population.

Table 1: Descriptive Statistics of the Population

Condition	b_{y1} 1	b_{y2} 2	$b_{y1.2}$ 3	$b_{y2.1}$ 4	r_{x_1,y_s} 5	r_{x_2,y_s} 6	$r_{y,y_s\|x,x_s}$ 7
1	0.46	0.40	0.37	0.27	0.45	0.50	0.60
2	0.46	0.40	0.37	0.27	0.45	0.50	-0.60
3	0.46	0.40	0.37	0.27	0.50	0.45	0.60
4	0.46	0.40	0.37	0.27	0.50	0.45	-0.60
5	0.46	-0.20	0.60	-0.41	0.45	0.50	0.60
6	0.46	-0.20	0.60	-0.41	0.45	0.50	-0.60
7	0.46	-0.20	0.60	-0.41	0.50	0.45	0.60
8	0.46	-0.20	0.60	-0.41	0.50	0.45	-0.60

Column 1 presents the coefficient in the simple regression of y on x_1. This value remains unchanged in all the conditions.

Column 2 presents the coefficient in the simple regression of y on x_2: In Conditions 1 through 4 this coefficient is positive; in Conditions 5 through 8 it is negative.

Columns 3 and 4 present the coefficients of the multiple regression of y on x_1 and x_2 respectively. These coefficients change as the validity of x_2 (Column 2) changes.

Columns 5 and 6 present the weights of the predictors in the selection equation. In Conditions 1, 2, 5 and 6, x_2 has a higher weight; in Conditions 3, 4, 7 and 8, x_1 has a higher weight.

Column 7 presents the correlation between the disturbances of the prediction equation and the selection equation. In Conditions 1, 3, 5 and 7 this correlation is positive; in Conditions 2, 4, 6 and 8 it is negative.

The parameters one wishes to estimate are the regression coefficients in Columns 1 through 4.

In the second stage of the simulation, the coefficients as obtained in a selected sample were evaluated. These coefficients are presented in Columns 1 to 4 of Table 2. Columns 5 to 8 present the coefficients which were obtained after applying Heckman's model for correction for selection.

Estimation of the regression coefficients using Heckman's model was conducted in two stages. In the first stage, a probit analysis was performed in which the dependent variable was defined as presence or absence in the selected sample. The independent variables were x_1 and x_2. The estimates obtained in this analysis were used to calculate the variable λ (see, for example, Berk, 1983) for each observation in the selected sample. In the second stage, a multiple regression , for predicting *y*, which included the additional predictor λ besides the original predictor(s) of the prediction equation, was estimated. The combinations of the predictors of the prediction equations were: x_1 or x_2, along the lines of Ayalon and Yogev's application, presented in Columns 5 and 6; and x_1 and x_2 together in a multiple regression, presented in Columns 7 and 8.

All the analyses were performed using SAS.

Table 2: Coefficients in simple and multiple regressions for predicting y in the selected sample

Condition	Without correction for selection				Corrected for selection, with prediction equation defined as:			
					Simple regression ('incorrect model')		Multiple regression ('correct model')	
	b_{y1}	b_{y2}	$b_{y1.2}$	$b_{y2.1}$	b_{y1}	b_{y2}	$b_{y1.2}$	$b_{y2.1}$
	1	2	3	4	5	6	7	8
1	0.31	0.21	0.28	0.14	0.19	-0.13	0.39	0.28
2	0.55	0.52	0.46	0.42	0.18	-0.14	0.36	0.28
3	0.30	0.24	0.26	0.17	0.07	0.02	0.36	0.25
4	0.58	0.49	0.49	0.39	0.06	0.03	0.38	0.30
5	0.41	-0.41	0.53	-0.53	0.90	-1.11	0.63	-0.40
6	0.64	-0.12	0.70	-0.29	0.88	-1.09	0.58	-0.44
7	0.39	-0.40	0.51	-0.51	1.06	-0.88	0.64	-0.41
8	0.65	-0.15	0.72	-0.31	1.02	-0.85	0.58	-0.42

A comparison of Columns 1 and 2 in Table 2 to Columns 1 and 2 in Table 1 demonstrates the effect of the selection process on the regression coefficients. Specifically, when the

disturbances in the selection equation and the prediction equation are positively correlated (the odd-numbered conditions), the regression coefficients in the sample are smaller than those in the population. When the disturbances in the two equations are negatively correlated (the even-numbered conditions), the regression coefficients in the sample are greater than those in the population. In addition, the higher the weight of the predictor in the selection process, the larger the bias which results in the sample. This may be concluded from comparing the regression coefficients of a predictor in the situation where its weight is high and the situation in which its weight is low, all other conditions being equal. (Compare within Column 1 or 2, between Conditions 1 and 3, 2 and 4, 5 and 7 and 6 and 8.) Similar phenomena exist regarding the partial coefficients in the multiple regression (Columns 3 and 4).

