Proceedings of Statistics Canada Symposium 2002
Modelling Survey Data for Social and Economic Research

EMBEDDING IRT IN STRUCTURAL EQUATION MODELS: A COMPARISON WITH REGRESSION BASED ON IRT SCORES

D.R. Thomas[1], I.R.R. Lu[1], and B.D. Zumbo[2]

ABSTRACT

This paper reviews the problems associated with using IRT-based latent variable scores for analytical modeling, discusses the connection between IRT and SEM-based latent regression modelling for discrete data, and compares regression parameter estimates obtained using predicted IRT scores and standardized Number-Right scores in OLS regression with regression estimates obtained using the combined IRT-SEM approach. The Monte Carlo results show the EAP approach is insensitive to sample size as expected but leads to appreciable attenuation in regression parameter estimates. On the other hand, the IRT-SEM method produced smaller finite sample bias, and as expected, generated 'consistent' regression estimates for suitably large sample sizes.

KEY WORDS: IRT; SEM; Latent Variable Modelling; Number-Right Scores; Personal Statistics; Scores.

1. INTRODUCTION

Researchers in the behavioural sciences frequently study constructs that are not directly observable, for example, job satisfaction, work stress, clinical depression levels, children's mathematical ability, children's reading ability, managerial competence, etc. To measure such constructs, referred to herein as latent variables, data consisting of responses to a set of scale items are obtained from each subject. If required, individual latent variable scores can then be obtained by means of a variety of techniques, which include factor analysis, classical psychometric scaling and item response theory (IRT). However, research interest usually focuses on relationships among latent variables, not on measurements of the latent variables themselves, and a variety of techniques have been developed to facilitate this requirement for latent variable modelling, most notably the structural equation model (SEM) methodology now available in programs such as LISREL (Joreskog and Sorbom, 1996), EQS (Bentler, 1995), AMOS (Arbuckle, Wothke, 1999), Mplus (Muthen and Muthen, 2001), LINCS (Schoenberg and Arminger, 1990). By combining factor analytic measurement models with structural equations of regression type, SEM methods can consistently estimate the parameters of a very general set of models. There is now a vast literature on SEM which intersects with the literature on measurement error models (Fuller, 1987) and in many cases extends it. However, though the use of SEM is growing, particularly for continuous measurement data, many researchers still prefer to use a simpler approach, in which latent variable scores are obtained and incorporated directly into regression and multivariate analyses. Though very common, this approach has serious drawbacks. It is well documented that the direct use of latent variable scores in regression and other statistical analyses can result in inconsistent or biased estimates of model parameters, and if the latent variable scales are based on too few items, this bias can be severe. This bias is likely to occur whenever a prediction is made of the value of a latent variable in a random effects model. (Note that in this paper, the term *prediction* is used to distinguish between the generation of scores on random latent variables and the *estimation* of fixed model parameters.) In general, the distribution of the predicted values does not converge to the distribution of the latent variable as the sample size grows, unless the number of scale items used is very large. Analyses based on the predicted values of latent variables are then subject to bias, regardless of sample size (Little and Rubin, 1983; Louis, 1984). Convergence occurs only when the number of scale

[1]Sprott School of Business, Carleton University, Ottawa, Ontario, Canada K1S 5B6.
(E-mail: rthomas@sprott.carleton.ca)

[2]University of British Columbia, Scarfe Building, 2125 Main Mall, Department of ECPS., Vancouver, B.C. Canada V6T 1Z4

items goes to infinity. This will be referred to as "finite item" bias in the paper. The problems associated with using latent variable scores derived from the factor analysis of continuous scale data in multiple regression have been extensively documented. The main conclusion of the research on this topic is that regression on factor scores can lead to serious finite item bias (Tucker, 1971; Shevlin, Miles, and Bunting, 1997), though there are situations where this can be avoided (Skrondal and Laake, 2001) or reduced (Croon, 2002). In contrast to the continuous scale case, relatively little attention has been focused on the problems of using binary scale data in regression analyses.

