Embedding IRT in Structural Equation Models as a Solution to Problems with Measurement Error in Predictor Variables in Validation Research

UBC-Paragon Webinar Series 2021-1

February 16, 2021

Bruno D. Zumbo

Professor & Distinguished University Scholar

Tier 1, Canada Research Chair in Psychometrics and Measurement

Paragon UBC Professor of Psychometrics and Measurement

University of British Columbia

Acknowledgements

- This presentation is an integration and extension of my programs of research on IRT and test scoring, and validation methods and variable ordering.

- Thank you for contributions from a collaborative research program with *Professor D. Roland Thomas* and *Dr. Irene Lu*, Carleton University, Sprott School of Business.

Outline

1. **Opening Remarks** – motivation
2. **Solutions**
   I. Observed Total Score
   II. Predicted Latent Variable Score
      a) Likelihood, b) Posterior Distribution Based, c) Plausible Values (Multiple Imputation)
   III. Embedding IRT in SEM
3. **Example**
   - Large-scale testing example, including variable ordering (Latent variable Pratt index)
4. **Conclusions**

Opening Remarks: Motivating the Problem

- Common research questions in assessment and validation studies use regression are:
  - which of the variables is most predictive of the criterion measure?
  - what variable(s) is most explanatory of our test scores?
- Let's consider an example to motivate our discussion of a validation study investigating predictors of a test of “knowledge of numbers”, a grade 8 mathematics test.

Opening Remarks: Motivating the Problem

- The criterion is:
  - Number Knowledge (9 items)
- Predictors are:
  1) availability of computers,
  2) time spent on math homework,
  3) tutoring or extra lessons,
  4) number of books at home -- all of which are observed variables, and
  5) a measure (test) of how much the student's value the "importance of mathematics" (4 items).

Opening Remarks

- There are three strategies that are typically discussed in the literature:

I. **Compute observed total scores (e.g., number correct scores on the math test and a scale score on the value of the importance of math).**
   - Fit a regression equation and test model parameters using the observed total scores in place of the discrete test items.

II. **Predicting Individual Abilities / Traits (IRT or factor scores) on the "math test" and separately on the "valuing math" scale.**
   a. Likelihood Based
   b. Posterior Distribution Based
   c. Plausible Values (Multiple Imputation) .. Not really a "prediction", like (a) or (b).
   - IRT external to the regression. Fit a regression equation and test model parameters using the predicted IRT or factor scores in place of the discrete test items.

III. **Embedding IRT in a Structural Equation Model.**

I. Observed total score
(number correct score)

- The common approach is to treat this test and/or scale scores as an observed score composite and use the composite score as variables in an analysis.
  - In ordinary least-squares regression, measurement error in the criterion variable can perhaps be absorbed into the error term.
- However, measurement in the predictors is not accommodated for and can be quite problematic.
  - Estimated regression coefficients are biased even as the sample size approaches infinity.
  - Type I error rates can be seriously inflated.
    - As a side note, far less widely known is that the Type I error is inflated; can approach 1.0 (correlation among predictors, and measurement error; see Shear & Zumbo, 2013, in *EPM*)
  - Cochran (1968), Fuller (1987), and Brunner & Austin (2009).

II. Predicting Individual Abilities / Traits

- **IRT is used externally to the regression model as a way of getting predicted theta ($\hat{\theta}$) scores to then input in the regression -- or if continuous item scores one could use linear factor analysis.**

- **2 types of methods of predicting ability (i.e., obtaining)**
  - **Likelihood Based**
  - **Posterior Distribution Based**

- **Will be assumed that item parameters and the parameters of the ability distribution are known exactly.**
- **In practice, item parameters will be replaced by their consistent estimates, which is current practice in IRT.**

Plausible Values (Multiple Imputation)

- Treat abilities as missing
- Use random draws from the predictive distribution to estimate regression parameters
- Used at Statistics Canada in PISA and TIMMS
- For multiple imputation in the IRT context, see Rubin (1987) and Mislevy, Johnson, & Muraki (1992)

