Discussion of “Rethinking Composite Indices: Reliability, Practical Alternatives, and an Application to Political Economy” by Daniel L. Millimet and Alfredo R. Paloyo

Michael Darden

November 8th, 2025

Examples Latent Variables in Economics

- **Utility**: Pareto (1906), Samuelson (1938), Debreu (1954, 1959)
- **Human Capital**: Becker (1964), Mincer (1974), Griliches (1977), Cunha *et al.* (2010).
- **Health**: Grossman (1972), Cutler & Richardson (1997), Gilleskie (1998), Dreider & Pepper (2007).
- **Social Capital**: Arrow (1972), Putnam (1993, 2000), Guiso *et al.* (2004), Tabellini (2008)

This Paper

Let X^* be a latent variable, and suppose you want to estimate:

$$y = \alpha + \beta X^* + w'\delta + \epsilon \tag{1}$$

This Paper

Let X^* be a latent variable, and suppose you want to estimate:

$$y = \alpha + \beta X^* + w'\delta + \epsilon \tag{1}$$

Instead of observing X^*, suppose you observed measures $z = z_1, \ldots, z_J$ and let

$$x = g(z)$$

This Paper

Let X^* be a latent variable, and suppose you want to estimate:

$$y = \alpha + \beta X^* + w'\delta + \epsilon \tag{1}$$

Instead of observing X^*, suppose you observed measures $z = z_1, \ldots, z_J$ and let

$$x = g(z)$$

Index-Creation Problem:

- Choose $g(\cdot)$ such that x is a *useful* proxy for X^* in Equation 1.
- Useful = consistent estimates of β and δ.
- Focus on linear indices because of literature.
  - $x = \sum_j \lambda z_j$

This paper represents a blend of "how-to" and novel econometrics.

Modeling Choices and Choice Set of $g(\cdot)$

Reflective Model

$$z_j = x^* + u_j \tag{2}$$

Formative Model

$$x^* = \sum_{j=1}^{\mathcal{J}} \lambda_j^* z_j \tag{3}$$

Modeling Choices and Choice Set of $g(\cdot)$

Reflective Model

$$z_j = x^* + u_j \tag{2}$$

Formative Model

$$x^* = \sum_{j=1}^{\mathcal{J}} \lambda_j^* z_j \tag{3}$$

Index Choices:

1. **Principal Component Analysis**
2. **Equal Weights**
3. **Mean Z-Score**
4. **Partial Least Squares (PLS)**
5. **Exploratory Factor Analysis**
6. **Outcome Driven Approaches.**
    - Lubotsky and Wittenberg (2006) and Yang, Jia, and Li (2023), and IV.

Evaluation Methods

1. Simulations.
2. Two applications:
    - Ortoleva and Snowberg (*AER*, 2015): Overconfidence in beliefs → ideological extremeness, increased voter turnout, and stronger partisan identification.
    - Republican share of 2020 and 2024 vote as a function of area level health.

Preview of Results

1. **Please stop using PCA!**
    - PCA yields the most attenuated associations.
2. Outcome-aligned approaches dominate on mean square error and coverage in nearly all designs.
3. Index mismeasurment causes bias in all coefficients.
4. Measurement error is nonclassical and attenuation bias is not guaranteed.
5. Bottom line: “index choice is an identification choice”

Super Helpful Intuition

Super Helpful Intuition

My most recent R&R

> *Would it be possible to explore the available information to minimize measurement error, perhaps by using factor analysis or principal component analysis?*

Preliminaries

Setup

Let X^* be a latent variable, and suppose you want to estimate:

$$y = \alpha + \beta X^* + w'\delta + \epsilon \tag{4}$$

Setup

Let X^* be a latent variable, and suppose you want to estimate:

$$y = \alpha + \beta X^* + w'\delta + \epsilon \tag{4}$$

Manifest Variables: Z

- $\mathcal{J} =$ all manifest variables related to X^*.
- $J =$ count of observed manifest variables.