The results obtained from applying Heckman's model as Ayalon and Yogev did (compare Columns 5 and 6 in Table 2 to Columns 1 and 2 in Table 1), demonstrate the systematic deviation caused by this type of application: The values which appear in Columns 5 and 6 are never close to the population values. For example, in Condition 1, the population parameters are 0.46 and 0.40, the uncorrected statistics obtained in a selected population are 0.31 and 0.21, and the estimates obtained with Ayalon and Yogev's model are 0.19 and -0.13. The deviations of the estimates obtained by the application of the incorrect model may be characterized as follows: When the predictive validity of both predictors is positive (Conditions 1 through 4), the estimates obtained for the regression coefficients of both predictors are biased downwards; when the predictive validity of one of the predictors is negative (Conditions 5 through 8), a large upward deviation is obtained in the estimation of the regression coefficient of the predictor whose validity is positive, and a large downward deviation is obtained in the estimation of the regression coefficient of the predictor whose validity is negative. The greater the weight of the predictor in the selection process, the larger is the deviation in the estimation.

In our opinion, Conditions 1 and 2 reflect parameters of the population which are compatible with the findings of Ayalon and Yogev: both predictors have positive predictive validity, and the predictor with the lower predictive validity has a larger weight in the selection process. Estimation of the model suggested by Ayalon and Yogev yields results which reproduce their findings: The estimates of the regression coefficients of the two predictors are both very low.

Also, the estimate obtained for the predictor whose predictive validity in the population is the lower of the two (a regression coefficient of 0.40 as compared to 0.46 for the other predictor, see Table 1, columns 1 and 2, Conditions 1 and 2) and whose weight in the selection process is higher (a correlation of 0.50 with the selection variable as compared to a correlation of 0.45 for the other predictor) is negative. Thus, Conditions 1 and 2 demonstrate how it is possible, in the application of the models defined by Ayalon and Yogev for the selection and prediction equations, to start from a population where both predictors have positive and even fairly high predictive validity and to obtain low and even negative estimates for their regression coefficients.

The fact that we can reproduce the findings of Ayalon and Yogev from a population where both predictors have positive and high predictive validity is a vital component in substantiating our position, namely, that Conditions 1 and 2 describe the real-life situation in which we operate. To complete our argument, we also note why the findings of Ayalon and Yogev could not be obtained from a population where one of the predictors has negative predictive validity, the other has positive predictive validity, and the predictor with negative validity has a higher weight in the selection process. This situation, which occurs in Conditions 5 and 6, is the situation which Ayalon and Yogev conclude actually exists. However, two phenomena which are not present in the data of Ayalon and Yogev characterize Conditions 5 and 6. First, the predictor whose corrected regression coefficient is negative (for example, -1.11, in Condition 5) also has a negative coefficient in the sample prior to correction (-0.41). Second, the corrected coefficient of the predictor whose predictive validity is positive (0.90, in Condition 5) is greater than its coefficient in the sample prior to correction (0.41). Thus, the findings of Ayalon and Yogev are not compatible with the population parameters, as follows: PET does not have a negative regression coefficient in samples prior to correction for selection, and the predictive validity of HSM subsequent to correction is lower than its predictive validity prior to correction.

We conclude based on the above results that an incorrect application of Heckman's method, involving the definition of a selection equation which includes predictors that were omitted from the prediction equation, leads to the obtaining of estimates which distort the population parameters prior to selection. Actually, the parameter estimates discussed by Ayalon and Yogev are not the ones that they wished to estimate. The values they reported represent the

marginal contribution of each predictor to prediction of the criterion, over and above its inclusion in a composite measure together with the other predictor (that composite measure being approximately a combination of the two predictors according to their weights in the selection process). Whatever the meaning of this measure, the estimates obtained do not express the coefficients of the predictors in simple regressions for predicting the criterion.

When the model does not suffer from specification error (i.e., its prediction equation is based on multiple regression), we may obtain precise estimates of the population regression coefficients, over all the conditions (Columns 7 and 8 in Table 2 as compared to Columns 3 and 4 in Table 1). Thus, correct definition of the model provides unbiased estimators of the partial regression coefficients in the multiple regression.

Of course, legitimate interest still exists regarding the coefficients of the simple regression of the criterion on each of the predictors. In order to avoid a situation in which applying Heckman's model also corrects for specification error, one must include the same predictor in both equations as a single predictor. (It is desirable, of course, to include additional predictors in the selection equation, on the condition that they do not have a unique contribution to y and/or are uncorrelated with x. Our discussion up to this point has assumed that such predictors, if they exist, are not available to the researcher.) The problem is that it is not correct to define the selection model based on a single predictor when the selection process is known to be based on an additional observed predictor. Such an insufficient definition of the selection process leads to the obtaining of imprecise estimates of the coefficients in the prediction equation (Gross & McGanney, 1987; Winship & Mare, 1992).

The recommended method for obtaining estimates of the simple regression coefficients for prediction of y is, therefore, to first estimate the partial regression coefficients of the two predictors in the multiple regression (as the values for Columns 7 and 8 in Table 2 were estimated). Through use of an appropriate formula (see Appendix), we may estimate the simple regression coefficients of y on each predictor separately, given the estimates of the partial regression coefficients for the multiple regression for prediction of y and the simple regression coefficients of each predictor on the other predictor (the latter are calculated directly, at the population level).

Handling Selection Bias: An Analysis of Empirical Data

The predictive validity of PET and HSM – including correction for selection bias – was calculated among applicants to Tel Aviv University for the academic year 1994-95.