A. Likelihood Based Methods

1. MLE (Maximum Likelihood Estimates):

- Treats a subject's ability as a fixed parameter, θ_i, and assumes that the subject's score (test outcomes *x*) is a random draw from a population of subjects having that particular ability

- The appropriate likelihood for ML estimation is,

$$P\left(X_i \mid \theta_i\right) = \prod_{j=1}^{n} P_j(\theta_i)^{x_{ij}} \left\{1 - P_j(\theta_i)\right\}^{(1-x_{ij})}$$

- Maximization yields the MLE $\hat{\theta}_i^{ML} = \hat{\theta}^{ML}(X_i)$

  - Asymptotic conditional bias is of order O(1/n)
  - Conditional mean squared error (MSE) with error term of order o(1/n)
  - Need the number of test items, n, large!

A. Likelihood Based Methods

2. WLE (Weighted Likelihood Estimates)

- **Obtained by maximizing the above likelihood, multiplied by a function of** θ
- **Has smaller bias properties as** $n \to \infty$
- **Smaller conditional bias than MLE, o(1/n) instead of O(1/n)**
- **But has same asymptotic conditional MSE.**

**** Please recall that n denotes the number of items or indicator variables**.

****** Please note that o() and O(), little-o and big-O, are used to symbolically express the asymptotic behavior of a given function; see Landau's formalism for "on the order of".

B. Posterior Distribution Based Methods

The posterior distribution of θ can be written as

$$P(\theta \mid x_i) = \frac{P(x_i \mid \theta) g(\theta)}{\int P(x_i \mid \theta) g(\theta) d\theta}$$

1. Bayes Modal Predictor – also known as the *maximum a posteriori* (MAP) predictor
   - This is obtained by maximizing the above posterior distribution for each subject
   - Conditional bias and unconditional bias is of order O(1/n)

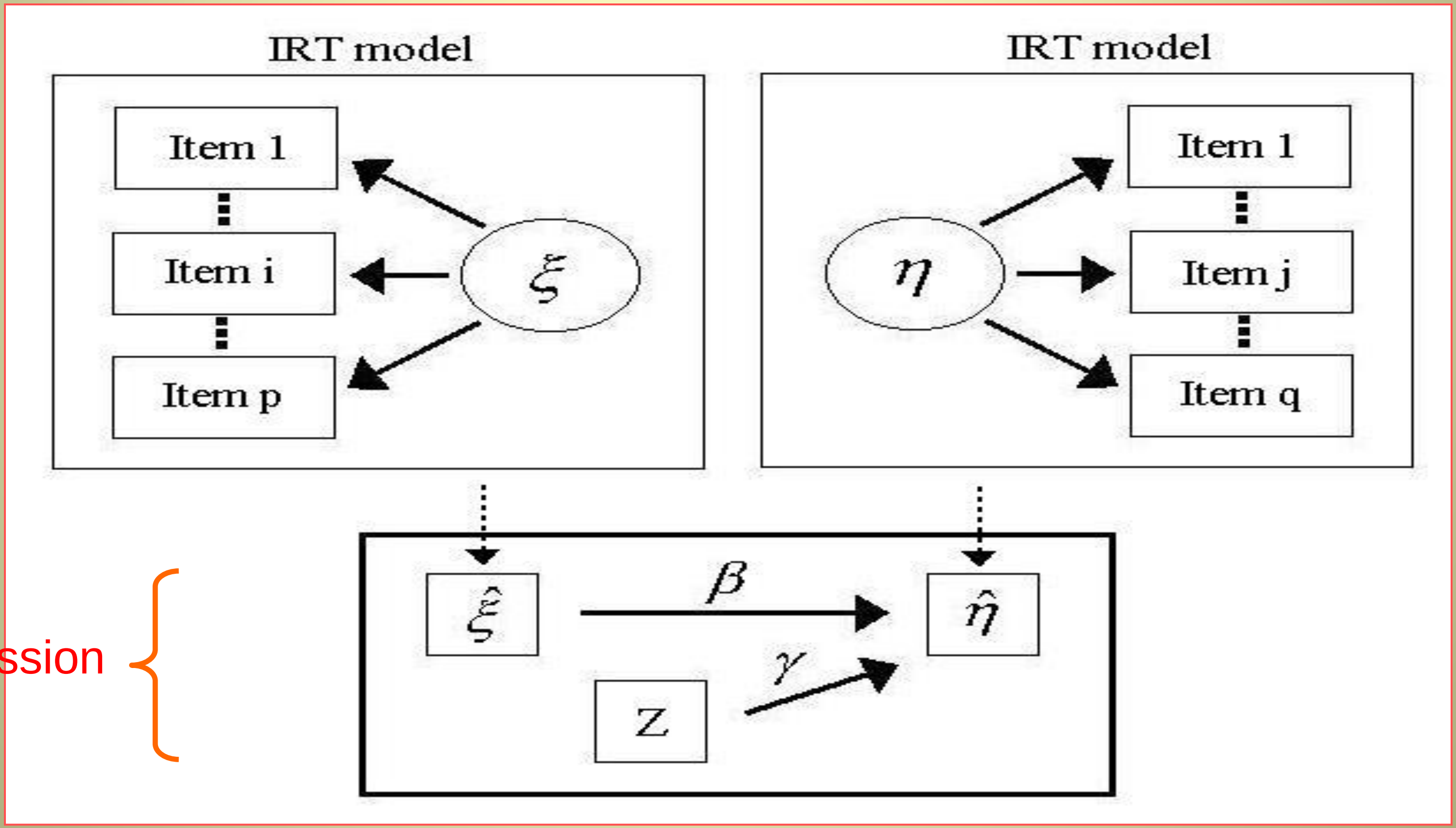
** Please recall that n denotes the number of items or indicator variables.