Linear index of standardized Z:

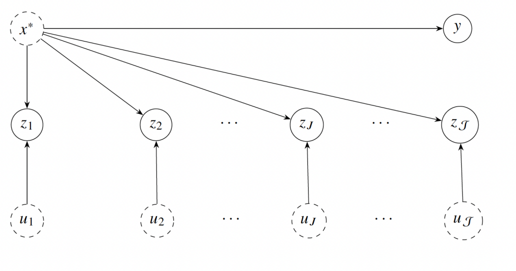
$$x = \sum_j \lambda_j \Big(\frac{z_j - \overline{z}_j}{\sigma_{z_j}} \Big)$$

Linear index of non-standardized:

$$x = \sum_j \lambda_j z_j$$

MD Guidance on when to standardize? By normalizing the variance of Z, you restrict the factor loadings.

Reflective Model

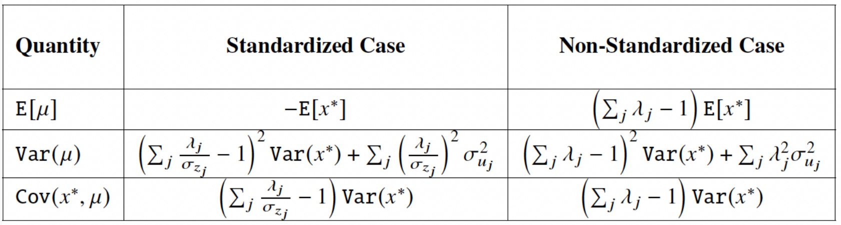


$$z_j = x^* + u_j \tag{5}$$

1. $E(u_j) = 0 \; \forall j$
2. $V(u_j) = \sigma^2_{u_j} \; \forall j$
3. $Cov(u_j, u_{j'}) = 0 \; \forall \; j \neq j'$
4. $Cov(X^*, u_j) = 0 \; \forall \; j$
5. $Cov(\epsilon, u_j) = 0 \; \forall \; j$
6. $E(wu_j) = 0 \; \forall \; j$

Index Error, Non-Standardized

$$\mu = X^* - X = \Big(\sum_j \lambda_j - 1\Big)x^* + \sum_j \lambda_j u_j \qquad (6)$$



Quantity	**Standardized Case**	**Non-Standardized Case**
$\texttt{E}[\mu]$	$-\texttt{E}[x^*]$	$\left(\sum_j \lambda_j - 1\right) \texttt{E}[x^*]$
$\texttt{Var}(\mu)$	$\left(\sum_j \frac{\lambda_j}{\sigma_{z_j}} - 1\right)^2 \texttt{Var}(x^*) + \sum_j \left(\frac{\lambda_j}{\sigma_{z_j}}\right)^2 \sigma^2_{u_j}$	$\left(\sum_j \lambda_j - 1\right)^2 \texttt{Var}(x^*) + \sum_j \lambda_j^2 \sigma^2_{u_j}$
$\texttt{Cov}(x^*, \mu)$	$\left(\sum_j \frac{\lambda_j}{\sigma_{z_j}} - 1\right) \texttt{Var}(x^*)$	$\left(\sum_j \lambda_j - 1\right) \texttt{Var}(x^*)$

Data-Generating Process

$$y = \alpha + \beta X^{*} + w'\delta + \epsilon \tag{7}$$

- $E[\epsilon] = 0$
- $E[\tilde{w}\epsilon] = 0$

where $\tilde{w} = [x^{*}w]$. OLS is unbiased and consistent estimator of β.

$$\text{plim}\,\hat{\beta}_{OLS} = \beta \left[\frac{\text{Var}(x^{*})\left(1 - R^{2}_{x^{*}|w}\right) + \text{Cov}(x^{*}, \mu) - \text{Cov}(x^{*}, w)' \left[\text{Var}(w)\right]^{-1} \text{Cov}(w, \mu)}{\text{Var}(x)\left(1 - R^{2}_{x|w}\right)} \right] \tag{8}$$

Index Creation

Table of contents

1. Preliminaries

2. Index Creation

    PCA

    Other Methods

3. Simulation

4. Replication

    Overconfidence in Political Behavior

    Health and Political Polarization

5. Comments

The Case Against PCA

$$x^{PCA} = \sum_{j=1}^{J} \nu_j z_j \tag{9}$$

where

$$\nu^{PCA} = \arg\max_{||\nu||=1} \nu' \sum_z \nu$$

- Ignores outcome y and latent x^* → unsupervised.
- Overweights noisy indicators if they have high variance.
- Induces nonclassical measurement error:

  Bias can be in any direction, not just attenuation.