This paper, therefore, will focus on issues relating to the use of IRT-based latent variable scores in analytical modelling. In addition to IRT scores, the commonly used standardized "number right" (NR) scores will also be examined. These consist of the total number of test items correctly answered by an individual, standardized to have a sample mean of zero and a sample standard deviation of one. In Section 2, basic IRT methodology will be summarized and methods for predicting latent variables from IRT models will be described. In Section 3, the existing literature on the use of binary IRT scores in OLS regression and other analyses will be reviewed. Methods of correcting for and avoiding finite item bias will be briefly discussed. Though the connection between factor analysis and SEM methodology is widely known, the analogous connection between IRT and SEM appears not to be as widely understood. In Section 4, this connection will therefore be reviewed and a method for embedding IRT measurement models into a SEM-based latent regression analysis will be described. This approach yields consistent estimates of all parameters, that is, it avoids the problem of finite item bias. Nevertheless, it will still be necessary to determine how well the theoretically consistent IRT-SEM estimates perform in practice, i.e., the extent of the finite sample bias. Finally, results will be presented from a preliminary simulation study that compares regression parameter estimates obtained using predicted IRT and standardized NR scores in OLS regression with regression estimates obtained using the combined IRT-SEM approach.

2. PREDICTING LATENT VARIABLE SCORES USING DISCRETE SCALE ITEMS

This section will focus on techniques specifically developed for constructing measurement scales based on discrete items, in particular the general family of item response theory (IRT) methods, which include methods designed for binary data as well as ordinal multi-category items. In the interests of space, only binary models will be considered. The use of binary IRT scores in regression analyses will be discussed in Section 3 of the paper.

2.1 The Binary IRT Model

Item response theory (IRT) methods were developed in the fields of education and psychometric testing, where the latent variable plays the role of a subject *ability* or *trait*. In psychometric testing, a subject is administered a number of test items, binary in the current discussion, and the item responses are subsequently transformed into an ability or trait score for that individual, using a previously calibrated IRT model. In the survey research context, the test items are administered to all survey respondents and predictions of individual abilities (if produced) and IRT model parameters are estimated from the same survey dataset. In both cases, the key to IRT methods is a model that links the characteristics of a given test item, and the true value of an individual's ability, to the probability that the subject will respond correctly to that test item. In the normal ogive binary IRT model (Bock and Liebermann, 1970; Bock and Aitkin, 1981) the probability of a correct response to the j-th item, given a subject ability η, is assumed to be

$$p_j(\eta) = \int_{-\infty}^{a_j\eta - b_j} \varphi(z)\,dz = \Phi(a_j\eta - b_j), \qquad (1)$$

where η represents a vector of latent subject abilities, a_j and b_j are parameters specific to each test item, and φ and Φ represent, respectively, the normal density and the normal cumulative distribution function (*cdf*). Under the assumption of local independence, the conditional probability of observing an outcome vector of n test responses $\boldsymbol{x}$ is given by

$$P(\boldsymbol{x}|\eta) = \prod_{j=1}^{n} [p_j(\eta)]^{x_j}[1 - p_j(\eta)]^{1-x_j}, \qquad (2)$$

so that the marginal likelihood for an observation $\boldsymbol{x}$ becomes

$$P(\boldsymbol{x}) = \int P(\boldsymbol{x}|\eta)\,g(\eta)\,d\eta, \qquad (3)$$

where $g(\eta)$ is an assumed form of the population marginal distribution of η. In the most commonly encountered form of the IRT model, it is assumed that the latent ability, η, is uni-dimensional, and that the normal *cdf* can be approximated by (or replaced by) the logistic *cdf*. This is referred to as the two-parameter binary logistic model. Other variants, including a three-parameter version of the logistic IRT model, are commonly used.

Several techniques have been developed for estimating the item parameters, $\boldsymbol{a}_j$ and b_j, the best known and most frequently used being the maximum marginal likelihood (MML) approach developed by Bock and Liebermann (1970), and later refined by Bock and Aitkin (1981). In the Bock and Aitkin version, marginalization over η is facilitated by treating $g(\eta)$ as discrete, and the resulting likelihood equations are solved using an EM scheme. More recent approaches include the use of collateral information in the estimation (Mislevy, 1987) as well as MCMC methods for simultaneously estimating item parameters and predicting latent abilities (Patz and Junker, 1999).

2.2 Prediction of IRT Scores

A variety of methods have been proposed for predicting the value of the latent ability, η, for a specific subject, given the subject's test outcomes $\boldsymbol{x}$. These methods include maximum likelihood "estimates" (MLE's), obtained by treating η as a fixed parameter and maximizing the "likelihood" (2) in which the locally independent items play the role of independent observations. The weighted likelihood estimator (WLE; Warm, 1989) is closely related to the MLE but has better bias properties as $n \to \infty$, where n is the number of test items. Its bias conditional on η is $o(n^{-1})$, compared to $O(n^{-1})$ for the MLE (Lord, 1983). Latent variable predictors can also be obtained from the posterior distribution of η, given by

$$P(\eta|\boldsymbol{U}_i) = \frac{P(\boldsymbol{x}_i|\eta)g(\eta)}{\int P(\boldsymbol{x}_i|\eta)g(\eta)\,d\eta}. \quad (4)$$

Two predictors of this type are the *maximum a posteriori* (MAP) predictor, obtained by maximizing equation (4) for each subject, and *expected a posteriori* (EAP) predictor, which is the mean of the posterior distribution (4). The latter is usually evaluated using Gauss-Hermite integration of the posterior latent distribution (see Stroud and Sechrest, 1966). It should be noted that the usual procedure is to fix the item parameters contained in the likelihood and the parameters of the latent distribution (if any) at their estimated values. In other words, predictors based on (4) will be of empirical Bayes type.