2. **Bayes Posterior Mean: also known as *expected a posteriori (EAP)* predictor**

- **EAP is the mean of the posterior distribution**
- **Sample mean (for i.i.d. samples) of the EAP predictor is an unbiased estimator of the mean of the ability distribution**
- **However, the unconditional variance < unconditional variance of the latent ability distribution**
- **This holds for the empirical Bayes predictor for large samples. A potentially troublesome property unless the number of test items is sufficient large.**

II. Graphical Depiction of Regression with Predicted Variables

- For example, use the EAP prediction $\hat{\eta}$ and $\hat{\xi}$ from the data file
- Perform the regression with $\hat{\eta}$, $\hat{\xi}$, and Z



Regression

The Problem with Predicting the Scores

- **Predicted ability scores are susceptible to a major source of error which is a function of the number of items (n) in the test.**

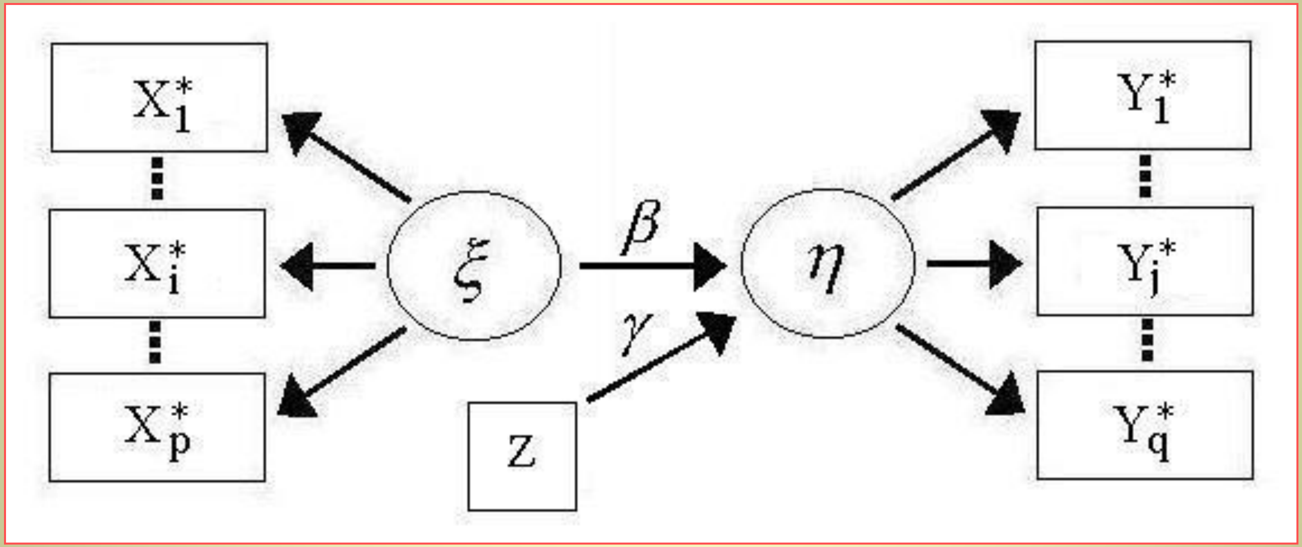
- **For a test of fixed n, the distribution of the predicted values (scores) does not converge to the distribution of the latent variable as sample size (N) goes to infinity (Lord, 1965, 1969; Little & Rubin, 1983; Louis, 1984).**