- “Convenience is not neutrality.” (Millimet & Paloyo)

Table of contents

1. Preliminaries

2. Index Creation

   PCA

   Other Methods

3. Simulation

4. Replication

   Overconfidence in Political Behavior

   Health and Political Polarization

5. Comments

1. Unit Weights

$$X^{\overline{z}} = 1/J \sum_j z_j \tag{10}$$

MD This is very common in health applications:

- CHADs score in atrial fibrillation.
- PDQ9 score for depression.

Mean z-Score

Common alternative to unit-roots

$$x^{\overline{\tilde{z}}} = \frac{1}{J} \sum_{j} \frac{z_j - \overline{z}_j}{\sigma_j} \tag{11}$$

Partial Least Squares

Orthogonal linear combinations of the Zs

$$x^{PLS} = \sum_j v_j^{PLS} z_j \tag{12}$$

Where the v

$$v^{PLS} = argmax[Cov(v'z, y)]^2 \tag{13}$$

subject to a normalization.

Lubotsky and Wittenberg (2006)

$$y = \alpha + \sum_j \beta_j z_j + w'\delta + \epsilon \tag{14}$$

produces $\hat{\beta}^{LW} = \sum_j \beta_j$

> *This result has seemingly gone unnoticed by applied researchers despite Lubotsky and Wittenberg (2006, p. 549) warning that the prevailing practice of creating summary measures such as PCA is "generally ad hoc and hardly ever optimal".*

Only valid in reflective model with classical measurement error.

Summary

- Replacing a latent construct with a proxy index generally leads to biased estimates of the coefficient on the latent construct.
- The index error is nonclassical and thus the estimated coefficient on the proxy index is not guaranteed to be attenuated.
- The direction and magnitude of the bias for the estimated coefficient on the proxy index depends on the underlying structural relationship between the manifest variables and the latent construct.

Simulation

Simulation Design

Sample of size N with J measures

$$y_i = \alpha + \beta x_i^* + w_i'\delta + \epsilon \tag{15}$$

Set $\alpha = 0$ and $\beta = 1$

$$z_{ji} = \omega_j x_i^* + u_{ij} \tag{16}$$

Parameters:

- $N \in \{500, 2000\}$
- $\delta \in \{0, 0.5\}$
- $x^*, w \sim N(\Upsilon, \sum_{uu})$
- $\{\omega_1, \omega_2, \omega_3\} = \{1, 1, 1\}, \{1, 1.5, 0.5\}, \{1, 5, 0.5\}, \{1, 10, 0.5\},$
- $u_1, \ldots, u_j \sim N_J(0, \sum_{uu})$
- $J = \{3, 5, 10, 25, 50\}$

Simulation

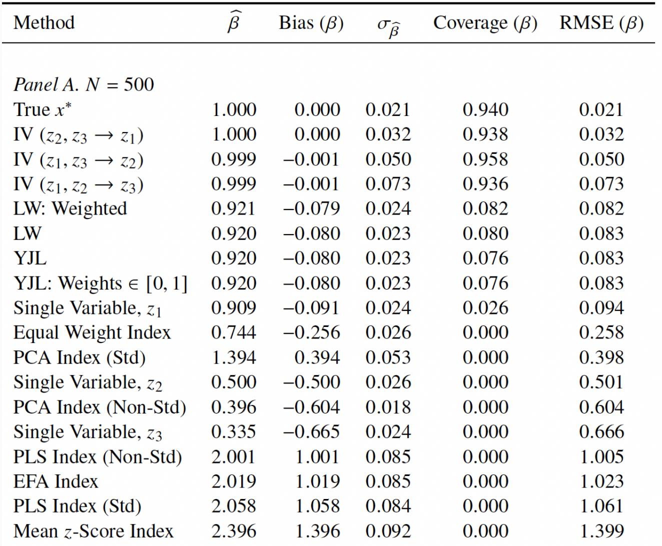
Evaluation

1. Mean bias
2. Mean Standard Deviation
3. Root Mean Squared Error
4. Empirical Coverage of nominal 95% CI

Estimators

1. Benchmark: true X^*.
2. OLS including each manifest variable individually in the regression,
3. OLS including an index obtained using the first principal component, with and without standardization.
4. OLS including an equally weighted average of Z with and without standardization.
5. OLS including the first component from partial least squares with and without standardization using the unit variance normalization,
6. OLS including the first factor score from EFA obtained after standardization (EFA Index),
7. Lubotsky and Wittenberg (2006).
8. Yang, Jia, and Li (2023)
9. IV using each combination of J-1 manifest variables to instrument for the remaining indicator