Method

Subjects

The current research was based on data from 6,359 applications to Tel Aviv University for the academic year 1994-95. From these data, 14,479 observations of applicants to 37 departments were obtained (since a candidate generally applies to more than one department, the total number of observations of candidacies by departments is larger than the number of candidates). These observations constituted the candidate population. First-year grades were available for 4,658 of the observations; these observations constituted the student sample. The selection process was therefore defined as the process through which observations were selected for the student sample (according to institutional and self- selections) from among the candidate population.

Predictors

These were:

-The total score on the Psychometric Test (PET).

-The average of the grades on the high school matriculation certificate (HSM).

For each department, the scores on the predictors of the candidates were standardized to a scale having a mean of 10 and a standard deviation of 1. The establishment of identical scales for the two predictors facilitated the comparison between their regression coefficients in each of the departments.

Criterion

The criterion was first-year grade average (FYGA).

Models for Parameter Estimation

As already mentioned, the goal of the current research was to estimate the predictive validity of PET and HSM in the population of applicants. In other words, the parameters we wished to

estimate were the coefficients in a simple regression of the criterion on each of the predictors separately.

This estimation was performed using two methods which differed in their definitions of the prediction equation. In the first method (the one we recommend), the prediction equation was defined as a multiple regression in which the independent variables were PET and HSM. The simple regression coefficients for each predictor were calculated based on the estimates of the partial regression coefficients of the predictors (see formula in the Appendix). In the second method (which was applied by Ayalon and Yogev), a separate prediction equation was defined for each predictor in a simple regression which included either PET or HSM as a single predictor.

The selection equation in the two methods was identical: the dichotomous dependent variable was inclusion or non-inclusion in the sample of students, and the independent variables were PET and HSM.

Estimation Method

The estimation was performed using the statistical program LIMDEP (Greene, 1995). Similar to Ayalon and Yogev's procedure, estimation of the coefficients was performed in a one-step analysis using maximum likelihood (ML) estimation. The advantage of this method over the two-step analysis (the original method proposed by Heckman, 1979) is that the estimators of the regression coefficients obtained through this method are more efficient (i.e., the coefficients have lower standard errors). The starting values of the one-step analysis are typically the estimates obtained in the two-step process, and, in this sense, the estimates of the two-step analysis may be viewed as approximations to the ML estimates. In the empirical data used in the current research, a large degree of fit (a correlation of approximately 0.80) was found between the estimates obtained in the one- and two-step methods.

For 12 of the original 37 departments, we were unable to produce the coefficient estimates using ML estimation from the LIMDEP program, given the basic defining conditions of that program. This was the case in both the prediction equation which included two predictors and in the prediction equation including a single predictor. These departments were not included in the analyses whose results will be presented here. The omitted departments were: Physics, Chemistry, Statistics, Mechanical Engineering, Special Education, Education of the Hearing Impaired, Social Foundations of Education, Methods of Teaching and Evaluation, Theatre, Art History and Social Work. When the standardized difference between the mean scores of

the student group and the non-student group on the predictors was calculated, and the differences ranked from highest to lowest, the results for these departments were all found in the lower half. In other words, these departments are characterized by a relatively small distinction between the selected group and the non-selected group in their scores on the predictors. What this means is that in these departments, we are less able to predict the selection process from the predictors. We may assume that this characterization limits the ability to estimate parameters along the lines of Heckman's model.

Results

In Table 3, the regression coefficients of PET and HSM for a prediction equation defined as a multiple regression which includes both predictors are shown for each department. Columns 3 and 4 contain the coefficients that were obtained in the sample; Columns 5 and 6 show the ML estimates of those coefficients, i.e., the regression coefficients in the population, corrected for sample selection bias. The standard deviations of the coefficients are given in parentheses. The last row of the table contains the average coefficients over all the departments, weighted by the number of students in each department.

Table 3: Coefficients in a Multiple Regression of FYGA on PET and HSM

Department	Number of Candidates	Number of Students	Statistics in the Selected Samples		ML Estimates of the Population Parameters	
			PET	HSM	PET	HSM
	1	2	3	4	5	6
Medicine	456	82	-1.25	1.26	0.20	2.56
			(1.18)	(0.79)	(7.49)	(6.75)
Communication	237	42	0.18	0.04	5.95	4.01
Disorders			(1.14)	(0.82)	(3.02)	(1.82)
Nursing (Basic)	214	110	1.26	1.40	-0.62	2.03
			(0.71)	(0.58)	(0.91)	(0.59)
Nursing	65	46	3.45	0.96	3.42	0.96
			(1.28)	(1.11)	(8.37)	(1.41)