- **Must have number of items n go to infinity.**

- **Thus, analyses based on predicted ability values (scores) are subject to bias of the resultant regression parameter estimates.**

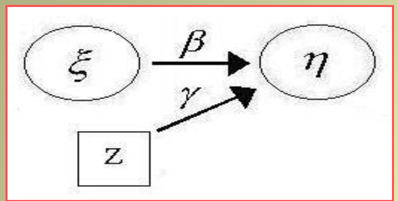
III. Embedding IRT in a Structural Equation Modeling (SEM)

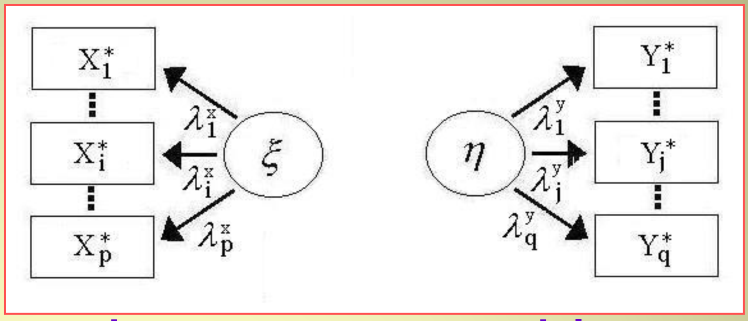
– **Simultaneous estimation without predicting latent variable scores. For example:**



Notice that this is a SEM on discrete variables

Specifically, we have





the structural model:

$$\eta = \beta\xi + \gamma Z + \zeta$$

the measurement models:

$$x^* = \Lambda_X \xi + \delta$$
$$x_i = 1 \text{ if } x_i^* \geq \tau_i$$
$$x_i = 0 \text{ otherwise}$$

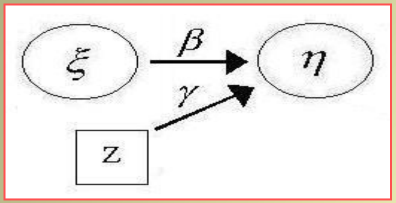
$$y^* = \Lambda_Y \eta + \varepsilon$$
$$y_j = 1 \text{ if } y_j^* \geq \tau_j$$
$$y_j = 0 \text{ otherwise}$$

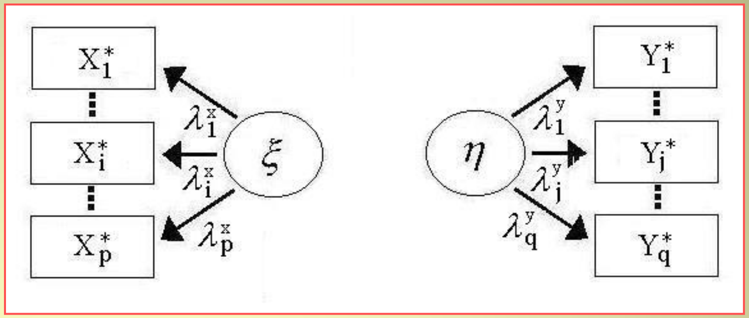
- **Estimate** $\tau's, \lambda's, \beta, \gamma$ **by optimizing a quadratic discrepancy measure** $(s-\sigma)'W^{-1}(s-\sigma)$