Some Simulation Results



Method	$\widehat{\beta}$	Bias (β)	$\sigma_{\widehat{\beta}}$	Coverage (β)	RMSE (β)
Panel A. $N = 500$					
True x^*	1.000	0.000	0.021	0.940	0.021
IV ($z_2, z_3 \rightarrow z_1$)	1.000	0.000	0.032	0.938	0.032
IV ($z_1, z_3 \rightarrow z_2$)	0.999	−0.001	0.050	0.958	0.050
IV ($z_1, z_2 \rightarrow z_3$)	0.999	−0.001	0.073	0.936	0.073
LW: Weighted	0.921	−0.079	0.024	0.082	0.082
LW	0.920	−0.080	0.023	0.080	0.083
YJL	0.920	−0.080	0.023	0.076	0.083
YJL: Weights $\in [0, 1]$	0.920	−0.080	0.023	0.076	0.083
Single Variable, z_1	0.909	−0.091	0.024	0.026	0.094
Equal Weight Index	0.744	−0.256	0.026	0.000	0.258
PCA Index (Std)	1.394	0.394	0.053	0.000	0.398
Single Variable, z_2	0.500	−0.500	0.026	0.000	0.501
PCA Index (Non-Std)	0.396	−0.604	0.018	0.000	0.604
Single Variable, z_3	0.335	−0.665	0.024	0.000	0.666
PLS Index (Non-Std)	2.001	1.001	0.085	0.000	1.005
EFA Index	2.019	1.019	0.085	0.000	1.023
PLS Index (Std)	2.058	1.058	0.084	0.000	1.061
Mean z-Score Index	2.396	1.396	0.092	0.000	1.399

Summary of Simulation Results

1. Lubotsky and Wittenberg (2006) approach seems highly recommended
2. PLS with non-standardized manifest variables does well when the manifest variables vary significantly in scale and the index is not the focus.
3. PCA never seems advantageous.
4. Mean z-score Index performs worse when additional noisy manifest variables are added.

Replication

Table of contents

1. Preliminaries

2. Index Creation

   PCA

   Other Methods

3. Simulation

4. Replication

   Overconfidence in Political Behavior

   Health and Political Polarization

5. Comments

Ortoleva and Snowberg (2015)

- **Media Index:** Individual's political information exposure
- **Overconfidence Index:** Latent trait of excessive certainty.

Authors use PCA for both indices. **MP**:

> *the sign of the behavioral relationships emphasized by Ortoleva and Snowberg (2015) is remarkably stable, but inferences about their size—and their sensitivity to additional controls capturing information sets—depend on how the latent constructs are operationalized.*

Table of contents

1. Preliminaries

2. Index Creation

   PCA

   Other Methods

3. Simulation

4. Replication

   Overconfidence in Political Behavior

   Health and Political Polarization

5. Comments

Millimet and Paloyo

At the Congressional District (n=429) level:

$$\Delta_{2020-2024}\text{Republican Vote Share} = f(Health, X) \tag{17}$$

Health index:

- PCA
- Factor Index
- PLS Index
- Equal Weights Index
- Mean z-Score Index
- Yang-Jia-Li
- Lubotsky-Wittenberg