Kim and Nicewander (1993) carried out a detailed investigation of the bias, standard errors and reliabilities of the above latent variable predictors using Monte Carlo techniques. They also examined the standardized NR score. They concluded that while the reliabilities of all five scores were very similar, all five exhibited large conditional bias for $|\eta| > 1$, with η measured on an $N(0,1)$ scale, the worst being the NR score and the MLE. Conditional bias, reliability and standard errors were similar for WLE and the Bayesian predictors MAP and EAP, with the WLE exhibiting slightly lower conditional bias but higher standard error than the other two.

3. FINITE ITEM BIAS IN LATENT VARIABLE REGRESSION

The goal of this section is to investigate the finite item bias in parameter estimates caused by using latent variable IRT scores in standard statistical analyses. Methods of correcting for finite item bias and methods for removing finite item bias by avoiding the prediction of individual latent variable scores will also be briefly described. Regression analysis will receive the greatest attention in this section, though some of the methods that have been proposed for overcoming bias are applicable to other analyses as well.

3.1 Direct Use of IRT Scores

Examples of the direct use of IRT scores in regression are less common in the literature than are examples of the use of factor scores derived from continuous item scales, which is not surprising given that IRT is not as well known as factor analysis outside the fields of psychometrics and educational testing. Though there seems to be general agreement

in the technical literature that using IRT scores directly is unadvisable (see, for example, Mislevy, Johnson and Muraki, 1992; Houtink and Boomsma, 1996), and leads to bias in parameter and standard error estimates, specific empirical evaluations of these effects are scarce. One study that does provide empirical evidence is that of Adams, Wilson and Wu (1997), who studied an extension of the Rasch model that encompasses a number of model types, including the partial credit model (Masters, 1982) that is designed to handle ordered categorical indicator variables. Their estimation scheme incorporated collateral information, i.e., they represented the latent variable, η, as a linear function of collateral information, $\boldsymbol{Y}$, such as gender, socioeconomic status, etc. Their model for η consisted of a simplified factor regression model of the form

$$\eta = \boldsymbol{Y}\beta + \zeta, \qquad (5)$$

where the vector of disturbances, ζ, was assumed to be normal with mean zero and variance σ^2. The Adams et al. (1997) approach is closely related to that used earlier by Mislevy (1987), and differs from that of Bock and Aitkin (1981) in that more prior information is assumed for η. Among other things, Adams et al. (1997) provided a comparison between three estimators of β, namely: (1) the estimator obtained using a simultaneous *ML*/*EM* estimation of the model item parameters and the regression coefficients; (2) a three-step estimator in which the item parameters were first estimated without using the collateral information, then used to generate EAP estimates of the latent trait, η, with the regression parameters, β, finally determined by OLS regression of the EAP scores on the collateral information $\boldsymbol{Y}$; (3) a two-step approach in which regression parameters were determined from a set of plausible values of η (see Section 3.3 below). Their simulation results demonstrated that OLS estimates based on EAP scores underestimated both R^2 and the magnitude of the regression coefficients, while the parameter estimates obtained using simultaneous estimation (method 1) and the plausible values approach (method 3) were close to the true model values.

3.2 Using Bias-Corrected IRT Scores

For the case of latent variable scores based on continuously measured indicator variables, Croon (2002) argued that there are some situations in which a two-step approach to latent variable regression and other analyses might be advantageous. For example, the search for appropriate measurement models, and the search for the correct functional form of a latent variable regression model, may be easier to undertake if the measurement models are estimated separately from the latent variable model. Similar arguments can be advanced for the case of discrete indicator variables, when IRT scores must be used as proxies for latent variables. If a two-step approach to latent variable regression is to be used in the discrete case, methods for incorporating bias corrected IRT scores become important.