❖ **Benefits:**

- **Obtains consistent estimates of fixed parameters**
- **No predictions of individual scores**

Let's take a closer look





the structural model:

$$\eta = \beta\xi + \gamma Z + \zeta$$

the measurement models:

$$x^* = \Lambda_X \xi + \delta$$
$$x_i = 1 \text{ if } x_i^* \geq \tau_i$$
$$x_i = 0 \text{ otherwise}$$

$$y^* = \Lambda_Y \eta + \varepsilon$$
$$y_j = 1 \text{ if } y_j^* \geq \tau_j$$
$$y_j = 0 \text{ otherwise}$$

We all recognize this part.

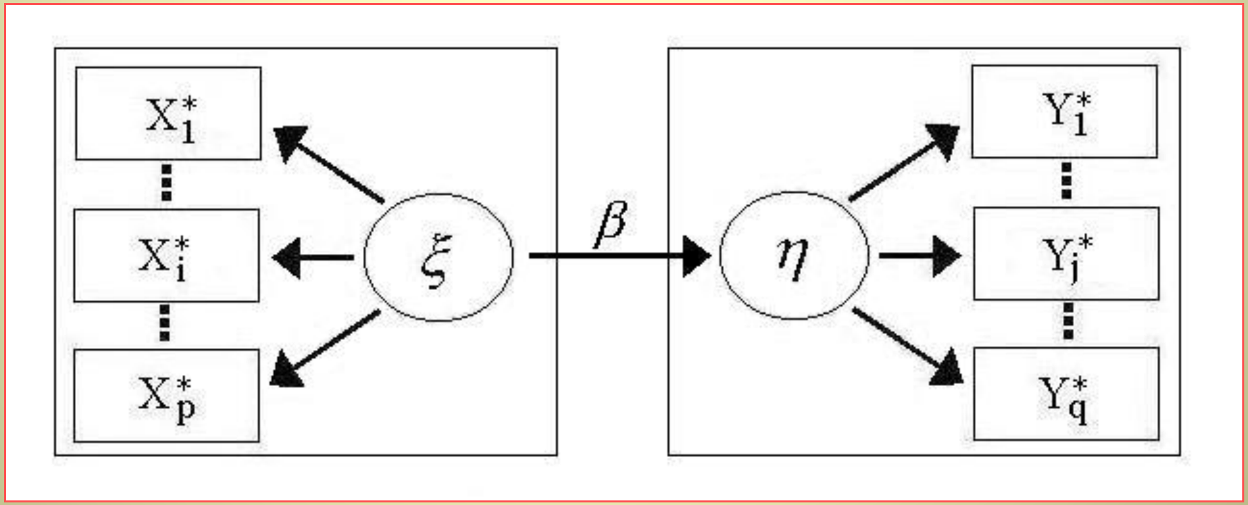
But what is this part?
It is a factor analysis on discrete variables.

This example is for binary items

- **But it is true that:**

Discrete factor analysis $\equiv$ Normal Ogive IRT model

- **And so, in this approach, when we perform SEM, we are actually fitting a simultaneous IRT/SEM model**



Some Remarks

Thus, parameter estimation in the discrete SEM is equivalent to simultaneous estimation of IRT item parameters and the latent regression parameters.

Will get consistent estimators of the structural parameters (Browne and Arminger, 1995) for finite number of items.

This approach side-steps the prediction of the latent variable scores.

Along the way, you can also see how one can order the predictors in terms of importance using a method introduced by Zumbo (2007).