Health Measures

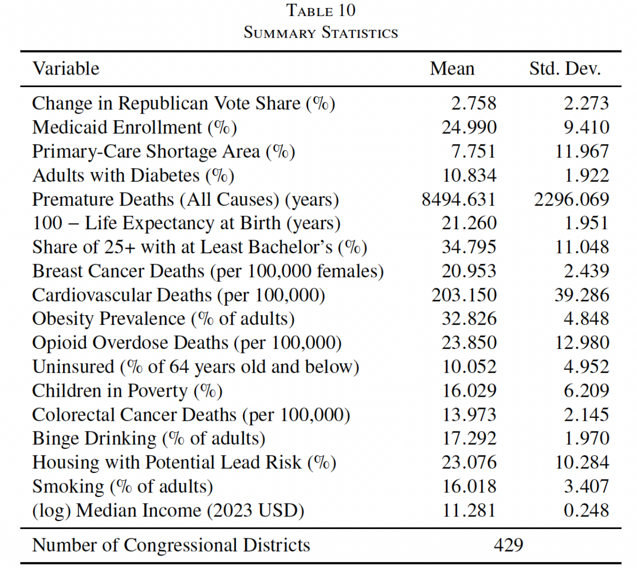


TABLE 10
SUMMARY STATISTICS

Variable	Mean	Std. Dev.
Change in Republican Vote Share (%)	2.758	2.273
Medicaid Enrollment (%)	24.990	9.410
Primary-Care Shortage Area (%)	7.751	11.967
Adults with Diabetes (%)	10.834	1.922
Premature Deaths (All Causes) (years)	8494.631	2296.069
100 – Life Expectancy at Birth (years)	21.260	1.951
Share of 25+ with at Least Bachelor's (%)	34.795	11.048
Breast Cancer Deaths (per 100,000 females)	20.953	2.439
Cardiovascular Deaths (per 100,000)	203.150	39.286
Obesity Prevalence (% of adults)	32.826	4.848
Opioid Overdose Deaths (per 100,000)	23.850	12.980
Uninsured (% of 64 years old and below)	10.052	4.952
Children in Poverty (%)	16.029	6.209
Colorectal Cancer Deaths (per 100,000)	13.973	2.145
Binge Drinking (% of adults)	17.292	1.970
Housing with Potential Lead Risk (%)	23.076	10.284
Smoking (% of adults)	16.018	3.407
(log) Median Income (2023 USD)	11.281	0.248
Number of Congressional Districts	429	

Results

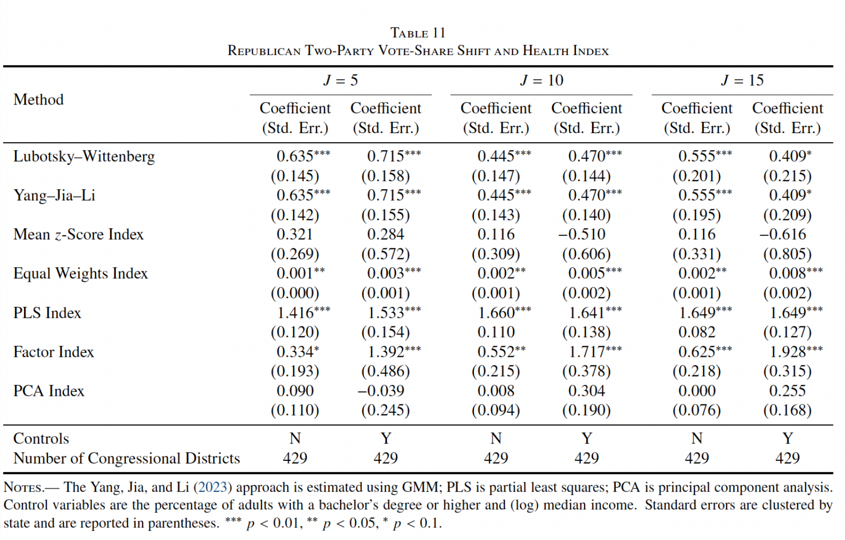


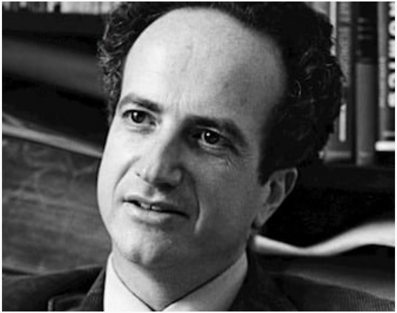
Table 11
Republican Two-Party Vote-Share Shift and Health Index

