MIRT as an Alternative/Supplement to CFA for Assessing Latent Factor Structure and Measurement Invariance

Philip Osteen, PhD
University of Maryland
School of Social Work
January 17, 2010

Outline

- CTT vs. IRT
- Participation in a Social Work Community of Practice (PSWCoP) Scale
- CFA Analysis
- MIRT Analysis
- Discussion/Q&A

CTT vs. IRT

- Have a shared goal
- Use different methods of evaluation
- Provide different types of information
- Not equivalent

CTT vs. IRT

- Strengths of CTT (Allen, & Yen, 2002; Fan, 1998)
  - Intuitive interpretation
  - Ease of calculation and application
  - Focused on test-level evaluation
  - Fewer (and less stringent) assumptions

- Strengths of IRT (Embretson, & Reise, 2000; Hambleton, et al., 1991; Harvey, & Hammer, 1999)
  - Sample independent
  - Log transformation creates interval-level data out of dichotomous/ordinal-level data
  - Assessment of item- and test-level bias
  - Greater flexibility in types of models

Participation in a Social Work Community of Practice (PSWCoP) Scale

- Based on Wenger's (1998) Community of Practice Theory
- Designed to address 3 types of motivation for participation (Wenger, McDermott, & Snyder, 2002)
- Domain Subscale, Community Subscale, & Practice Subscale
- 18 items
- Likert response format

Sample

- Purposive sample of 17 CSWE-accredited programs around the country
- Convenience sample of 528 MSW students

Item Selection

- Domain Subscale – items removed based on internal consistency (CFA) and MNSQ Infit (IRT) results
- Community Subscale – items removed based on internal consistency (CFA) and MNSQ Infit (IRT) results
- Practice Subscale – items removed based on internal consistency (CFA) and MNSQ Infit (IRT) results
  - reevaluated using EFA, yielding two factors
    - Competency Subscale
    - Skills Subscale

CFA

- Model Evaluation
  - Model Fit
  - Analysis of Nested Models
  - Parameter Estimation
  - Measurement Invariance

CFA

- Estimating Model Fit (Jackson, et al., 2009; Sun, 2005; Kline, 2005)
  - Population-based
    - RMSEA
    - CFI
  - Sample-based
    - Model χ^2
    - SRMR
    - AIC
    - GFI

CFA

- Analysis of Nested Models
  - Individual Model Fit
  - Relative Model Fit
  - Theory

CFA

- Lisrel 8.8 used to evaluate nested models
- 3 models tested (trimmed items)
  - 4-factor model with cross-loadings
  - 4-factor model without cross-loadings
  - Theoretical 3-factor model

CFA

- Analysis of Nested Models

	Model 1: 4 Factors w/Cross-Loadings	*Model 2: 4 Factors w/o Cross-Loadings**	*Model 3: 3 Factors w/o Cross-Loadings***
$\chi^2_{(df)}$	$64.48_{(35)}$	$185.52_{(48)}$	$359.90_{(51)}$
p-value (model)	.00175	<.001	<.001
$\chi_1^2 - \chi_2^2{}_{(df1 - df2)}$		$121.04_{(13)}$	$174.38_{(3)}$
p-value (model diff)		<.001	<.001
RMSEA	.042	.077	0.112
CFI	0.98	.91	0.8
SRMR	0.043	.094	0.12
AIC	150.48	245.82	413.90
GFI	0.97	0.91	0.85
		*Compared to Model 1	**Compared to Model 2

CFA

- Analysis of Nested Models

	Model 1: 4 Factors w/Cross-Loadings	*Model 2: 4 Factors w/o Cross-Loadings**	*Model 3: 3 Factors w/o Cross-Loadings***
$\chi^2_{(df)}$	$64.48_{(35)}$	$185.52_{(48)}$	$359.90_{(51)}$
p-value (model)	.00175	<.001	<.001
$\chi_1^2 - \chi_2^2{}_{(df1 - df2)}$		$121.04_{(13)}$	$174.38_{(3)}$
p-value (model diff)		<.001	<.001
RMSEA	.042	.077	0.112
CFI	0.98	.91	0.8
SRMR	0.043	.094	0.12
AIC	150.48	245.82	413.90
GFI	0.97	0.91	0.85