AN EXAMPLE OF EMBEDDING IRT IN SEM

Method

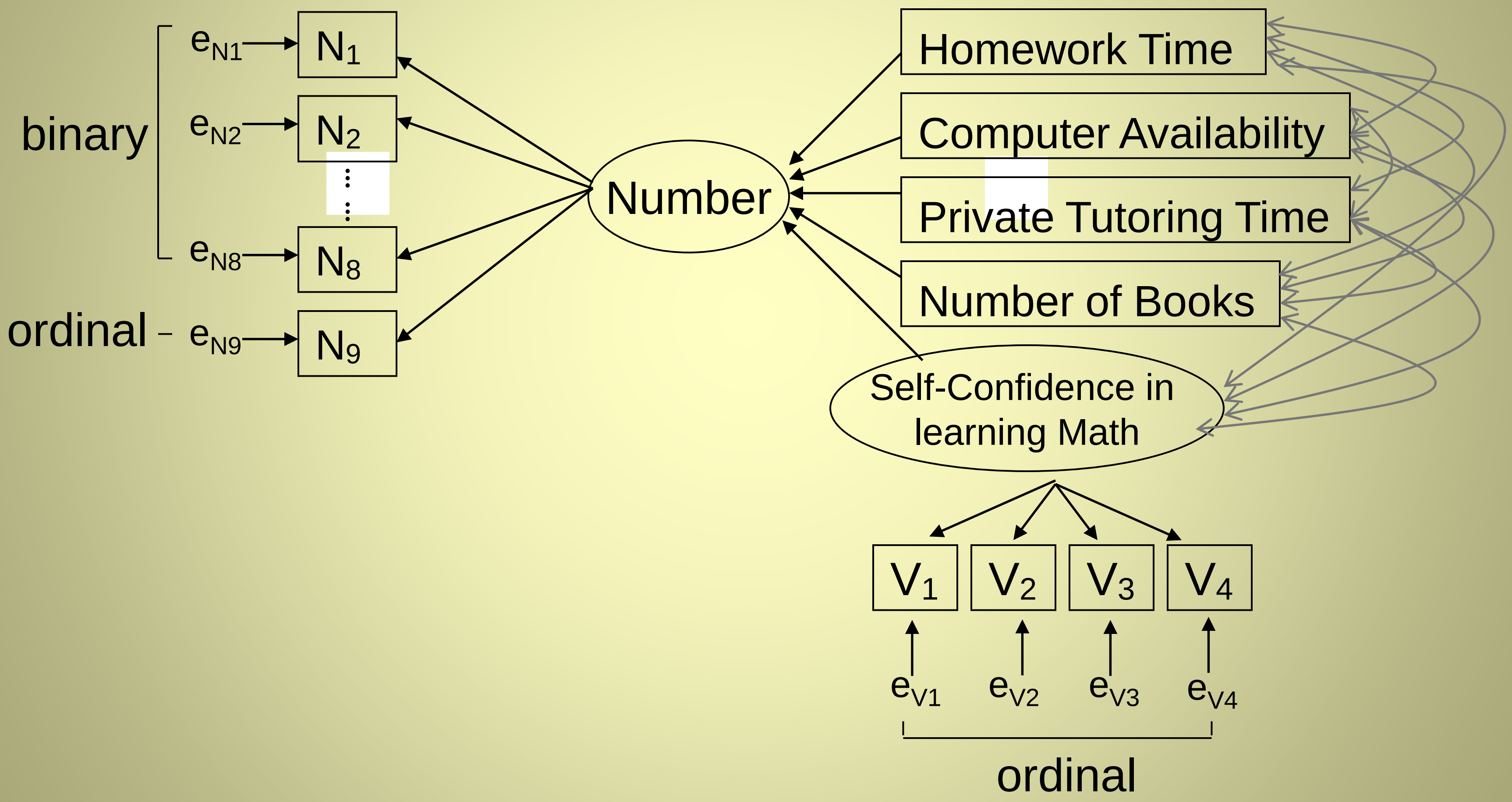
Instrument

- TIMSS 2003 grade-8 mathematics tests were used as our mathematic achievement measure. A total of 12 booklets were used in the test. As for the purpose of demonstration, we only used one booklet (booklet 8) in the present study.
- Student's questionnaire was used to obtain students' background variables.