Department	Number of Candidates	Number of Students	Statistics in the Selected Samples		ML Estimates of the Population Parameters	
			PET	HSM	PET	HSM
	1	2	3	4	5	6
Physical Therapy	243	59	1.84	2.02	2.27	2.34
			(1.40)	(1.00)	(11.90)	(8.47)
Occupational Therapy	236	63	0.70	0.46	1.10	0.57
			(1.10)	(0.86)	(9.00)	(2.85)
Dentistry	269	28	-9.04	0.17	-21.54	-1.17
			(3.59)	(1.41)	(4.71)	(2.45)
Mathematics	460	109	2.64	7.45	3.50	3.30
			(2.53)	(2.37)	(3.33)	(5.82)
Computer Science	761	234	4.62	5.37	17.66	15.58
			(1.46)	(1.26)	(3.30)	(2.92)
Biology	573	205	3.36	6.04	6.66	7.62
			(1.25)	(1.05)	(2.37)	(1.75)
Electrical Engineering	304	117	1.43	6.85	1.53	7.26
			(1.28)	(1.46)	(4.03)	(13.78)
Industrial Engineering	303	58	3.49	4.90	-3.54	-4.09
			(1.83)	(2.23)	(2.74)	(3.66)
Humanities	1,425	915	2.19	2.52	2.19	2.51
			(0.30)	(0.29)	(0.91)	(5.04)
Guidance Counseling	219	47	0.01	0.69	3.92	1.18
			(1.26)	(1.07)	(2.88)	(1.90)
Educational Administration	151	63	0.44	0.76	0.58	1.23
			(1.01)	(1.16)	(2.66)	(7.95)
Academy of Music	40	26	-1.70	4.51	-1.86	4.80
			(1.80)	(1.82)	(9.61)	(15.17)
Film and Television	401	163	1.23	1.38	0.45	0.72
			(0.37)	(0.41)	(0.59)	(0.51)
Architecture	190	57	1.65	0.93	0.75	0.44
			(0.35)	(0.33)	(0.61)	(0.54)

Department	Number of Candidates	Number of Students	Statistics in the Selected Samples		ML Estimates of the Population Parameters	
			PET	HSM	PET	HSM
	1	2	3	4	5	6
Economics	1,255	342	2.07	2.93	-5.85	-1.81
			(0.94)	(0.70)	(1.20)	(0.90)
Political Science	638	167	0.00	3.27	11.68	2.48
			(1.03)	(0.84)	(2.48)	(1.06)
Sociology	653	137	0.29	3.01	0.31	3.02
			(0.96)	(0.97)	(31.76)	(16.18)
Psychology	695	170	0.06	4.23	-4.76	-0.98
			(1.00)	(0.88)	(1.25)	(0.95)
Accounting	593	157	2.95	2.82	4.02	3.52
			(1.00)	(0.74)	(5.19)	(3.58)
Management	786	98	-0.23	2.37	-10.19	1.87
			(1.19)	(0.62)	(2.03)	(0.78)
Law	1,128	391	1.61	2.26	1.78	2.37
			(0.47)	(0.30)	(1.57)	(1.03)
Across all Depts.	12,295	3,886	1.71	3.06	1.96	2.93

Table 4 presents the coefficients of a simple regression of FYGA on each of the predictors, PET and HSM, separately. Columns 1 and 2 present coefficients obtained in the sample. Columns 3 through 6 present estimates of the population parameters, which were calculated using two different methods. The estimates in Columns 3 and 4 were calculated using the method employed by Ayalon and Yogev: they were estimated directly as the coefficients of the predictors in the prediction equations defined as simple regressions, one equation for each variable. The estimates in Columns 5 and 6 were calculated using the method we recommend: in the first step, regression coefficients of the predictors in the prediction equation defined as a multiple regression were calculated (these estimates are shown in Table 3, Columns 5 and 6). From these coefficients, the simple regression coefficients of FYGA on each of the predictors separately were calculated (see Appendix for the formula used). The last row in the

table presents the average coefficients over all the departments, weighted by the number of students in each department.

Table 4: Coefficients in Simple Regressions of FYGA on PET or HSM

Department	Statistics in the Selected Samples		Estimates of Population Parameters			
			Prediction Equation Defined as Simple Regression		Prediction Equation Defined as Multiple Regression	
	PET	HSM	PET	HSM	PET	HSM
	1	2	3	4	5	6
Medicine	-1.70	1.46	-2.42	2.43	1.13	2.63
	(1.16)	(0.77)	(1.23)	(1.14)		
Communication	0.18	0.03	1.80	0.26	8.09	7.19
Disorders	(1.12)	(0.80)	(1.61)	(0.96)		
Nursing (Basic)	2.03	1.86	3.05	1.78	0.35	1.73
	(0.64)	(0.52)	(1.00)	(0.54)		
Nursing	4.01	2.49	4.07	1.53	3.93	2.76
	(1.10)	(1.02)	(7.16)	(1.26)		
Physical Therapy	0.18	1.25	-1.42	0.45	3.09	3.13
	(1.16)	(0.81)	(1.55)	(1.21)		
Occupational	0.54	0.31	-0.03	0.26	1.33	1.02
Therapy	(1.06)	(0.82)	(2.62)	(0.82)		
Dentistry	-9.17	1.31	-20.94	1.06	-21.84	-6.70
	(3.34)	(1.47)	(4.17)	(2.38)		
Mathematics	6.72	8.72	-5.52	-6.67	5.38	5.30
	(2.26)	(2.04)	(2.59)	(2.83)		
Computer	6.52	6.59	-6.99	-5.91	25.64	24.62
Science	(1.44)	(1.22)	(1.99)	(2.01)		
Biology	4.95	6.66	-3.63	1.31	10.65	11.10
	(1.31)	(1.04)	(1.80)	(1.51)		