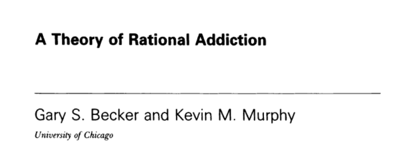
Method	$J = 5$		$J = 10$		$J = 15$	
	Coefficient (Std. Err.)	Coefficient (Std. Err.)	Coefficient (Std. Err.)	Coefficient (Std. Err.)	Coefficient (Std. Err.)	Coefficient (Std. Err.)
Lubotsky–Wittenberg	0.635***	0.715***	0.445***	0.470***	0.555***	0.409*
	(0.145)	(0.158)	(0.147)	(0.144)	(0.201)	(0.215)
Yang–Jia–Li	0.635***	0.715***	0.445***	0.470***	0.555***	0.409*
	(0.142)	(0.155)	(0.143)	(0.140)	(0.195)	(0.209)
Mean z-Score Index	0.321	0.284	0.116	−0.510	0.116	−0.616
	(0.269)	(0.572)	(0.309)	(0.606)	(0.331)	(0.805)
Equal Weights Index	0.001**	0.003***	0.002**	0.005***	0.002**	0.008***
	(0.000)	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)
PLS Index	1.416***	1.533***	1.660***	1.641***	1.649***	1.649***
	(0.120)	(0.154)	0.110	(0.138)	0.082	(0.127)
Factor Index	0.334*	1.392***	0.552**	1.717***	0.625***	1.928***
	(0.193)	(0.486)	(0.215)	(0.378)	(0.218)	(0.315)
PCA Index	0.090	−0.039	0.008	0.304	0.000	0.255
	(0.110)	(0.245)	(0.094)	(0.190)	(0.076)	(0.168)
Controls	N	Y	N	Y	N	Y
Number of Congressional Districts	429	429	429	429	429	429

Notes.— The Yang, Jia, and Li (2023) approach is estimated using GMM; PLS is partial least squares; PCA is principal component analysis. Control variables are the percentage of adults with a bachelor's degree or higher and (log) median income. Standard errors are clustered by state and are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Comments

Quick Aside

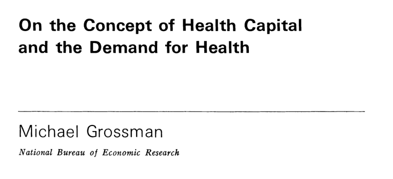




A Theory of Rational Addiction

Gary S. Becker and Kevin M. Murphy

University of Chicago



On the Concept of Health Capital and the Demand for Health

Michael Grossman

National Bureau of Economic Research

Quick Aside

A Dynamic Stochastic Model of Lifetime Smoking Behavior

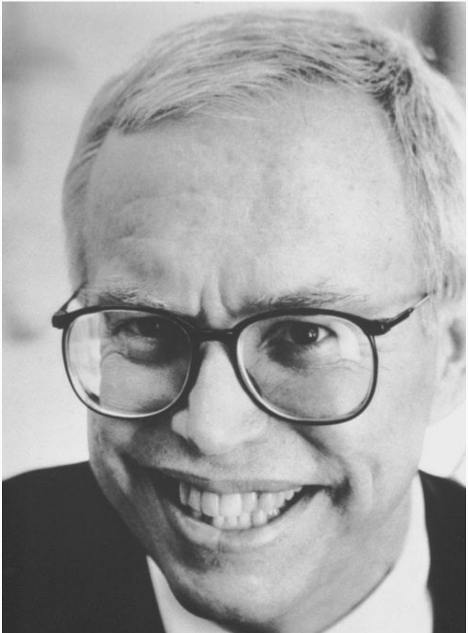
Michael Darden

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Economics.