The results of Kim and Nicewander (1993) reported earlier showed that the least biased IRT predictors were the posterior mode (MAP), the posterior mean (EAP) and the weighted likelihood predictor (WLE). Hoijtink and Boomsma (1996) investigated the bias in MAP and WLE predictors both empirically and theoretically, and developed asymptotic expressions, correct to order $O(1/n)$, for the conditional mean and variance of the latent variable, η, the covariance between η and a measured covariate, $\boldsymbol{y}$. For the predictor WLE, these expressions can be readily evaluated and used in many kinds of analyses, including analysis of variance and multiple regression. Based on an ANOVA simulation, Hoijtink and Boomsma (1996) concluded that their asymptotic bias correction method was necessary, but that it still required at least 15 binary items in the IRT model to keep biases in the ANOVA means and error estimates to an acceptable level.

3.3 Plausible Values

The *plausible values* technique differs from the techniques described above in that multiple sets of scores are provided for each latent variable, which should not be interpreted as individual predictors. Plausible values in fact represent multiple imputations (Rubin, 1987) drawn from the predictive distribution of the latent variable, the latent variable in this case being treated as missing data. The technique was designed for surveys such as the bi-annual U.S. National Assessment of Education Progress (NAEP) survey, in which individual students may be asked to answer as few as eight questions from a particular test, so that none of the likelihood based or posterior based predictions will provide sufficient accuracy. The predictive distribution incorporates the variability arising from using a finite number of scale items. Therefore, plausible values estimators, which consist of estimated expected values taken with respect to the predictive

distribution, will be free of finite item bias. A detailed discussion of the plausible values methodology used in the NAEP is given by Mislevy, Johnson and Muraki (1992).

3.4 Simultaneous Estimation of Item and Regression Parameters

The work of Adams et al. (1997) for the ordinal partial credit data, and the earlier work of Mislevy (1987) for binary indicator variables, are methods whereby the IRT model item parameters and the regression parameters are simultaneously (and consistently) estimated, bypassing the need to predict latent variable scores, and thus avoiding finite item bias. Similar simultaneous modelling of the regression parameters was also undertaken by Zwinderman (1991) in the context of a Rasch model. These methods were developed to improve precision of estimation by incorporating collateral information via equation (5), which features a latent response variable, but predictor variables measured without error. Thus they do not provide a general approach to analyzing latent variable regression models. Nevertheless, general methods do exist for obtaining consistent parameter estimates of systems of linear latent variable equations with discrete indicator variables, and these can be used to simultaneously estimate IRT item parameters and the parameters of the latent regression models, as described below.

4. SIMULTANEOUS ESTIMATION OF IRT AND LATENT REGRESSION MODEL PARAMETERS

From the mid 1970's on, extensive work was carried out by numerous authors (for example, Christopherson, 1975; Muthen, 1983, 1984) to extend the techniques of SEM to handle discrete indicator variables. The methodology currently implemented in programs such as Mplus (Muthen amd Muthen, 2001) allows for the analysis of mixtures of discrete and continuous variables and includes robust procedures for estimating standard errors and fit statistics that do not rely on multivariate normal indicators. This methodology is very general and subsumes IRT modeling and the simultaneous estimation of both IRT item parameters and the parameters of latent regression models. The IRT connection to general SEM appears not to be as well known as the corresponding connection between factor measurement models and SEM in the continuous case, despite the fact that the IRT connection was very clearly spelled out by Takane and de Leeuw (1987). It was used by Muthen (1988) to relate IRT models to external variables such as grouping indicators as well as to continuous background variables, and was subsequently extended by Muthen, Kao and Burstein (1991).

4.1 SEM-Based Latent Regression Models with Binary Indicators

The structural SEM model is given by

$$\eta = \Gamma\xi + \zeta \qquad (6)$$

where η is a vector of latent response variables, ξ is a vector of latent explanatory variables, Γ is a matrix of regression coefficients and ζ represents a vector of disturbances independent of ξ, with covariance matrix Ψ. The complete specification of system (6), which is a special case of the general model available via Mplus (Muthen and Muthen, 2001), requires the specification of measurement models for η and ξ, namely

$$\boldsymbol{x}^* = \Lambda_x\xi + \delta \quad \text{and} \quad \boldsymbol{y}^* = \Lambda_y\eta + \varepsilon \quad , \qquad (7)$$

where $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$ are themselves unobservable variables consisting of vectors of items (or indicator variables) of dimension p and q, respectively , and the "loadings" Λ_x and Λ_y comprise matrices of fixed parameters. The vectors δ and ε contain random disturbances that are independent of the latent variables and of each other. For the single population model discussed here, the means of $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$ can be set to zero. Since $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$ cannot be observed, their variance is arbitrary and usually set to one for convenience. The observable discrete indicator variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are then modeled as

$$x_j = 1 \text{ if } x_j^* \geq \tau_j \quad , \quad x_j = 0 \text{ otherwise, } \quad j = 1, \ldots, p, \qquad (8)$$

$$y_j = 1 \quad \text{if} \quad y_j^* \geq \tau_j \quad , \quad y_j = 0 \ \text{otherwise}, \quad j = p+1, \ldots, p+q,$$

where the τ's represent thresholds that must be estimated from the data, and p and q are the number of indicator variables, or items, in the measurement models for η and ξ, respectively.