Department	Statistics in the Selected Samples		Estimates of Population Parameters: Prediction Equation Defined as Simple Regression		Estimates of Population Parameters: Prediction Equation Defined as Multiple Regression	
	PET	HSM	PET	HSM	PET	HSM
	1	2	3	4	5	6
Electrical	1.10	6.76	-1.05	-1.23	4.24	7.83
Engineering	(1.39)	(1.46)	(1.83)	(2.22)		
Industrial	2.47	3.82	-2.32	-2.77	-5.77	-6.02
Engineering	(1.83)	(2.21)	(2.27)	(3.06)		
Humanities	3.75	3.80	3.66	3.52	3.79	3.91
	(0.25)	(0.24)	(0.28)	(0.30)		
Guidance	0.50	0.70	-3.69	0.75	4.65	3.59
Counseling	(1.00)	(0.85)	(1.31)	(1.53)		
Educational	0.58	0.87	0.29	0.71	1.04	1.45
Administration	(0.99)	(1.13)	(1.21)	(4.13)		
Academy	1.82	3.15	1.39	2.16	1.85	3.36
of Music	(1.21)	(1.11)	(1.39)	(1.31)		
Film and	1.68	1.86	0.46	0.71	0.87	0.98
Television	(0.36)	(0.39)	(0.55)	(0.50)		
Architecture	1.92	1.35	0.74	0.47	0.94	0.77
	(0.36)	(0.38)	(0.56)	(0.56)		
Economics	1.36	2.65	-5.40	-0.66	-6.70	-4.54
	(0.95)	(0.70)	(1.09)	(0.80)		
Political Science	1.83	3.27	-2.28	3.31	13.10	9.19
	(0.96)	(0.74)	(1.10)	(1.31)		
Sociology	1.21	3.10	-3.27	2.86	2.17	3.21
	(0.94)	(0.92)	(1.35)	(1.32)		
Psychology	-0.75	4.22	-4.49	4.14	-5.30	-3.61
	(1.05)	(0.87)	(1.20)	(1.03)		
Accounting	2.85	2.77	-0.96	1.14	5.34	5.03
	(1.04)	(0.75)	(1.36)	(0.82)		

Department	Statistics in the Selected Samples		Estimates of Population Parameters: Prediction Equation Defined as Simple Regression		Estimates of Population Parameters: Prediction Equation Defined as Multiple Regression	
	PET	HSM	PET	HSM	PET	HSM
	1	2	3	4	5	6
Management	-1.07	2.39	-26.71	2.38	-9.21	-3.46
	(1.25)	(0.60)	(4.17)	(0.57)		
Law	1.78	2.31	-1.97	0.40	2.83	3.16
	(0.50)	(0.31)	(0.55)	(0.35)		
Across all Depts.	2.41	3.51	-1.84	1.00	3.49	4.04

The following observations may be made regarding the results documented in Tables 3 and 4:

1. Comparison of the regression coefficients calculated in the sample (Columns 3 and 4 in Table 3 for the multiple regression, and Columns 1 and 2 in Table 4 for the simple regressions) and the regression coefficients which were corrected - by means of a correct application of the model - for the selection process (Columns 5 and 6 in Table 3 for the multiple regression and Columns 5 and 6 in Table 4 for simple regressions) demonstrates that the corrected coefficients are generally higher than the coefficients obtained in the sample. In other words, the coefficients calculated in the selected sample underestimate the coefficients in the population.

   The results obtained with regard to the predictive validity of the two predictors in the population are as follows: the predictive validity of each of the two predictors is positive, with the validity of HSM being somewhat higher than the validity of PET (the simple regression coefficients of the two predictors, averaged over all departments, are 4.04 and 3.49 respectively). The marginal contribution of the two predictors, HSM and PET, to the prediction of the criterion is positive (the multiple regression coefficients, averaged over all departments, are 2.93 and 1.96 respectively).
2. Applying the model as Ayalon and Yogev did reproduced their findings. As expected, a negative coefficient was obtained for PET and a small, positive coefficient was obtained for HSM (the values obtained, averaged over all the departments, were -1.84 and 1.00 respectively). However, this application of the model is incorrect. As was demonstrated in

the simulation, such an application leads to the results mentioned above, whereas the population parameters are completely different: In the population, the two predictors each have a positive and a fairly high predictive validity (and the predictor with the lower predictive validity has a greater weight in the selection process).

3. The model presented here, although clearly preferable to the model defined by Ayalon and Yogev, does possess a drawback. The variance of the regression coefficients obtained by applying this model is quite large (see the values appearing in parentheses in Columns 5 and 6 in Table 3) compared to the variance obtained when no correction is employed (Columns 3 and 4 in Table 3). The large variance results from the existence of a high correlation between λ and the predictors of the prediction equation (multicollinearity). This correlation increases as the correlation between the predictors of the prediction equation and the predictors of the selection equation increases (since λ is highly correlated with the predictors of the selection equation). Clearly, the degree of multicollinearity is especially high when the predictors of the two equations are identical, as in our application. A proper course for dealing with this difficulty would be, for example, to attempt to improve the estimation of the selection process by finding additional predictors which play a role in the selection process (and whose non-inclusion in the prediction equation does not entail a specification error). In contrast to this, Ayalon and Yogev dealt with this difficulty by defining an incorrect model, the consequences of which were more severe than the difficulty it was supposed to solve.

