Title
Selection and Predictive Validity with Latent Variable Structure

Authors
Bengt O. Muthén
Jin-Wen Yang Hsu

Publication Date
2011-10-24

British Journal of Mathematical and Statistical Psychology (1993) 46, 255–271 Printed in Great Britain

Selection and predictive validity with latent variable structures†

Bengt O. Muthén‡ and **Jin-Wen Yang Hsu§**

Graduate School of Education, University of California, Los Angeles, Los Angeles, CA 90024-1521, USA

Estimators of the predictive validity of a multifactorial test are considered. These estimators take into account the selectivity of the sample of those who have observations on the criterion measure. It is pointed out that the selectivity problem can be viewed as a missing data situation. The relationships between the classic Pearson–Lawley adjustment, regression based on factor scores, and maximum-likelihood estimation under ignorable missingness are described. The estimators are compared in a study of artificial population data.

1. Introduction

This paper considers the selection of individuals using a multifactorial test and the assessment of criterion-related validity of the selection instrument. Such selection routinely takes place in placing military personnel in various specialties with job performance as criterion and in admitting students to various schools with first-year grade point average as criterion. Most often a composite measure related to the total test score or subtests are used in such selection. We will argue, however, that the use of a multiple factor latent variable model for the observed variables comprising the test can make more efficient use of the test information. This is in line with arguments for latent variable modelling of broad and narrow abilities recently presented by Gustafsson (1988*a*). Explicit use of the latent variable model for selection may also be beneficial. Even when the latent variable model is not used for selection, it can provide more detailed information in the validation stage.

Correctly assessing the predictive validity in traditional selection studies, without latent variables, is a difficult task involving adjustments to circumvent the selective nature of the sample to be used for the validation. Adjustment for range restriction is commonly carried out by Pearson–Lawley corrections. As we will see, the use of a latent variable model produces further complications related to selection. This paper will focus on the technical issues involved in criterion-related validity assessment

†The research described in the paper was funded by the Graduate Management Admission Council, Los Angeles, USA. The GMAC encourages researchers to formulate and freely express their own opinions, and the opinions expressed here are not necessarily those of the GMAC.
‡Requests for reprints.
§Now at the Foundations of Education, University of Florida.

using a latent variable model. While the work of Gustafsson (1988*a*, *b*) dealt with latent variable modelling in predictive validity contexts, it did not address selectivity. Muthén (1989) dealt with selectivity in a latent variable model for an admissions test, but not in the context of predictive validity. In this paper, the two issues will be studied together. An efficient estimator which appears not to have been previously used in predictive validity studies will be proposed.

2. The latent variable models

Latent variable modelling of the components of a test in relation to a criterion variable provides more precise predictor variables, and may include factors which have a small number of measurements. The strength of latent variable models in identifying individual differences in both broad and narrow abilities has recently been stressed by Gustafsson (1988*a*), also reviewing related literature on the predictive value of tests and aptitude–treatment interaction. For many ability and aptitude tests it is relevant to postulate a model with both a general factor influencing all components of the test, and specific factors influencing more narrow subsets. For selection into special training programs the added information of the specific factors may be important. Consider as an artificial example the model of Fig. 1. The observed variables of the test are denoted by x and the criterion by y. The general factor is denoted as η_G and the specific factors are denoted η_{S1}, η_{S2}, and η_{S3}. On the x-side of the model, the fact that these factors are assumed to be uncorrelated means that a multivariate variance component modelling is achieved, where the relative contributions of these factors to the variances in the x variables can be studied.

The various paths from the ηs to y reflect the differential predictive power of the general factor versus the specific ones as well as among the specific ones. While the general factor may always be important, different specific factors, or combinations thereof, presumably have different importance for different selection purposes, such as different specialties in the military. If the values of the factors were known, a selection procedure might involve choosing individuals who have particularly high values on the relevant specific factor, or combination of specific factors, while maintaining a certain minimum standard with respect to the general factor.

Another example of a latent variable structure for a test used for selection is given by Muthén, Shavelson, Hollis, Kao, Muthén, Tam, Wu & Yang (1988), presenting a standard confirmatory factor model with oblique factors to describe 24 item composites created from items of the GMAC admissions test for graduate school of business and management, the GMAT. The estimated, simple structure five-factor model for a sample of 55 279 test takers is given in Table 1.

Formally, a latent variable model used to predict a criterion variable may be written as

$$\mathbf{x} = \upsilon + \Lambda\eta + \varepsilon \tag{1}$$

$$\mathbf{y} = \alpha_y + \beta'\eta + \xi. \tag{2}$$

In (1) $\mathbf{x}$ is a vector of p test variables, Λ is a $p \times m$ matrix of factor loadings, η is the