Maximum likelihood estimation of the parameters of the above model is difficult as it involves integration of $p+q$ correlated multivariate normal variables over the $p+q$ dimensional space defined by the thresholds. For this reason, a limited information generalized least squares approach to estimation is used, based on one-way and two-way marginal probabilities of the discrete indicator variables (Christopherson, 1975; Muthen, 1984). Consistency of the parameter estimates has been established.

4.2 The Connection between IRT and SEM

Takane and de Leeuw (1987) showed that a discrete measurement model of the form given in equation (8) is formally equivalent to the normal ogive IRT model given in equation (1). In particular, for a discrete measurement model of the form $y^* = \Lambda\eta + \varepsilon$, with $var(\varepsilon) = \Theta$, they showed that the item parameters $\boldsymbol{a}_j$ and b_j of equation (1) can be expressed in terms of the parameters of the discrete measurement model as

$$\boldsymbol{a}_j = \lambda_j/\theta_j^{1/2} \qquad \text{and} \qquad b_j = -\tau_j/\theta_j^{1/2}, \tag{9}$$

where λ_j is the j-th row of Λ, and θ_j is the j-th diagonal element of Θ. For uni-dimensional latent IRT traits, $\boldsymbol{a}_j$ will be scalar. Thus estimation of the parameters of the discrete SEM defined by equations (6), (7) and (8) can be regarded as a simultaneous and consistent estimation of IRT item parameters and the parameters of a latent regression model. In other words, the IRT model is automatically embedded in the structural equation model. Whenever the IRT parameters $\boldsymbol{a}_j$ and $\boldsymbol{b}_j$ are known, the corresponding measurement model parameters, λ_j and τ_j, can be calculated and held fixed during the SEM estimation. This will be referred to as the fixed IRT-SEM approach to distinguish it from the case where both IRT and structural parameters are simultaneously estimated. The fixed IRT-SEM approach will again provide consistent estimates of the structural equation parameters.

5. MONTE CARLO STUDY

The Monte Carlo study has two parts. In the first part, the simulation capabilities of Mplus (Muthen and Muthen, 2001) were used to investigate the finite sample bias of the estimate of the latent regression parameter, Γ, obtained using both the simultaneous and fixed IRT-SEM approaches described in Section 4.2. In the second part, EAP and NR scores for both response and explanatory latent variables were predicted from simulated binary observations, and the latent regression parameter, Γ, was then estimated using OLS regression. The focus of this part of the study was to investigate the finite item bias attributable to using predicted scores, as a function of the number of test items.

5.1 Design of Monte Carlo Study

A simple structural regression model was selected for study, consisting of one latent response variable, η, and one latent explanatory variable, ξ, i.e., equation (6) with Γ equal to a scalar parameter γ, set at the value $\sqrt{2}/2$. The variances of η and ξ were set to one, yielding a coefficient of determination of 0.5 for the structural regression model. The measurement models for the latent response and explanatory variables were $y_j^* = \lambda\eta_j + \varepsilon_j$ and $x_j^* = \lambda x_j + \delta_j$, $j = 1, \ldots, n$, with n being the number of items in each scale. They have identical loading parameters. The variances of the individual y_j^* and x_j^* were set at one.

The process of generating binary item response data for the simulation involved three steps. In the first step, the means and tetrachoric correlations of y_j^* and x_j^* were calculated to correspond to the structural and measurement models. In

the second step, multivariate normal item responses were generated using these means and tetrachoric correlations. In the third step, the multivariate item response data was dichotomized into binary item response data, using item thresholds as described in Section 4.1. The thresholds were initially drawn from a standardized normal distribution, truncated to provide values in the range -1.5 and 1.5, and then fixed for the duration of the study. For the first part of the study, the last two steps were carried out using the simulation capabilities of Mplus.