Discussion and Conclusion

Measures of the predictive validity (regression coefficients and correlation coefficients) of tests used for selection decisions are usually calculated in samples which have undergone selection. This situation necessitates consideration of the question of whether the measures calculated in the sample correctly estimate the parameters in the population. The common treatment of this problem in the professional literature (for example, Donlon, 1984) assumes that the regression lines (criterion on the predictors) calculated for the sample do not change as a result of the selection process. Yet it is known that correlation coefficients calculated in a selected sample are lower than correlation coefficients in the population, due to the restriction of range of scores on the predictor. The formulas developed in the professional literature for correcting the correlations for restriction of range (for example, Guilford, 1967; Lord &

Novick, 1968) are based on the assumption mentioned above: that the regression coefficients in a sample are unbiased estimators of the regression coefficients in the population.

The current work dealt with this assumption. First, we presented evidence that this assumption in fact holds in the context of selection processes for institutions of higher education. Studies conducted in the United States (Donlon, 1984) and in Israel (Ben-Shakhar & Beller, 1981, 1983; Kiderman, 1987) provide empirical evidence that the correlations obtained in selected samples, and corrected for restriction of range under the assumption that the regression coefficients do not change in the course of the selection process, accurately reflect correlations calculated for an unselected population. In other words, the available empirical evidence confirmed the assumption that the nature of the selection process for institutions of higher education is such that the regression lines of the criterion on the predictors remain unchanged. However, the existence of such empirical evidence (which is limited by its nature, since criterion data are usually available only for people who have undergone some sort of selection) does not in and of itself obviate the concern regarding the validity of the assumption under discussion.

We may suspect the existence of bias in the regression lines (calculated in the sample) of the criterion on the predictor when some of the variables underlying selection are unobservable. When these variables are unrelated to the criterion which we wish to predict (FYGA), the selection will not result in biased estimates of the regression parameters. However, when these variables are correlated with FYGA, sample statistics become biased estimates of the population parameters. The direction of bias (downward or upward) will depend on whether the relationship between the unobservable variables and the selection variable is in the same direction as the relationship between these variables and FYGA. When the direction of the relationship is the same (in the terminology of Heckman's model, a situation in which the disturbances in the selection equation and the prediction equation are positively correlated), the regression coefficients in the sample will be biased downward. When FYGA and the likelihood of being included in the sample are correlated with the unobservable variables in opposite directions (when a negative correlation exists between the disturbances of the two equations), the regression coefficients in the sample will be biased upward. Understanding the source of the bias which is liable to result from the selection process leads to the expectation that the selection process will lead to a *downward* bias in the regression coefficients

calculated in the sample, rather than to an upward bias. It makes more sense to expect that an unobservable variable involved in the selection process will be related to success in studies and to the likelihood of being included in the sample of students in the same direction, rather than in the opposite ones. The work of researchers familiar with the variables typically involved in selection processes for higher education and with the relationships among these variables (Levin, 1972; Linn, 1983a, 1983b; Linn, Harnisch & Dunbar, 1981) supports this a priori assumption concerning downward bias in the regression coefficients calculated in a selected sample.

Given the a priori assumption, which maintains that the selection process causes a downward bias in the regression coefficients (and, therefore, an *underestimation* of the correlation coefficients calculated according to the traditional formulas for correction for restriction of range), it is desirable to have a statistical method which allows empirical investigation into the existence and direction of the bias resulting from selection, and enables us to correct for it. Such a model was proposed by Heckman (1979), and Ayalon and Yogev's attempt to apply it in the investigation of the predictive validity of the components of the selection process for Tel-Aviv University is of interest.

Before turning to a discussion of the results of applying Heckman's model, it is worth noting that this model for correction for selection bias has met with some criticism. The criticism has been based both on theoretical grounds and on results of studies carried out on simulated data (a literature review, including critiques, may be found in Stolzenberg & Relles, 1990; Winship & Mare, 1992; Breen, 1996). Here we shall cite just a few problems, some particularly relevant to the current context:

1. The corrected regression coefficients are not efficient (they have a large variance). The variance of the coefficients is influenced by the correlation between λ and the predictors in the prediction equation: the larger the correlation, the larger the variance of the corrected coefficients. This correlation is in turn influenced by – among other factors – the degree of overlap between the predictors of the selection equation and the predictors of the prediction equation; the greater the overlap, the greater the correlation under discussion. This problem is especially severe, of course, when the prediction and the selection equations are based on identical combinations of predictors, as in our application. The recommended method for dealing with this difficulty – although not easy to put into

practice – is to attempt to locate additional variables related to the selection process whose non-inclusion in the prediction equation does not lead to specification error.

We note that the correlation between λ and the predictors in the prediction equation is also influenced by the selection ratio. The smaller the proportion of observations included in the sample of students, the larger the correlation. This makes application of the model problematic in departments where the selection ratio is extreme.