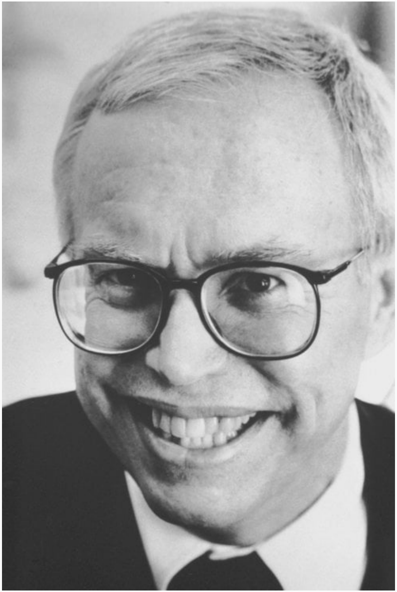
Chapel Hill
2011

Submission

Submission

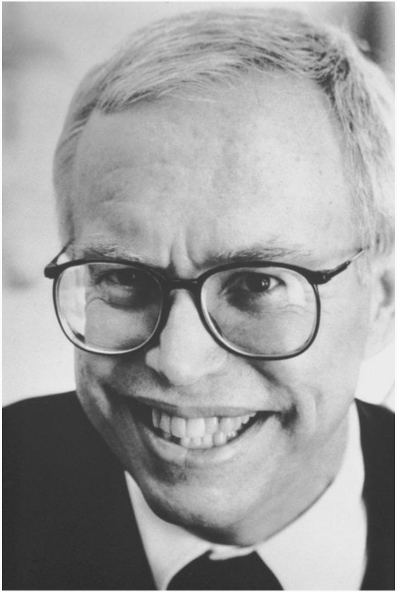


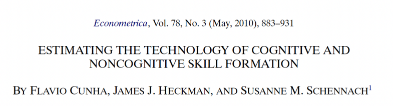
Submission



“PCA is Bullsh!t”

Submission





Econometrica, Vol. 78, No. 3 (May, 2010), 883–931

ESTIMATING THE TECHNOLOGY OF COGNITIVE AND NONCOGNITIVE SKILL FORMATION

BY FLAVIO CUNHA, JAMES J. HECKMAN, AND SUSANNE M. SCHENNACH[1]

- The latent variable is the object of interest.

Comment #1: Measurement System is Unspecified

$$z_j = \mu_j + \alpha_j X^* + u_j$$

Assumptions

1. $E(u_j) = 0 \ \forall j$
2. $V(u_j) = \sigma^2_{u_j} \ \forall j$
3. $Cov(u_j, u_{j'}) = 0 \ \forall \ j \neq j'$
4. $Cov(X^*, u_j) = 0 \ \forall \ j$
5. $\alpha_1 = 1$
6. $E(X^*) = 0$
7. $J > 2$

Consider estimating α_j for $j \in \{2, \ldots, J\}$ jointly with β and δ. E-M algorithm to estimate the joint likelihood function, integrating out the common latent factor.

Identification of the loadings: $\alpha_j = \frac{Cov(Z_j, Z_k)}{Cov(Z_1, Z_k)}$ for $j \neq k$

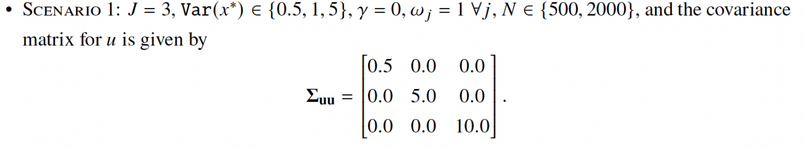
Comment #1a: Unclear to me how simulation works

In the $J = 3$ case, scenario 1:

- Fixes all three factor loadings to 1
- Fixes the mean and variance of X^*
- Assumes uncorrelated measurement errors.

Comment # 2: Correlated Errors

- “correlated measurement errors are likely to be the norm in practice”



- SCENARIO 1: $J = 3$, $\texttt{Var}(x^*) \in \{0.5, 1, 5\}$, $\gamma = 0$, $\omega_j = 1\ \forall j$, $N \in \{500, 2000\}$, and the covariance matrix for u is given by

$$\mathbf{\Sigma_{uu}} = \begin{bmatrix} 0.5 & 0.0 & 0.0 \\ 0.0 & 5.0 & 0.0 \\ 0.0 & 0.0 & 10.0 \end{bmatrix}.$$

Key result from Agostinelli and Wiswall (2025, JPE): Allowing for correlated errors significantly attenuates the degree of complementarity relative to Cunha, Heckman, and Schennach (2010, Econometrica).

Comment #3

This paper represents a blend of “how-to” and novel econometrics, and as such, it’s hard to identify an audience.