Sample

- A total of 682 USA participants responding to booklet 8 with 324 boys and 358 girls.

Example

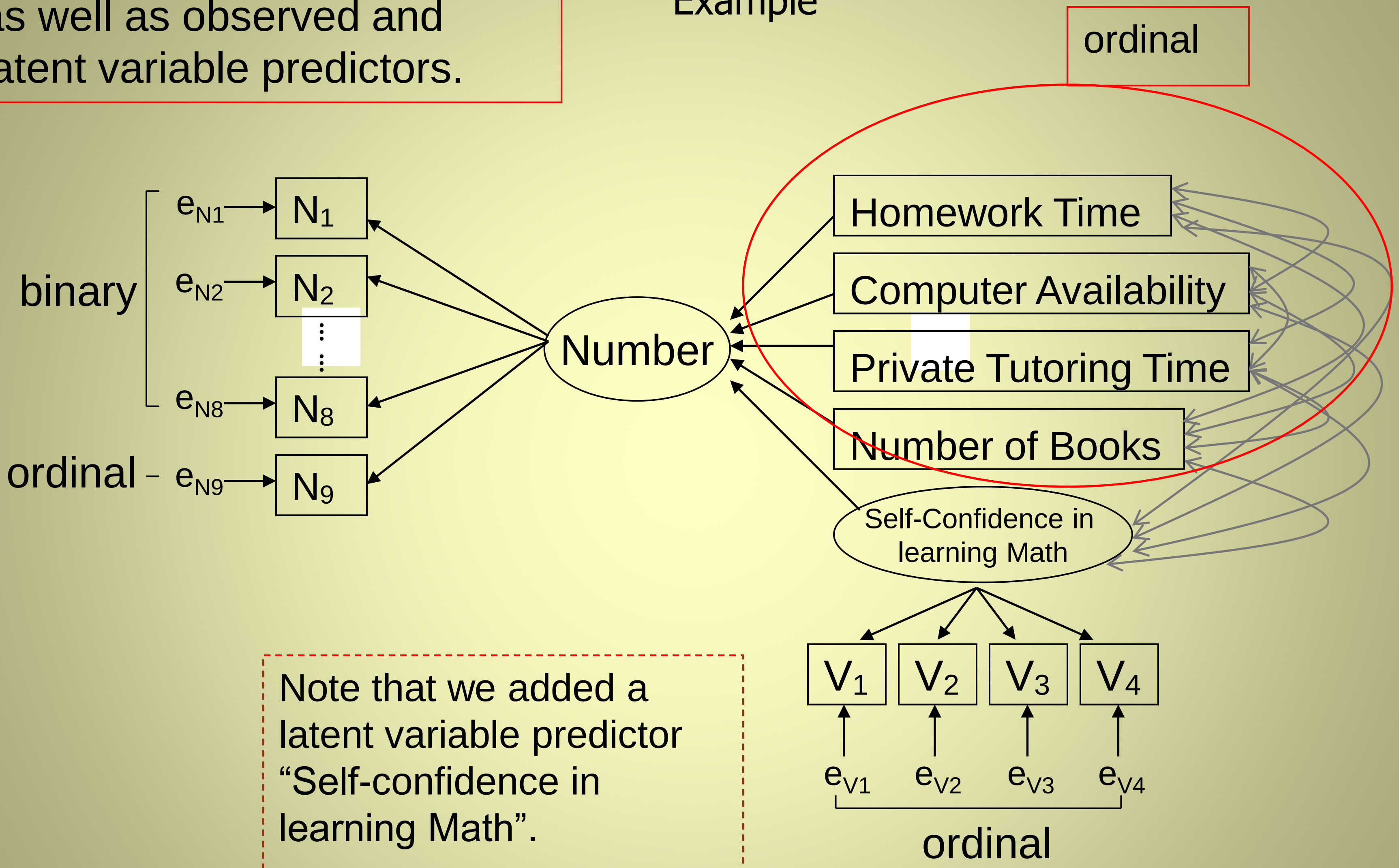

binary
ordinal
eN1
eN2
eN8
eN9
N1
N2
N8
N9
Number
Homework Time
Computer Availability
Private Tutoring Time
Number of Books
Self-Confidence in learning Math
V1
V2
V3
V4
eV1
eV2
eV3
eV4
ordinal

How is the model estimated?