Several characteristics were varied in this stage of the study: (a) the coefficients of determination of the measurement models, namely *cd* = 0.3 and 0.6; (b) the number of simulated items in the test, namely *n* = 10, 20, and 30; and (c) the sample sizes, namely *N*=300, 500, 800, and 2000. For each experimental condition, final results were based on the means of 1000 independent replications.

The second part of the study again involved generating binary indicators for both response and explanatory latent variables. EAP predictions were then calculated for each of the *N* cases in a given experimental condition, based on the normal ogive IRT model corresponding to the measurement models used in the first stage (see Section 4.2). Thus the IRT item parameters were treated as known and fixed in this study. However, the work of Kim and Nicewander (1993) suggests that results obtained using fixed parameters will not differ greatly from results based on estimated parameters, for the sample sizes considered here. The EAP scores were calculated using 24 point Gauss-Hermite integration of the posterior latent distribution. OLS regression using the EAP scores was then used to estimate the latent regression parameter, γ, with summary results based on 1000 independent simulations. This stage of the experiment was also repeated using standardized NR scores.

The following tables provide a comparison of the theoretically consistent estimates obtained using the IRT-SEM methodology with the results obtained using latent score regression.

5.2 Results of Monte Carlo Study

Table 1 shows that with 10 items in each measurement model, the EAP approach is insensitive to sample size as expected but leads to appreciable attenuation in regression parameter estimates and coefficients of determination. On the other hand, the simultaneous IRT-SEM approach produced smaller finite sample bias even for the smaller sample sizes, and as expected it generated 'consistent' regression estimates for suitably large sample sizes.

Table 2 shows the effect of test length on finite item bias in the regression estimates, for a sample of size 800. As predicted, the parameter bias with EAP regression decreases as number of items increases. Table 2 also shows that the higher the coefficient of determination in the measurement model, the smaller the finite item bias in EAP regression. For simultaneous IRT-SEM estimation, finite item bias was small for all test lengths, reaching a maximum of about 2% for a 10 item test.

Table 1. Biases in Regression Parameter Estimates

# of items:	EAP Regression Estimation			Simultaneous IRT-SEM Estimation		
10-10	$\hat{\gamma}$	*%bias* of $\hat{\gamma}$	*%bias* of R^2	$\hat{\gamma}$	*%bias* of $\hat{\gamma}$	*%bias* of R^2
CD=0.3						
N=300	0.466 (.002)	-34.1	-55.7	0.764 (.01_)	+8.06	+16.1
N=500	0.466 (.001)	-34.1	-56.0	0.733 (.007)	+3.68	+7.36
N=800	0.467 (.001)	-33.9	-56.0	0.722 (.005)	+2.12	+4.24
N=2000	0.468 (.001)	-33.8	-56.0	0.711 (.003)	+0.57	+1.14
CD=0.6						
N=300	0.590 (.001)	-16.6	-30.0	0.751 (.006)	+6.22	+12.4
N=500	0.590 (.001)	-16.6	-30.2	0.734 (.004)	+3.82	+7.64
N=800	0.590 (.001)	-16.6	-30.3	0.720 (.003)	+1.84	+3.68
N=2000	0.591 (.001)	-16.4	-30.3	0.709 (.002)	+0.28	+0.56

Note: N denotes sample size. CD denotes coefficient of determination of both measurement models. Standard errors of γ appear in parentheses.

Table 2. Effect of Test Length on Regression Parameter Estimates

N=800	EAP Regression Estimation			Simultaneous IRT-SEM Estimation		
# of items	$\hat{\gamma}$	*%bias* of $\hat{\gamma}$	*%bias* of R^2	$\hat{\gamma}$	*%bias* of $\hat{\gamma}$	*%bias* of R^2
CD=0.3						
10-10	0.467 (.001)	-33.9	-56.0	0.722 (.005)	+2.12	+4.24
20-20	0.562 (.001)	-20.5	-36.6	0.711 (.003)	+0.57	+1.13
30-30	0.604 (.001)	-14.6	-27.0	0.714 (.003)	+0.99	+1.98
CD=0.6						
10-10	0.590 (.001)	-16.6	-30.3	0.720 (.003)	+1.84	+3.68
20-20	0.640 (.001)	-9.48	-18.0	0.715 (.003)	+1.13	+2.26
30-30	0.660 (.001)	-6.65	-13.0	0.716 (.003)	+.003	+.005

Table 3 displays a comparison of simultaneous and fixed IRT-SEM estimation. It appears that fixed IRT-SEM estimation yields smaller finite sample bias than the simultaneous version for the smaller sample sizes. Even for 10 items and a sample size of only 300, finite sample bias is less than 5% if the item parameters are fixed at their known values.