Another factor influencing the variance of the estimators is the degree to which the predictors in the prediction equation explain the selection process. The smaller the degree to which the selection process is explained by the predictors, the smaller will be the variance of λ. In such a situation, there will be a high collinearity between λ and the intercept in the prediction equation. The implication of high collinearity, as already mentioned, is that the variance of the estimators will be high (Winship & Mare, 1992). We should therefore strive to build a model which manages to explain the variation in the selection process, i.e., one which effectively predicts which observations are selected into the sample. It is not clear to what extent this requirement is fulfilled in the circumstances of the current research. In any event, this requirement goes well with the recommendation expressed above to locate additional variables relevant to the selection process (in particular, variables which will explain the self-selection process).

(It should be noted that the problem of multicollinearity and its implications for the efficiency of the estimators is discussed in the literature in the context of estimators obtained in the two-step analysis. However, because the two-step estimators are starting values for the iterative computation of the ML estimators, their statistical characteristics will influence the quality of the ML estimators. A discussion of the efficiency of the estimators obtained from Heckman's two-step analysis and from ML estimation may be found in Nelson, 1984.)

2. The application of Heckman's model is based on several assumptions:
   a) Homoscedasticity (constant variance) of the disturbances in the selection equation.
   b) Bivariate normality for the joint distribution of the disturbances in the prediction and selection equation.

   Violation of these assumptions is likely to have serious implications (see Maddala, 1983, with reference to heteroscedasticity and Goldberger, 1983, with reference to non-normality).
3. The corrected regression coefficients are consistent estimators; consistency is a

> characteristic relevant to large samples. Analysis of simulated data has shown that application of Heckman's model requires large samples (Gross & McGanney, 1987). Those findings showed that, in a situation in which the selection process did not produce bias (the correlation between the errors of the selection and prediction equations was set to zero), the two-step estimators and the ML estimators were less precise than the values obtained in the selected sample. This was the case given a population of 200 candidates and a sample of 100 accepted candidates.

Thus, in certain contexts, the accuracy of corrections made with Heckman's method should be questioned, and extreme care is required in applying this method. While we have noted these general reservations about Heckman's model, the following discussion will focus on one fundamental shortcoming of the manner in which Heckman's model was applied by Ayalon and Yogev.

The central problem is that Ayalon and Yogev have applied Heckman's model incorrectly. They defined a prediction equation with a specification error (based on a single predictor: PET or HSM), and a selection equation including predictors which were omitted from the prediction equation (the selection equation they defined was based on two predictors, PET and HSM). While the purpose of Heckman's model is to correct for the selection process, in such a combination of equations, it also causes an unintended and uncontrolled correction for the specification error. As a result, Ayalon and Yogev have reported incorrect values as estimates for the parameters they intended to estimate (regression coefficients in a simple regression of the criterion on each of the predictors separately). Analysis of simulated data illustrated the deviation of the obtained estimates from the true population values when the application of the model was that espoused by Ayalon and Yogev. Furthermore, that analysis showed how, from a population in which the predictive validity of both of the predictors is positive and rather high, an application similar to the one proposed by Ayalon and Yogev yields results similar to those that they reported.

A correct application of Heckman's model is based on a definition of prediction and selection equations each including both predictors: PET and HSM. Such a definition of the model was also investigated through analysis of simulated data, where it was found that this definition accurately estimates the desired parameters (the partial regression coefficients in the multiple regression for prediction of the criterion). Given the partial regression coefficients, we can

calculate the desired parameters: the regression coefficients in the simple regression of the criterion on each of the predictors separately.

Heckman's model for correcting for selection bias was applied according to the definition suggested above in an investigation of the predictive validity of PET and HSM among candidates for admission to Tel Aviv University for the academic year 1994-95. It was found that the predictive validity of the two predictors was positive (the simple regression coefficients of PET and HSM, over all departments, were 3.49 and 4.04, respectively). Each of the two predictors had a marginal positive contribution to predicting the criterion (regression coefficients of the multiple regression of PET and HSM, over all departments, were 1.96 and 2.93, respectively).
The corrected regression coefficients of both of the two predictors were higher, on average, than the regression coefficients which were calculated in the sample. In other words, the coefficients calculated in the selected sample *underestimate*, on average, the coefficients in the population.

Before concluding, we would like to address two issues: the definition of candidacy and the inclusion of departments in the study.
The results presented above were obtained under one particular definition of "who is a candidate." According to this definition (see the "Subjects" paragraph in the Method section), candidacy for admission to a department exists when the candidate marks that department on his or her application to the university (notwithstanding the preference marked for a department, or whether the candidate ultimately studies in that department). This definition of candidacy matches the definition used by the registration and admissions officers of the universities. However, we are aware that other legitimate or relevant definitions of candidacy can be formulated (for example, a definition based on candidacies for admission to departments that were marked on the application form as first or second choice only, and omitting candidacies of applicants who were accepted but did not study in the department).
The other issue pertains to the departments which were not included in the original analyses. It will be recalled that the results documented above omitted departments for which the LIMDEP program did not produce ML estimators (see "Estimation Method" in the Method section). In cases in which the program does not succeed in producing ML estimators, it provides estimators produced by the two-step analysis. By using the estimates produced by

the two-step method, it is possible to include in the analyses departments for which no ML estimators were produced.