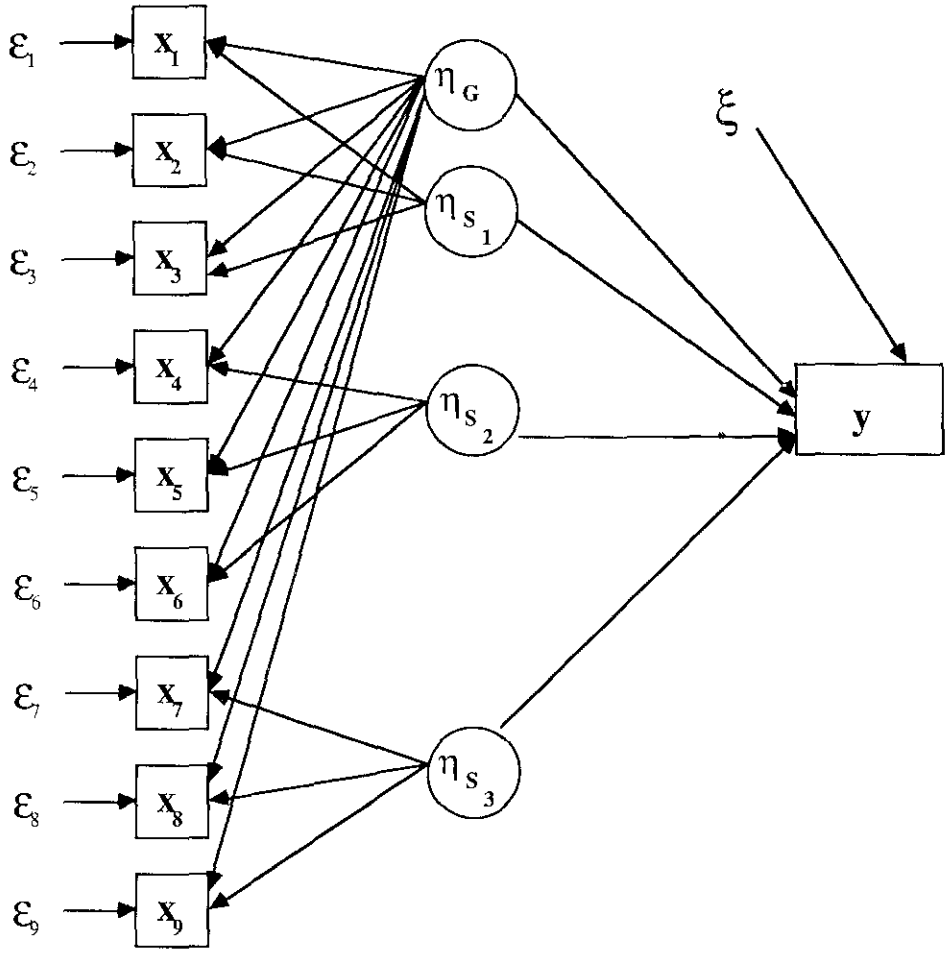

Figure 1. Artificial latent variable model.

vector of m factors, ε is a vector of p measurement error variables; in (2) α_y is an intercept parameter, β is a vector of m slopes, and ξ is a residual. Let $E(\eta)=\alpha$, $V(\eta)=\Psi$, $V(\varepsilon)=\Theta$, and $V(\xi)=\psi_y$. With ordinary assumptions,

$$\Sigma_{xx}=\Lambda\Psi\Lambda'+\Theta, \tag{3}$$

$$\sigma_{xy}=\Lambda\Psi\beta, \tag{4}$$

$$\sigma_{yy}=\beta'\Psi\beta+\psi_y. \tag{5}$$

Let

$$\Sigma_{zz}=\begin{bmatrix}\sigma_{yy}\sigma'_{xy}\\ \sigma_{xy}\Sigma_{xx}\end{bmatrix} \tag{6}$$

We note that this model assumes that the factors of η and the residual ξ are the

Table 1. Standardized estimates for GMAT simple structure factor model ($N = 55279$)

		Factor				
		Verbal			Quantitative	
		General	Specific		General	Specific
Section		Sentence Corr. & reading comp.	Minor key	Other key	Accuracy	Speed & accuracy
Verbal						
Sentence correction	V1	.588				
	V2	.599				
	V3	.625				
	V4	.656				−.024
Analysis of situations	V5	.044	.598			
	V6		.662			
	V7	.356	.454			−.002
	V8	−.044		.647		
	V9			.765		
	V10	.076		.603		.109
Reading comprehension	V11	.573				
	V12	.630				
	V13	.622				
	V14	.597				.080
Quantitative problem solving 1	Q1	−.123			.711	.084
	Q2	−.047			.525	.335
	Q3	.046			−.148	.838
Data sufficiency	Q4				.635	
	Q5	.119			.492	.034
	Q6	−.026			.244	.530
	Q7	.178			−.088	.657
Problem solving 2	Q8	−.054			.739	−.071
	Q9	−.026			.645	.218
	Q10					.846
Factor correlations		1.000				
		.577	1.000			
		.668	.674	1.000		
		.583	.620	.557	1.000	
		.268	.411	.433	.687	1.000

Note. Empty entries in the factor loading matrix correspond to elements fixed at zero.

only relevant predictors of the criterion y. The model is misspecified whenever omitted predictors of y are correlated with η or ε (or correlated with $\mathbf{x}$).

3. Selection issues

In standard predictive validity studies, using observable predictors $\mathbf{x}$, it is clearly recognized that the assessment of validity must take selectivity into account. Observations on y from a random sample of the population of those who took the

test are not available, but only observation from those who were selected. These observations may be viewed as a random sample from a selected population. We will call the sample observations the matriculant sample corresponding to a matriculation population. Given estimates from a matriculant sample, the classic Pearson–Lawley selection formulas are frequently used for adjustments to avoid bias and make the inference to the test taker population more appropriate. Given a random vector $\mathbf{z}$ and a vector of selection variables $\mathbf{s}$, these formulas give the relation between the test-taker population and the matriculant population (indicated by asterisks)

$$\mu_z^* = \mu_z + \mathbf{B}(\mu_s^* - \mu_s), \tag{7}$$

$$\Sigma_{zz}^* = \Sigma_{zz} + \mathbf{B}(\Sigma_{ss}^* - \Sigma_{ss})\mathbf{B}', \tag{8}$$

$$\Sigma_{zs}^* = \mathbf{B}\Sigma_{ss}^*, \tag{9}$$

where $\mathbf{B} = \Sigma_{zs}\Sigma_{ss}^{-1}$ (see for example Johnson & Kotz, 1972). These relations provide the corrections needed to obtain Σ_{zz} from Σ_{zz}^*, namely the 'Pearson–Lawley correction' formula

$$\Sigma_{zz} = \Sigma_{zz}^* - \Sigma_{zs}^*\Sigma_{ss}^{*-1}(\Sigma_{ss}^* - \Sigma_{ss})\Sigma_{ss}^{*-1}\Sigma_{zs}^{*\prime}. \tag{10}$$