Table 3. Comparison of Finite Sample Biases with Fixed and Simultaneous IRT-SEM Estimation

# of items: 20-20	Fixed IRT-SEM Estimation			Simultaneous IRT-SEM Estimation		
	$\hat{\gamma}$	*%bias* of $\hat{\gamma}$	*%bias* of R^2	$\hat{\gamma}$	*%bias* of $\hat{\gamma}$	*%bias* of R^2
CD=0.3						
N=300	0.739 (.004)	+4.53	+9.06	0.764 (.01_)	+8.06	+16.1
N=500	0.724 (.003)	+2.40	+4.80	0.733 (.007)	+3.68	+7.36
N=800	0.717 (.003)	+1.41	+2.82	0.722 (.005)	+2.12	+4.24
N=2000	0.710 (.002)	+0.42	+0.84	0.711 (.003)	+0.57	+1.14
CD=0.6						
N=300	0.738 (.002)	+4.38	+8.76	0.751 (.006)	+6.22	+12.4
N=500	0.731 (.002)	+3.39	+6.78	0.734 (.004)	+3.82	+7.64
N=800	0.723 (.002)	+2.26	+4.52	0.720 (.005)	+1.84	+3.68
N=2000	0.712 (.001)	+0.71	+1.42	0.709 (.002)	+0.28	+0.56

Table 4 shows that finite item bias is very similar for standardized NR and EAP regression estimates. A side-study showed that the IRT scores and the NR scores were almost perfectly correlated.

Table 4. Comparison of Finite Item Biases with EAP and Standardized NR Scores

CD	N	# of items: 10-10		# of items: 20-20	
		EAP	Standardized NR	EAP	Standardized NR
0.3	300	0.466 (.002)	0.468 (.001)	0.562 (.002)	0.562 (.001)
	500	0.466 (.001)	0.467 (.001)	0.562 (.001)	0.561 (.001)
	800	0.467 (.001)	0.467 (.001)	0.562 (.001)	0.562 (.001)
0.6	300	0.590 (.001)	0.589 (.001)	0.641 (.001)	0.638 (.001)
	500	0.590 (.001)	0.589 (.001)	0.642 (.001)	0.639 (.001)
	800	0.590 (.001)	0.588 (.001)	0.640 (.001)	0.638 (.001)

6. CONCLUSION

In general, the distribution of the predicted values does not converge to the distribution of the latent variable as the sample size increases unless the number of the test items is sufficiently large. Analyses based on the predicted values of latent variables are subject to finite item bias, and thus predicted latent variable scores should be used with caution. Our preliminary Monte Carlo study showed that for binary items, both simultaneous and fixed IRT-SEM methodology generated 'consistent' regression parameter estimates for suitably large sample sizes. On the other hand, when IRT and standardized NR scores are used in regression, large numbers of test items (n>30) are required to reduce finite item bias in regression coefficients to acceptable levels. Thus, given large enough samples, IRT-SEM modelling may provide an alternative to direct estimation and analysis using IRT scores, and also an alternative to the use of plausible value methods.

Like plausible value methods, the IRT-SEM modelling approach side-steps the prediction of individual latent variable scores, but unlike the plausible value methods, it does not require a large set of conditioning variables and a major data processing effort on the part of the agency providing the data set. Thus IRT-SEM methods may provide an alternative way for agencies like Statistics Canada to make consistent latent variable regression analysis feasible for users of their data. Sample size requirements of the simultaneous IRT-SEM methods may be limiting in some contexts. However, using previously calibrated items will offset this effect to some extent.

Another finding of this study was that the regression estimates obtained using standardized NR and EAP scores are highly comparable, regardless of test length and measurement model precision. This requires further research.

REFERENCES

Adams, R. J., M. Wilson, and M. Wu. (1997), "Multilevel Item Response Models: An Approach to Errors in Variables Regression", *Journal of Educational and Behavioral Statistics*, 22(1), pp. 47-76.

Arbuckle, J. L. and W. Wothke (1999), *AMOS 4.0 User's Guide*, SPSS.

Bentler, P. M. (1995), *EQS Structural Equations Program Manual*, Encino, Calif.:Multivariate Software.

Bock, R. D. and M. Lieberman (1970), "Fitting a Response Model for N Dichotomously Scored Items"*, Psychometrika,* 35(2), pp. 179-197.

Bock, R. D. and M. Aitkin (1981), "Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM algorithm", *Psychometrika*, 46(4), pp. 443-459.

Christoffersson, A. (1975), "Factor Analysis of Dichotomized Variables", *Psychometrika*, 40(1), pp. 5-32.