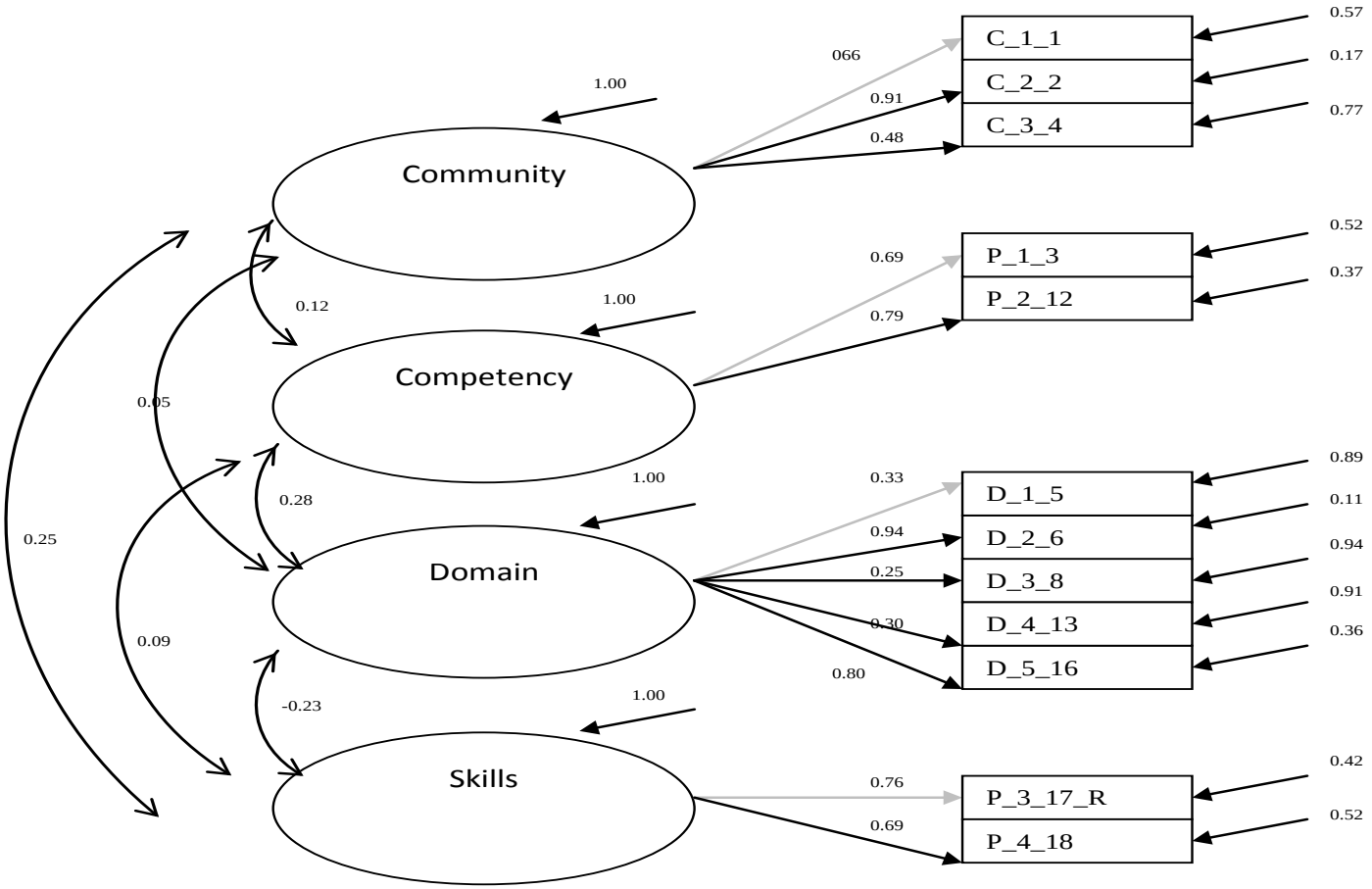
*Compared to Model 1

**Compared to Model 2

CFA

- Parameter Estimates
  - Parameter estimation is done by comparing the actual covariance matrix representing the relationships between variables and the estimated covariance matrices of the best fitting model
  - A variety of estimation procedures available