These formulas assume that the regression of $\mathbf{z}$ on $\mathbf{s}$ is linear and homoscedastic. Given an estimate of Σ_{zz}^* based on a random sample of the selected population, (10) shows that an estimate of Σ_{zz} is obtained as soon as estimates of Σ_{ss}^*, Σ_{ss} and Σ_{zs}^* are available. Estimates of the latter three matrices can be obtained, for example through a suitable, large reference sample and knowledge about the selection procedure. In many instances, the selection process is not exactly known and $\mathbf{x} = \mathbf{s}$ is assumed as an approximation. It is well known that when the true selection process is such that $\mathbf{x} = \mathbf{s}$, the selective nature of the sample does not cause bias in the regression of y on $\mathbf{x}$. The resulting incidental selection on the dependent variable y is fully accounted for by the exogenous variables of $\mathbf{x}$. Here, Pearson–Lawley corrections are needed only to avoid bias in the correlations between y and the xs due to the restriction of range in $\mathbf{x}$.

Consider now a latent variable model for the set of xs. The latent variable modelling presented in the previous section can be estimated with standard covariance structure techniques, see for example Muthén (1984, 1987). As before, however, standard application of such modelling will give biased results vis-à-vis the population of test takers if applied to the matriculant sample. We will study this bias by considering the following three kinds of selection procedures. In each case we assume a univariate $\mathbf{s}$ that is linearly related to a set of variables that influence the selection, either in the form of a linear regression or deterministically.

3.1. Selection based on the factors η

Consider first selection determined by the latent variables of η as it affects the $\mathbf{x}$-part

of the model, Σ_{xx}. Although in practice η is unknown, this case is of practical importance for two reasons. First, an approximation of η may be obtained for each individual in terms of estimated factor scores. Second, this case includes the situation where selection is based on predictors of η, where the predictors have no direct effects on **x** but affect **x** only indirectly via η. This case was for example studied by Meredith (1964) and Muthén & Jöreskog (1983), who pointed out that this type of selection does not distort the factor model structure per se, but leaves ν, Λ, and Θ unchanged with a change in $E(\eta)=\alpha$ and $V(\eta)=\Psi$. The factor variables of η act as exogenous variables in the **x**-part of the model. We note that this brings the results in line with the standard regression situation with observed exogenous variables, where the regressions are unaffected by selection on the exogenous variables, while the exogenous variable distribution is naturally affected. Adding y to the variables considered, it is clear that this extra variable may be viewed as one more measurement of the factors. This reinterprets the slopes βs as loadings and the residual variance of ψ_y as measurement error variance and it follows that these parameters are unchanged. Hence, if selection was based on the factors of η, perhaps as approximated by estimated factor scores, the regression of y on η would not be distorted. However, the correlations between y and the ηs are biased due to restriction of range in η. Regular Pearson–Lawley corrections could again be carried out for the estimated correlations among these variables.

3.2. Selection based on the observed predictors **x**

Consider next selection determined by the observed variables of the test **x**. This covers both cases where selection is made on the test alone and where selection is in addition based on variables not observed (or not entered into the model) which are uncorrelated with **x** and y. This situation was studied in Muthén (1989), where it was pointed out that a distorted factor structure for **x** results in the matriculant population.

3.3. 'Non-ignorable' selection

Consider last the case of selection determined by **x** and other unobserved variables correlated with the residual ξ. This covers cases where the test is an important factor in the selection but other variables such as high school graduation, grade point average, and the like are important. When such other variables can be assumed to influence both selection and y, correlations with ξ arise. This may be the most common selection situation. This implies selection that is directly related to all endogenous variables of the model and it follows that distortion of all parts of the model will arise.

4. Estimation

Let us now focus on the estimation of the latent variable model of **x** and y. We will assume that information on a random matriculant sample is available as well as a

random sample of test takers, part of whom may be in the matriculant sample. Four estimation procedures will be mentioned.

4.1. Using the matriculant sample (the LQL estimator)

Consider first the straightforward estimation using only the matriculant sample. From the discussion of the three selection situations we conclude that unless selection is determined by η, the test taker population model will not be correctly estimated. This estimator will be called LQL (see 4.4).

4.2. The factor score estimator

A second estimation approach is as follows. Given that the ultimate interest is in assessing the predictive strength of the factors of η, it may be natural to attempt the use of estimated factor scores as proxies for η and regress y on these proxies for η, and other predictors. The regression method of factor score estimation (see, e.g., Lawley & Maxwell, 1971, p. 109) takes the estimated η, $\mathbf{f}$ say, as

$$\mathbf{f} = \alpha + \Psi\Lambda'\Sigma_{xx}^{-1}(\mathbf{x} - \nu - \Lambda\alpha). \tag{11}$$

Here it is assumed that Σ_{xx} follows the factor analysis model of (3). In the case of a random sample it is well known (Tucker, 1971) that with the regression method of estimating factor scores, regressing y on $\mathbf{f}$ and other predictors would give consistent estimates of these regression coefficients. While sample covariances between y and $\mathbf{f}$ and the sample covariance matrix of $\mathbf{f}$ are in this case both inconsistent estimators of the corresponding population quantities, these biases cancel out in the regression coefficients.

In the present case of a selective sample, it is clear that the case of selection based on $\mathbf{x}$ discussed in Section 3.2 will still give consistent estimates of the regression coefficients. The estimated factor scores are simply a linear transformation of the observed predictors, and selection based on observed predictors does not bias the regression. Note that this assumes that the model parameters are known or are estimated from large enough samples to be considered essentially non-stochastic. The unbiasedness is clearly seen when considering the case of $s = \mathbf{w}'\mathbf{x}$, where $\mathbf{x}$ is the set of test variables. This will now be shown.