Croon, M. (2002), "Using Predicted Latent Scores in General Latent Structure Models", in G. A. Marcoulides and I. Moustaki (eds.) *Latent Variable and Latent Structure Models*, Mahwah, NJ: Lawrence Erlbaum Associates, pp. 195-223.

Fuller, W. A. (1987), *Measurement Error Models*, New York: Wiley.

Hoijtink, H. and A. Boomsma (1996), "Statistical Inference Based on Latent Ability Estimates", *Psychometrika*, 61(2), pp. 313-330.

Jöreskog, K. G. and D. Sörbom (1996), *LISREL 8 User's Reference Guide*, Chicago: Scientific Software International.

Kim, J. K. and W. A. Nicewander (1993), "Ability Estimation for Conventional Tests." *Psychometrika*, 58(4), pp. 587-599.

Little, R. J. A. and D. B. Rubin (1983), "On Jointly Estimating Parameters and Missing Data by Maximizing the Complete Data Likelihood", *American Statistician*, 37, pp. 218-220.

Lord, F. M. (1983), "Unbiased Estimators of Ability Parameters, of their Variance, and of their Parallel-forms Reliability", *Psychometrika*, 48(2), pp. 233-245.

Louis, T. A. (1984), "Estimating a Population of Parameter Values Using Bayes and Empirical Bayes Methods", *Journal of the American Statistical Association*, 79(386), pp. 393-398.

Masters, G. N. (1982), "A Rasch Model for Partial Credit Scoring", *Psychometrika*, 47(2), pp. 149-174.

Mislevy, R. J. (1987), "Exploiting Auxiliary Information about Examinees in the Estimation of Item Parameters", *Applied Psychological Measurement,* 11(1), pp.81-91.

Mislevy, R. J., E. G. Johnson, and E. Muraki (1992), "Scaling Procedures in Naep." *Journal of Educational Statistics* 17(2), pp. 131-154.

Muthen, B. (1983), "Latent Variable Structural Equation Modeling with Categorical Data", *Journal of Econometrics*, 22, pp. 43-65.

Muthen, B. (1984), "A General Structural Equation Model with Dichotomous, Ordered Categorical, and Continuous Latent Variable Indicators", *Psychometrika*, 49(1), pp. 115-132.

Muthen, B. (1988), "Some Uses of Structural Equation Modeling in Validity Studies: Extending IRT to External Variables", in H. Wainer and H. I. Braun (eds.) *Test Validity,* Princeton, NJ: Educational Testing Service, pp. 213-238.

Muthen, B., C. F. Kao, and L. Burstein (1991), "Instructionally Sensitive Psychometrics: Application of a New IRT-Based Detection Technique to Mathematics Achievement Test Items", *Journal of Educational Measurement*, 28(1), pp. 1-22.

Muthen, L. K. and B. O. Muthen (2001), *Mplus User's Guide,* Los Angeles, CA: Muthen & Muthen.

Patz, R. J. and B. W. Junker (1999), "A Straightforward Approach to Markov Chain Monte Carlo Methods for Item Response models", *Journal of Educational and Behavioral Statistics*, 24(2), pp. 146-178.

Rubin, D. B. (1987), *Multiple Imputation For Non-Response in Surveys*, New York: Wiley.

Schoenberg, R. and G. Arminger (1990). *LINCS2.0*. Kent. WA, RJS Software.

Shevlin, M., J. N. V. Miles, and B.P. Bunting (1997), "Summated Rating Scales: A Monte Carlo Investigation of the Effects of Reliability and Collinearity in Regression Models", *Personality and Individual Differences*, 23(4), pp. 665-676.

Skrondal, A. and P. Laake (2001), "Regression among Factor Scores", *Psychometrika*, 66(4), pp. 563-576.

Stroud, A. H. and D. Secrest (1966), *Gaussian Quadrature Formulas*, Englewood Cliffs, N.J.: Prentice-Hall.

Takane, Y. and J. Deleeuw (1987), "On the Relationship between Item Response Theory and Factor Analysis of Discretized Variables", *Psychometrika,* 52(3), pp. 393-408.

Tucker, L. (1971), "Relations of Factor Score Estimates to their Use", *Psychometrika*, 36(4), pp. 427-436.

Warm, T. A. (1989), "Weighted Likelihood Estimation of Ability in Item Response Theory", *Psychometrika,* 54(3), pp. 427-450.

Zwinderman, A. H. (1991), "A Generalized Rasch Model for Manifest Predictors", *Psychometrika*, 56(4), pp. 589-600.