CFA


Community
Competency
Domain
Skills
1.00
1.00
1.00
1.00
066
0.91
0.48
0.69
0.79
0.33
0.94
0.25
0.30
0.80
0.76
0.69
C_1_1
C_2_2
C_3_4
P_1_3
P_2_12
D_1_5
D_2_6
D_3_8
D_4_13
D_5_16
P_3_17_R
P_4_18
0.57
0.17
0.77
0.52
0.37
0.89
0.11
0.94
0.91
0.36
0.42
0.52
0.12
0.05
0.25
0.28
0.09
-0.23

CFA

- Measurement Invariance (Horn & McArdle, 1992)
  - Total Invariance: Everything constrained across groups (Are group covariance matrices essentially the same)
  - Metric Invariance: Factor loadings and uniqueness terms (and item intercepts) fixed. Factor variances and covariances free across group.
  - Configural Invariance: Uniqueness terms fixed. Factor variances, covariances, loadings free across group.
  - Not reasonable to test means with configural invariance allowed because we are no longer sure whether factors measure same trait across group.
  - Begin with Total Invariance and compare to progressively less restrictive models

CFA

- Two sample characteristics selected for assessment of invariance based on qualitative analysis
  - Academic Standing
    - Foundation
    - Advanced (includes Concentration and Advanced Standing)
  - Age
    - Under 30
    - Over 30

CFA

- Measurement Invariance for Academic Standing

	Total Invariance	*Metric Invariance**	*Metric Invariance **	*Configural Invariance****
$\chi^2_{(df)}$	296.95(126)	235.59(115)	213.67(102)	184.13(93)
p-value (model)	<.001	<.001	<.001	<.001
$\chi_1^2 - \chi_2^2{}_{(df1-df2)}$		60.66(11)	21.91(13)	29.54(9)
p-value (model diff)		<.001	>.05	<.001
RMSEA	.077	.073	.067	.067
CFI	.85	.89	.90	.91
SRMR	.11	.09	.084	.082
AIC	348.75	315.76	311.81	310.11
GFI	.89	.92	.92	.93
		*Free Factor Var and Covar	**Free Factor Var and Covar; Free Item Var	***Free Factor Var and Covar; Free Item Var; Free Factor Loadings

CFA

- Measurement Invariance for Age Group

	Total Invariance	*Metric Invariance**	*Metric Invariance **	*Configural Invariance****
$\chi^2_{(df)}$	327.96(126)	293.65(116)	234.81(107)	204.12(95)
p-value (model)	<.01	<.001	<.001	<.001
$\chi_1^2 - \chi_2^2{}_{(df1 - df2)}$		34.31(10)	58.85(9)	30.69(12)
p-value (diff)		<.001	<.001	<.01
RMSEA	.096	.09	.078	.073
CFI	.81	.83	.88	.90
SRMR	.19	.15	.12	.10
AIC	437.96	399.02	345.23	327.56
GFI	.82	.84	.88	.90
		*Free Factor Var and Covar	**Free Factor Var and Covar; Free Item Var	***Free Factor Var and Covar; Free Item Var; Free Factor Loadings

MIRT

- Model Evaluation
  - Model Fit
  - Analysis of Nested Models
  - Parameter Estimation
  - Measurement Invariance

MIRT

- Estimating Model Fit (Wright, & Masters, 1982, Hambleton, et al., 1991; Wu, et. al, 2002)
  - Model χ^2
  - Item-fit/Person-fit
  - Deviance statistic (for model comparison)

MIRT

- Analysis of Nested Models

	Four Factor (Between)	Three Factor* (Between)	Two Factor* (Within)	One Factor*
Deviance (G^2)	17558.64	17728.83	17963.99	17962.52
Parameters	26	22	19	17
G^2_Δ		-170.19	-405.35	-403.88
$Parameter_\Delta$		4	7	9
$G^2_\Delta/Parameter_\Delta$		42.55	57.91	44.88
p-value		< 0.001	< 0.001	< 0.001

* Compared to the Four Factor, Between-Items Model

MIRT

- Parameter Estimation
  - Parameter estimation is done by comparing observed responses to expected responses based on a specific IRT model
  - A variety of estimation procedures available