In the matriculant population the covariance matrix of $\mathbf{f}$ is

$$\Sigma_{ff}^{*} = \Psi\Lambda'\Sigma_{xx}^{-1}\Sigma_{xx}^{*}\Sigma_{xx}^{-1}\Lambda\Psi, \tag{12}$$

where by (8)

$$\Sigma_{xx}^{-1}\Sigma_{xx}^{*}\Sigma_{xx}^{-1} = \Sigma_{xx}^{-1} + \omega\mathbf{w}\mathbf{w}'. \tag{13}$$

The covariances between $\mathbf{f}$ and y in the matriculant population are

$$\mathrm{Cov}(\mathbf{f}, y)^{*} = \Psi\Lambda'\Sigma_{xx}^{-1}\,\mathrm{Cov}(\mathbf{x}, y)^{*}, \tag{14}$$

where by (12)

$$\begin{aligned}\mathrm{Cov}(\mathbf{x}, y)^* &= \Lambda\Psi\beta + \omega\Sigma_{xx}\mathbf{w}\mathbf{w}'\Lambda\Psi\beta \\ &= \Sigma_{xx}(\Sigma_{xx}^{-1} + \omega\mathbf{w}\mathbf{w}')\Lambda\Psi\beta. \end{aligned} \tag{15}$$

Collecting terms we find that the regression of y on $\mathbf{f}$ is unbiased,

$$\begin{aligned}&\mathrm{Cov}\,(\mathbf{f}, y)^{*\prime}\Sigma_{ff}^{*-1} \\ &= \beta'\Psi\Lambda'(\Sigma_{xx}^{-1} + \omega\mathbf{w}\mathbf{w}')\Lambda\Psi[\Psi\Lambda'(\Sigma_{xx}^{-1} + \omega\mathbf{w}\mathbf{w}')\Lambda\Psi]^{-1} \\ &= \beta'. \end{aligned} \tag{16}$$

The factor score estimation approach would use the test taker sample to estimate the parameters involved in (11), compute $\mathbf{f}$ for the matriculant sample, and compute the regression of y on $\mathbf{f}$ and other predictors from the matriculant sample. A standardized solution can be obtained by using the covariance matrix for the predictors estimated from the test taker sample, which would avoid the biases of using the matriculant sample. For example, for $\mathbf{f}$ the test taker estimate of Ψ would be used.

4.3. The Pearson–Lawley estimator

A third estimation approach is as follows. One may use the matriculant sample to estimate Σ_{zz}^* as the matriculant sample covariance matrix. Using sample information from the test takers and assuming that selection is determined by $\mathbf{x}$, a Pearson–Lawley correction can then be made to this covariance matrix to obtain an estimate of the test-taker population Σ_{zz}. The latent variable model may then be fitted to the estimated Σ_{zz} using standard covariance structure methods.

4.4. Maximum likelihood under ignorability (the FQL estimator)

A fourth, and more refined estimator is available. For this estimator we note that the test-taker and matriculant samples may be viewed as an example of missing data; test takers that are not among the matriculants may be viewed as having missing data on y. Missing data theory is discussed in Little & Rubin (1987). It is shown that under the assumption of 'ignorability' of the missing data mechanism, correct maximum likelihood (ML) estimation can be provided by using the matriculant and test-taker sample information jointly in the estimation. In our formulation of the three selection situations, ignorability is obtained when the missingness on y can be predicted by $\mathbf{x}$, but not in the other two cases. For purposes of estimating the model parameters, the log likelihood of the sample can then be simplified as

$$\log L = \sum_{i=1}^{N} \log f(\mathbf{x}_i) + \sum_{i=1}^{Nm} \log f(y_i | \mathbf{x}_i), \tag{17}$$

where N is the total sample size of the test-taker sample, Nm is the part of the test-

taker sample that constitutes the matriculant sample, and f represents various densities assumed to correspond to the multivariate normal distribution.

Consider equation (17) for the case of no latent variable structure on $\mathbf{x}$ with y regressed on $\mathbf{x}$. The second term on the right-hand side of (17) contains the regression parameters, while the first term contains the parameters of the marginal distribution of $\mathbf{x}$. It follows that, under ignorability, proper ML estimation of regression parameters is obtained using only the matriculant sample and that the test-taker sample provides ML estimation of the parameters of the $\mathbf{x}$ distribution.

The maximum-likelihood theory under ignorability presented by Little & Rubin (1987) gives the following maximum-likelihood estimators when the latent variable structure is not imposed ('unrestricted' case) and selection is determined by $\mathbf{x}$. While Σ_{xx} is estimated by the test taker sample covariance matrix S_{xx} we have the ML estimates for the unrestricted case

$$\hat{\sigma}_{yx} = s^*_{yx}\mathbf{S}^{*-1}_{xx}\,\mathbf{S}_{xx}, \tag{18}$$

$$\hat{\sigma}_{yy} = s^*_{yy} - \mathbf{s}^*_{yx}\mathbf{S}^{*-1}_{xx}(\mathbf{S}^*_{xx} - \mathbf{S}_{xx})\mathbf{S}^{*-1}_{xx}\mathbf{s}^{*\prime}_{yx}, \tag{19}$$

where the asterisk refers to statistics obtained from the matriculant sample of size Nm. The first term on the right-hand side of both (18) and (19) is the estimate obtained by the matriculant sample alone while other terms represent the corrections needed to obtain ML estimates. We note then that sample information on $\mathbf{x}$ for individuals who do not matriculate is actually used to obtain ML estimation of both $\mathbf{x}$- and y-related parameters.

Comparing equations (18), (19) with (9), (10) shows that the Pearson–Lawley estimation of Σ_{zz} in the third estimator scheme that we suggested above is actually the same as the ML estimator of (18) and (19) when using selection based on $\mathbf{x}$.