Please note that throughout we used the *Mplus* framework and structural equation (covariance) modeling software to fit the MIMIC models and get the appropriate regression and correlation coefficients <u>using the correct correlation matrix</u> involving binary, ordinal, and continuous observed variables (Muthén & Muthén, 2010, Version 6).

Modeling of ordinal
and binary item responses
as well as observed and
latent variable predictors.

Example



Note that we added a latent variable predictor "Self-confidence in learning Math".

Results

Model-
DV: Number Knowledge

IV	b_j	r	Pratt
timehw	0.157	0.210	0.072
avlcompu	0.170	0.260	0.097
tutorhr	-0.219	-0.244	0.117
numbook	0.275	0.354	0.213
self-conf	0.479	0.517	0.542
R-square	0.457		

"Self-confidence in learning math" is the most important predictor.

These 4 variables together = 0.458

Within this set "Number of Books" is most important

Note:
Zumbo (2007) describes the latent variable Pratt index as $d_j = \frac{\hat{b}_j \times r_j}{R^2}$,

Conclusions

- **Predicted latent variable scores should be used with caution**
  - **in discrete case, large number of items needed**
  - **In NLSCY, the bias in EAP variance found by Thomas and Cyr (2002) for 20 items is equivalent to a R^2 bias of 40%.**
  - **for continuous items, Skrondal & Laake (2001) conclude that *"conventional factor score regression performs badly and should definitely be abandoned."***

Conclusions

- **In discrete case, simultaneous IRT-SEM approach may provide an alternative to plausible value methods**
  - **IRT-SEM models do not require large set of conditioning variables**
  - **sample size requirements may be limiting in some contexts**
  - **however, using reliable calibrated items will get better convergence even with small sample size**

References

- Brunner, L. J., & Austin, P. C. (2009). Inflation of type I error rate in multiple regression when independent variables are measured with error. *Canadian Journal of Statistics, 37(1)*, 33-46.
- Cochran, W. G. (1968). Errors of measurement in statistics. *Technometrics, 10*, 637-666.
- Fuller, W. A. (1987). *Measurement error models*. New York: Wiley.
- Little, R. J. A., & Rubin, D. B. (1983). On jointly estimating parameters and missing data by maximizing the complete data likelihood, *American Statistician, 37*, 218–220.
- Louis, T. A. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *Journal of the American Statistical Association, 79(386)*, 393–398.
- Lu , I.R.R., Thomas, D.R., & Zumbo, B.D. (2005) Embedding IRT in Structural Equation Models: A Comparison With Regression Based on IRT Scores. *Structural Equation Modeling, 12:2*, 263-277.
- Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17*, 131–154.
- Rubin, D. B. (1987). *Multiple imputation for non-response in surveys*. New York: Wiley.
- Shear, B. R., & Zumbo, B.D. (2013). False Positives in Multiple Regression: Unanticipated Consequences of Measurement Error in the Predictor Variables. *Educational and Psychological Measurement, 73*, 733-756.
- Zumbo, B.D. (2007). Validity: Foundational Issues and Statistical Methodology. In C.R. Rao and S. Sinharay (Eds.) *Handbook of Statistics, Vol. 26: Psychometrics*, (pp. 45-79). Elsevier Science B.V.: The Netherlands.

THE END!

Citation:

Zumbo, B.D. (2021). Embedding IRT in Structural Equation Models as a Solution to Problems with Measurement Error in Predictor Variables in Validation Research [Video]*. UBC-Paragon Webinar Series 2021-1*. Vancouver, B.C.

https://www.youtube.com/channel/UCCkvlPbkn15FVDKRplm-fuw