Statistical analyses were conducted with regard to the two points described above (i.e., the analyses were repeated using other definitions of candidacy, and including departments for which ML estimators were not produced). The results of these analyses were similar to those reported above: the regression coefficients of the two predictors after correction were higher than the coefficients prior to correction. In other words, regression coefficients obtained in a sample of students provide an underestimate of the coefficients in the population. The stability of this finding over different definitions of the candidate population and departments demonstrates that these results are not a product of one manipulation or another in the composition of the data.

The results of the current research concerning the typical direction of bias resulting from the selection process is in accord with the a priori assumptions (Linn, 1983a; 1983b), and is consistent with the findings of previous research in this area (Linn & Hastings, 1984). From this finding, we may conclude that the traditional correction of the validity coefficients of the predictors for restriction of range (a correction which assumes that regression coefficients do not change as a result of the selection process), which has been adopted in the validity studies of the National Institute for Testing and Evaluation (Beller, 1994; Kennet-Cohen, Bronner & Oren, 1995), provides a conservative estimate of the correlations in the population.

Appendix

Calculation of simple regression coefficients by means of multiple regression coefficients

The calculation of simple regression coefficients, given multiple regression coefficients, is carried out according to the following formula (Pedhazur, 1982, p. 226):

$b_{y1} = b_{y1.2} + b_{y2.1}b_{21}$ where:

b_{y1} is the coefficient in a simple regression of y on x_1 only.

$b_{y1.2}$ and $b_{y2.1}$ are the partial regression coefficients obtained in a regression of y on x_1 and x_2.

b_{21} is the regression coefficient obtained from the regression of x_2 on x_1.

The relationship expressed by the formula exists at the population level. For $b_{y1.2}$ and $b_{y2.1}$, we used ML estimators of the population parameters (Columns 5 and 6 in Table 3). b_{21} was calculated directly in the population.

References

Ayalon, H., & Yogev, A. (1997).

[Causes of bias in the prediction of success in university studies] (Conference paper No. 3-97). Tel Aviv, Israel: The Pinhas Sapir Center for Development at Tel Aviv University.

Ayalon, H., & Yogev, A. (1994).

[Psychometric tests, matriculation grades, and the prediction of success in university studies: why the known is actually unknown]. *Megamot, 36*, 109-122.

Beller, M. (1994).

[Psychometric and sociological issues regarding the process of selecting candidates for Israeli universities]. *Megamot, 36*, 88-108.

Ben-Shakhar, G., & Beller, M. (1981).

[Evaluation of the selection process of students in the Hebrew University in Jerusalem]. *Megamot, 27*, 22-36.

Ben-Shakhar, G., & Beller, M. (1983). An application of decision-theoretic model to a quota-free selection problem. *Journal of Applied Psychology, 68*, 137-146.

Berk, R. A. (1983). An introduction to sample selection bias in sociological data. *American Sociological Review, 48,* 386-398.

Breen, R. (1996). *Regression models: Censored, sample selected, or truncated data.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-111. Thousand Oaks, CA: Sage.

Donlon, F. T. (Ed.). (1984). *The college board technical handbook for the Scholastic Aptitude Test and Achievement Tests*. New-York: College Entrance Examination Board.

Goldberger, A. S. (1983). Abnormal selection bias. In S. Karlin, T. Amemiya, & L. A. Goldman (Eds.), *Studies in econometrics, time series and multivariate statistics* (pp. 67-84). New York: Academic Press.

Greene, W. H. (1995). *LIMDEP.* New York: Econometric Software.

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

Guilford, J. P. (1967). *Fundamental statistics in psychology and education*. New York: McGraw-Hill.

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

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47,* 153-161.

Kennet-Cohen, T., Bronner, S., & Oren, C. (1995). *A meta-analysis of the predictive validity of the selection process to universities in Israel* (Report No. 202). Jerusalem: National Institute for Testing and Evaluation.

Kiderman, A.. (1987).

[Application of a decision-making model to a comparison of the utility of two selection systems when selection is for an unlimited number of places]. Master thesis in Psychology, The Hebrew University of Jerusalem.

Levin, J. (1972). The occurrence of an increase of correlation by range restriction. *Psychometrika, 37*, 93-97.

Linn, R. L. (1983a). Pearson selection formulas: Implications for studies of predictive bias and estimates of educational effects in selected samples. *Journal of Educational Measurement, 20*, 1-15.

Linn, R. L. (1983b). Predictive bias as an artifact of selection procedures. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A Festschrift for Frederic M. Lord.* Hillsdale, NJ: Lawrence Erlbaum.

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

Linn, R. L. & Hastings, C. N. (1984). Group differentiated prediction. *Applied Psychological Measurement, 8*, 165-172.

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

Maddala, G. S. (1983). *Limited dependent and qualitative variables in econometrics*. Cambridge: Cambridge University Press.

Nelson, F. D. (1984). Efficiency of the two-step estimator for models with

endogenous sample selection. *Journal of Econometrics, 24*, 181-196.

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

Stolzenberg, R. M., & Relles, D. A. (1990). Theory testing in a world of constrained research design. *Sociological Methods and Research, 18*, 395-415.

Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual Review in Sociology, 18,* 327-350.