Muthén, Kaplan & Hollis (1987) discussed related missing data issues applied to latent variable structural equation modelling. Muthén *et al.* also described an analysis technique which is directly applicable to our situation and uses existing structural modelling software that handles mean structures in conjunction with covariance structures. Hence, this approach can be used to go beyond the unrestricted estimation above and to apply the latent variable structure in a 'restricted' analysis that uses the hypothesized factor model. The idea is based on rewriting the log likelihood equation of (17) as

$$\log L = \sum_{i=1}^{Nm} \log f(\mathbf{x}_i, y_i) + \sum_{i=Nm+1}^{N} \log f(\mathbf{x}_i). \tag{20}$$

While equations (17) and (20) are algebraically equivalent, (20) has the interesting implication that standard multiple group structural modelling can be used for the estimation. The two terms of the right-hand side of (20) correspond to the test takers who do matriculate and the test takers who do not matriculate, respectively. A simultaneous analysis of these two groups, with different number of observed variables in the two groups and across- group equality restrictions on common parameters yields ML estimates of the latent variable model parameters. Muthén *et al.* describe how to set up this analysis using the LISCOMP program (Muthén,

1987). They also show how the model can be tested against the alternative of an unrestricted covariance matrix for **x** and y.

It is interesting to note that the LQL estimation scheme that we discussed in 4.1, estimating the latent variable model from matriculants only, corresponds to using only the first term on the right-hand side of (20). The selection bias of this estimator may then be rephrased as follows. When a latent variable structure is imposed on the marginal distribution of **x**, as opposed to the situation of an ordinary regression of y on **x**, the second term must also be included in order not to obtain biased estimates. While the ordinary regression case avoids bias in the regression parameters, this way of estimating the latent variable model does not avoid bias in any of the model's parameters.

Muthén *et al.* termed the estimation approach based on both terms of (20) estimation by FQL (full, quasi likelihood) to emphasize that it may be used also when ignorability is not at hand, in which case true maximum likelihood estimation is not obtained. They showed that in cases of non-ignorability the FQL estimator compares favourably with the standard listwise present estimator, LQL (listwise, quasi likelihood). The LQL estimation is the same as the first estimation scheme that we have discussed, using only the matriculant sample, or the first term of (20).

4.5. Comparing estimators

It is interesting to note the differences in assumptions behind the Pearson–Lawley corrections and ignorability of the missing data approach used for the FQL-estimator as they relate to our three simple selection schemes in sections 3.1, 3.2, and 3.3. Pearson–Lawley corrections build on the assumption of the **x** and y being linearly related to the selection variables s with homoscedasticity in these regressions. Ignorability assumes that conditional on **x**, y and s are independent. Assume for simplicity that all variables of the model, including s, have a multivariate normal distribution. In our first selection scheme, with selection determined by η, the assumptions behind Pearson–Lawley corrections are fulfilled, but not the assumptions behind ignorability. In our second selection scheme, with selection being determined by **x**, the assumptions behind both Pearson–Lawley and ignorability are fulfilled. In our third selection scheme, assuming selection determined by **x** and variables correlated with the residuals ξ in the y regression, the assumptions behind Pearson–Lawley are fulfilled while those behind ignorability are not. The advantages of Pearson–Lawley are, however, to some extent lost if the selection process is not known and selection on **x** has to be assumed as is usually the case.

It is of interest to compare all four of the estimators discussed: LQL, the factor score approach, analysis of the Pearson–Lawley corrected estimate of Σ_{zz}, and FQL. The LQL estimator has known biases but it is of interest to see how large these biases will be since the use of the matriculant sample only corresponds to a very common way of doing validations studies. The factor score approach gives unbiased estimates as long as the ignorability assumption holds, since the incidental selection on y is then fully accounted for by the predictors. The third estimation technique uses the Pearson–Lawley assumptions to correct the biases of LQL while the fourth

technique, FQL, uses the ignorability assumption. If both assumptions are true, the maximum-likelihood property of FQL makes it the most efficient estimator. If only the Pearson–Lawley assumptions are true or if none of the two sets of assumptions is true, FQL's advantage may be lost. We will illustrate the biases of these estimators in a population study that builds on the latent variable models discussed in Section 2. Further understanding of the comparative behaviour of the estimators can be obtained by a Monte Carlo study to investigate sampling variability with various forms of violations of the two sets of assumptions. This is beyond the scope of this paper.

5. Examples

Two examples will be studied. They correspond to the two latent variable measurement structures of Section 2. The first model involves a general factor and three orthogonal specific factors, while the second model involves five oblique factors. Two of the three types of selection discussed in Section 3 will be illustrated. For the first model, selection based on **x** will be illustrated, where **x** contains the set of test variables for which the latent variable model holds. For this type of selection, the factor score and Pearson–Lawley estimation approaches will give the same results as FQL since a population study is carried out and the assumptions are fulfilled for all these estimators. For the second model we illustrate selection based on a variable which is a function of not only the test variables but also another observed predictor, as well as an unobserved component which is correlated with the residual in the prediction equation of y. In this case, each of the estimators is in violation of its assumptions. These two examples will give a rough indication of how the estimation procedures compare in terms of large-sample bias.

5.1. Example 1

Consider the latent variable model represented by Fig. 1. Assume the test-taker population values for the parameters of this model given in the leftmost column of Table 2. The measurement parameters of the factor model for the **x** distribution represent a set of highly reliable indicators of the factors. The variance component interpretation of the model can be explicated as follows. The general factor accounts for 59–76 per cent of the variation in each test variable, while the specific factors account for 24–41 per cent. All factors are uncorrelated. The criterion variable regression on the factors has a population R^2 of 50 per cent. The true population slopes are all .4, corresponding to standardized values (unit variances for the ηs and for y) of .535 for the general factor and .267 for each of the three specific factors. Hence, we assume that the general factor is the most important one in predicting y. Measurement intercept values of zero and factor means of zero were chosen.

Given this model and its parameter values, the population mean vector μ_z and covariance matrix Σ_{zz} were created for **x** and y. Selection based on an unweighted sum of the nine x variables was assumed, so that the assumption of ignorability holds. Inclusion in the matriculant population corresponded to exceeding a threshold

Table 2. LQL estimates for an artificial latent variable model

Parameter	Parameter value	Estimate 50% selection	25% selection	10% selection
		Measurement parameter loadings		
$\lambda(1,1)$	0.600	0.256	0.148	0.017
$\lambda(2,1)$	0.700	0.329	0.219	0.087
$\lambda(3,1)$	0.800	0.403	0.290	0.149
$\lambda(4,1)$	0.900	0.477	0.363	0.215
$\lambda(5,1)$	0.600	0.255	0.146	0.013
$\lambda(6,1)$	0.700	0.328	0.217	0.087
$\lambda(7,1)$	0.800	0.400	0.284	0.146
$\lambda(8,1)$	0.900	0.475	0.357	0.212
$\lambda(9,1)$	0.600	0.254	0.141	−0.003
		Regression parameters estimate		
Parameter	Parameter value	50% selection	25% selection	10% selection
		Raw solution		
β	0.400	0.153	0.084	0.001
$\beta(1)$	0.400	0.358	0.336	0.312
$\beta(2)$	0.400	0.357	0.334	0.308
$\beta(3)$	0.400	0.356	0.330	0.303
$\psi(y)$	0.280	0.289	0.294	0.297
		Standardized solution		
β	0.535	0.240	0.135	0.001
$\beta(1)$	0.267	0.279	0.268	0.246
$\beta(2)$	0.267	0.279	0.268	0.245
$\beta(3)$	0.267	0.279	0.267	0.244
$\psi(y)$	0.500	0.709	0.767	0.820

on s, where s was assumed to be normally distributed. Thresholds corresponding to the matriculant population consisting of the upper 10, 25 and 50 per cent of the s distribution were used. In each case, a matriculant population mean vector μ_z^* and covariance matrix Σ_{zz}^* were also created in line with the Pearson–Lawley formulas of (7) and (8). A non-matriculating test-taker population was also created in each case corresponding to the complementary group of those who did not exceed the threshold on s.

Analyses using the FQL and LQL estimators were then carried out on these population mean vectors and covariance matrices. We note again that the Pearson–Lawley and factor score approaches would give the same results as FQL. For FQL, equation (20) postulates the use of the $\mathbf{x}$, y mean vector and covariance matrix for matriculants (see the first term) and the $\mathbf{x}$ mean vector and covariance matrix for the non-matriculating test takers (see the second term). For LQL only the matriculant covariance matrix need be used. In each case the model was tested against the unrestricted alternative to obtain a chi-square test of model fit. The analyses were carried out by the ML estimator of the LISCOMP program.

For FQL the chi-square test of fit indicated perfect fit and the estimates were

identical to the population values in all cases. This verifies the consistency of the FQL estimator in this case where ignorability holds. For LQL non-zero chi-square test values indicated that the model does not hold for the matriculant covariance matrix. The LQL estimates are given in Table 2 above.

The measurement parameter estimates show a sharp decrease in LQL estimated reliability for the x variables. While the estimated loadings decrease with increasing selectivity the measurement error variances and factor variances remain constant at the true population values (and are therefore not reported). The estimates of primary interest in a validity study are those of the regression of y on the ηs and these are given both in raw and standardized form in Table 2. The β coefficient without a subscript refers to the slope for the general factor. We note from the standardized solution that the selection on the sum of xs induces a strong reduction of the importance of the general factor, so that the specific factors incorrectly appear as the more powerful predictors. This is natural since the total test score is a proxy for the general factor and its variation is strongly reduced by the selection. It is interesting to note that in contrast the standardized slopes for the specific factors decrease much less dramatically as the selection percentage decreases. The unexplained portion of the variation in the criterion variable is overestimated and the overestimation increases with increasing degree of selectivity.

5.2. Example 2

Consider next the measurement model of Table 1 corresponding to the five-factor structure of the GMAT test (cf. Muthén, 1989). In our present example this measurement model will be augmented by another observed predictor variable which may be thought of as undergraduate grade point average (UGPA). The values of the correlations between the five factors of the GMAT and UGPA are taken from previous analyses by Muthén *et al.* (1988). The GMAT and UGPA scores are used for selection into MBA programs and it is assumed that these variables predict first-year grade-point averages (FYA). Selection is however influenced by many other factors. We will consider the model for selection and FYA prediction indicated in Fig. 2. Here, the selection variable s is influenced by each of the test variables, by UGPA, and by a residual δ. The residual δ is taken to be correlated with the FYA residual ξ, to illustrate the influence of left-out variables which influence both FYA and s and which are correlated. It is assumed that FYA is linearly related to a weighted sum of the five factors and UGPA. The R^2 in FYA is taken to be 50 per cent. The factors and UGPA are taken to have variance one. UGPA is taken to have the weight .4, so that its direct effect is contributing to 16 per cent of the FYA variance. The weights of the factors are taken to be equal. The selection variable s is taken to be linearly related to a weighted sum of the 24 test variables and UGPA. The weights of the test variables are chosen so that the total contribution to s corresponds closely to the GMAT test (see also Muthén, 1989). The weight of UGPA is chosen to be about half of that of the GMAT test (this is in line with findings in Muthén, Hollis, Muthén & Tam, 1991). Two R^2 values for s are used, 50 per cent and 75 per cent. The residuals in the s and FYA equations are taken to have a

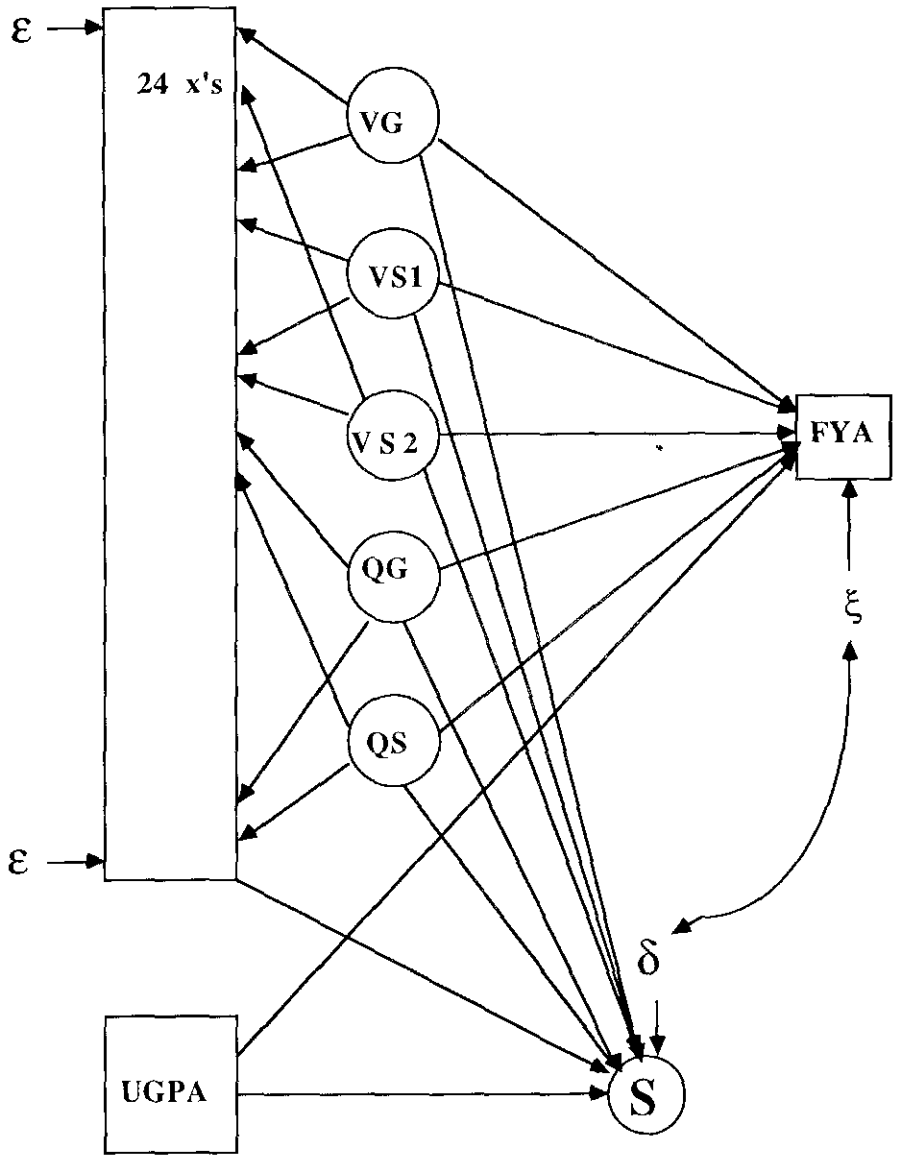

Figure 2. Selection and FYA prediction.

correlation of .25. A correlation of .50 is also studied. To give a more realistic picture, a distortion in the factor model is also introduced. The population values are here generated by the sample covariance matrix for the 24 test variables analysed in Muthén *et al.* (1988), while the estimators assume the simple factor structure of Table 1. Since the model fit is not perfect, although quite good, a model misspecification is introduced. This also has the advantage of avoiding artificial agreement in the results of some of the estimators due to using a population study with a correct model.

Table 3. Parameter estimates by FQL, LQL and FS. R^2 in s is .50, 25% selection

Structural	Parameter	Estimator FQL (%)	LQL (%)	FS (%)
		Raw solution		
Residual	.500	.484 (−03.20)	.483 (−03.40)	.496 (−00.80)
B for VG	.116	.084 (−27.69)	.077 (−33.62)	.083 (−28.45)
B for VS1	.116	.106 (−08.62)	.098 (−15.52)	.106 (−08.62)
B for VS2	.116	.107 (−07.76)	.099 (−14.66)	.110 (−05.17)
B for QG	.116	.098 (−15.52)	.088 (−24.14)	.099 (−14.66)
B for QS	.116	.095 (−18.10)	.091 (−21.55)	.094 (−18.97)
B for UGPA	.400	.343 (−14.25)	.348 (−13.00)	.343 (−14.25)
		Standardized solution		
Residual	.500	.570 (14.00)	.674 (34.80)	.688 (37.60)
B for VG	.116	.092 (−20.69)	.091 (−21.55)	.107 (−07.76)
B for VS1	.116	.116 (00.00)	.115 (−00.86)	.127 (09.48)
B for VS2	.116	.117 (00.86)	.117 (00.86)	.136 (17.24)
B for QG	.116	.106 (−08.62)	.104 (−10.34)	.128 (10.34)
B for QS	.116	.103 (−11.21)	.107 (−7.76)	.123 (06.03)
B for UGPA	.400	.372 (−07.00)	.360 (−10.00)	.476 (19.00)

Table 4. Parameter estimates by FQL, LQL and FS. R^2 in s is .75, 25% selection

Structural	Parameter	Estimator FQL (%)	LQL (%)	FS (%)
		Raw solution		
Residual	.500	.490 (−02.00)	.489 (−02.20)	.501 (00.20)
B for VG	.116	.077 (−33.62)	.069 (−40.52)	.076 (−34.48)
B for VS1	.116	.104 (−10.34)	.091 (−21.55)	.104 (−10.34)
B for VS2	.116	.106 (−08.62)	.094 (−18.97)	.108 (−06.90)
B for QG	.116	.093 (−19.83)	.080 (−31.03)	.095 (−18.10)
B for QS	.116	.090 (−22.41)	.085 (−26.72)	.089 (−23.28)
B for UGPA	.400	.330 (−17.50)	.341 (−14.75)	.330 (−17.50)
		Standardized solution		
Residual	.500	.593 (18.60)	.757 (51.40)	.772 (54.40)
B for VG	.116	.085 (−26.72)	.086 (−25.86)	.108 (−06.90)
B for VS1	.116	.114 (−01.72)	.113 (−02.59)	.138 (18.97)
B for VS2	.116	.117 (00.86)	.116 (00.00)	.149 (28.45)
B for QG	.116	.103 (−11.21)	.100 (−13.79)	.137 (18.10)
B for QS	.116	.099 (−14.66)	.106 (−08.62)	.129 (11.21)
B for UGPA	.400	.363 (−09.25)	.342 (−14.50)	.509 (27.25)

Tables 3 and 4 give the estimates from the three estimators LQL, factor score, and FQL. Pearson–Lawley is not reported, since it turned out to be very close to FQL. This is because we have a population study where the model holds true, except for the minor deviations in the factor structure of the test. In line with the discussion in Section 4.4, Pearson–Lawley can be seen as merely a less efficient estimator than FQL, not fully utilizing the model structure.

For simplicity, measurement parameter estimates are not reported, but only

structural coefficients. Unstandardized and standardized values will be reported only for 25 per cent selection and residual correlation of .25, since it was found that the relative performance of the estimators was the same for other values. For the coefficients of the structural equation of FYA, percentage bias is given in parentheses. Table 3 refers to an R^2 of 50 per cent in s. We note that there is a remarkable degree of similarity of FQL and LQL bias. This calls into question whether or not the extra computational effort of FQL is worthwhile. Note, however, that FQL does better than LQL in terms of estimated R^2 for FYA. The FS estimator is for some coefficients better than both FQL and LQL.

Table 4 gives the corresponding results for an R^2 in s of 75 per cent. This higher R^2 value was chosen for the following reason. The full potential of FQL may be achieved only when the observed background variables explain selection well. In practice this would mean that more variables related to selection should be included in the model. In our study, we can simulate that situation by increasing R^2. As is seen in a comparison of Table 3 and Table 4 results, this change in s R^2 has a dramatic impact. The advantage of FQL over LQL is now clear. For the estimated R^2 in FYA, there is a considerable difference. The correct R^2 is 50 per cent. FQL obtains a value of 41 per cent whereas LQL obtains a value of 24 per cent. The standardized solution now shows the factor score estimator to be more biased overall than FQL.

6. Discussion

The proposed FQL estimator works best when selecting on $\mathbf{x}$, where $\mathbf{x}$ determines selection to a high degree. It will be useful in studies with traditional selection, say in the form of a simple sum of test variables. In such cases, a latent variable model is presumably not contemplated. Here, latent variable modelling may come into play in secondary analyses with the aim of investigating which factors are most important in the prediction. Given such successful latent variable modelling suggesting differential impact of different factors, the researcher might in a future study attempt to select individuals based on η rather than on $\mathbf{x}$. This may be done by estimating factor scores for the test takers, although this will result in estimation errors and will not be the same as selecting on η. Validation should then not be carried out by regression of y on the estimated scores, since according to the model these are not the assumed predictors but the variables of η are. Since factor scores are computed as a function of the observed $\mathbf{x}$ vector, such selection is determined by $\mathbf{x}$ and the assumption behind FQL is still fulfilled. Hence, FQL would be the appropriate validation technique also in this case. Although selection then takes place on variables closely related to the predictor variables of η, there may well be sufficient matriculant variation in these predictors for stable estimation of the latent variable model.

The usefulness of the FQL estimator for latent variable models has wider implications than for predictive validity problems. For example, longitudinal studies frequently result in data where not all subjects have observations on all variables at later time points. This may be a result of the design, e.g. using adaptive testing where test forms with different difficulty level are administered depending on performance at

previous time points, or it may be due to self-selection and attrition. The FQL estimator has general applicability to latent variable modelling with missing data.

References

Gustafsson, J. E. (1988*a*). Broad and narrow abilities in research on learning and instruction. In R. Kanfer, P. L. Ackerman & R. Cudeck (Eds), *Abilities, Motivation, and Methodology: The Minnesota Symposium on Learning and Individual Differences*. Hillsdale, NJ: Erlbaum.

Gustafsson, J. E. (1988*b*). Hierarchical models of individual differences in cognitive abilities. In R. J. Sterberg (Ed.), *Advances in the Psychology of Human Intelligence*, vol. 4, pp. 35–71. Hillsdale, NJ: Erlbaum.

Johnson, N. & Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*. New York: Wiley.

Lawley, D. N. & Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method*. London: Butterworth.

Little, R. J. & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Toronto: Wiley.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika*, **29,** 177–185.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, **49,** 115–132.

Muthén, B. (1987). LISCOMP. *Analysis of Linear Structural Equations with a Comprehensive Measurement Model*. Theoretical integration and user's guide. Mooresville, IN: Scientific Software.

Muthén, B. (1989). Factor structure in groups selected on observed scores. *British Journal of Mathematical and Statistical Psychology*, **42,** 81–90.

Muthén, B., Kaplan, D. & Hollis, M. (1987). On structural equation modelling with data that are not missing completely at random. *Psychometrika*, **42,** 431–462.

Muthén, B., Hollis, M., Muthén, L. & Tam, T. (1991). *Applying Logistic Regression to Choice-based and Supplementary Samples for the Prediction of MBA Matriculation and Persistence*. GMAC Occasional Papers. Los Angeles: Graduate Management Admission Council.

Muthén, B. & Jöreskog, K. (1983). Selectivity problems in quasi-experimental studies. *Evaluation Review*, **7,** 139–173.

Muthén, B., Shavelson, R., Hollis, M., Kao, C., Muthén, L., Tam, T., Wu, S. & Yang, J. (1988). *Relationship between Applicant Characteristics, MBA Program Attributes, and Student Performance. The Psychometric Study*, GMAC Occasional Papers. Los Angeles: Graduate Management Admission Council.

Tucker, L. R. (1971). Relations of factor score estimates to their use. *Psychometrika*, **36,** 427–436.

Received 13 September 1990; revised version received 19 August 1992; final version received 18 January 1993