MIRT

- Measurement Invariance (Wu, et al., 2002)
  - DIF
    - An indication of whether or not an item performs the same for members of different groups who have the same level of ability
    - X^2 test of item, group, and item*group parameter estimates
    - DIF values may be positive (+) or negative (–); positive DIF values indicate that it was easier for the reference group to endorse an item, while negative DIF values indicate that it was easier for the comparison group to endorse an item.
    - Wilson (2005) classifies the magnitude of DIF as
      - "negligible" (DIF<0.426)
      - "intermediate" (0.426<DIF<0.638)
      - "large" (0.638<DIF)

MIRT

- Measurement Invariance (Wu, et al., 2002)
  - Step calibrations for response categories
    - chi-square difference statistic (χ^2_D) for models with constrained versus non-constrained step calibrations

MIRT

- DIF for Academic Standing
  - Based on the item*group analysis, there was evidence of DIF ($x^2_{(22)}$=123.75, p<0.001)
  - Seven items met the criterion for DIF as defined as [(estimate/error)>2]
  - Magnitude of DIF is "negligible" (Wilson, 2005)

Item	(Est/Err)	DIF	Magnitude
C_1_1	3.23	-0.139	Negligible
C_2_2	2.84	-0.146	Negligible
P_2_12	3.93	0.072	Negligible
D_2_6	3.76	0.293	Negligible
D_3_8	2.79	-0.136	Negligible
D_5_16	3.37	0.228	Negligible
P_4_17_R	2.69	-0.184	Negligible

MIRT

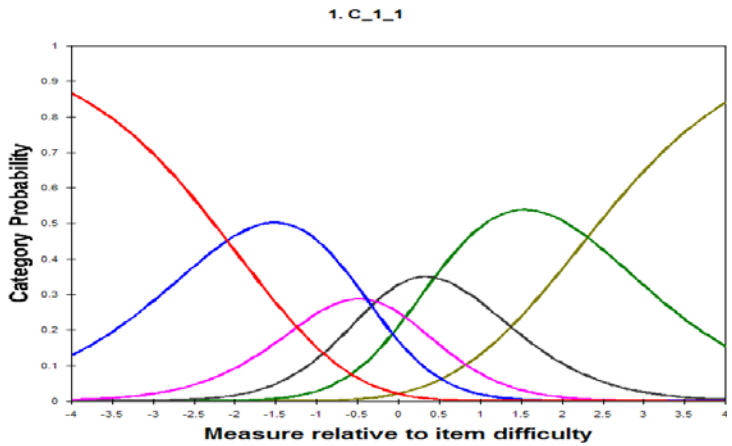
- Step Calibrations for Academic Standing

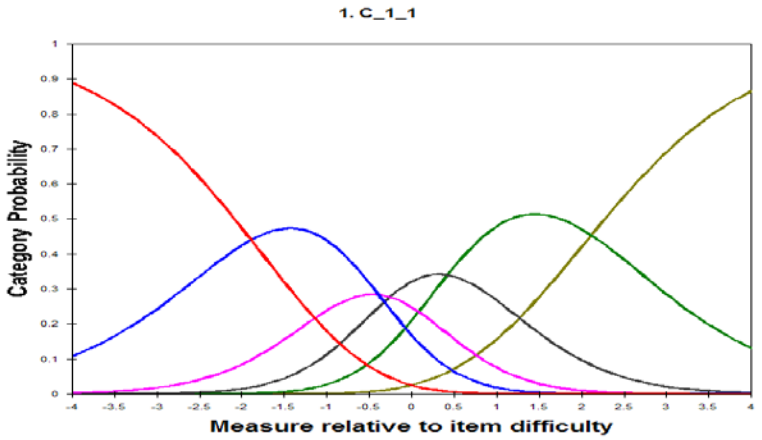
	Invariant Step Structure	Free Step Structure
Deviance (G^2)	15568	15493
Parameters	73	118
G^2_Δ		75
$Parameter_\Delta$		45
$G^2_\Delta/Parameter_\Delta$		1.67
p-value		.003

MIRT

- Step Calibrations for Academic Standing

Step	Foundation Students	Advanced Students
1	NONE	NONE
2	-1.72	-1.38
3	-.02	-.18
4	.10	-.02
5	.81	.68
6	2.68	2.15





MIRT

- DIF for Age Groups
  - Based on the item*group analysis, there was no evidence of DIF ($x^2_{(11)}=16.44$, $p=0.125$)

MIRT

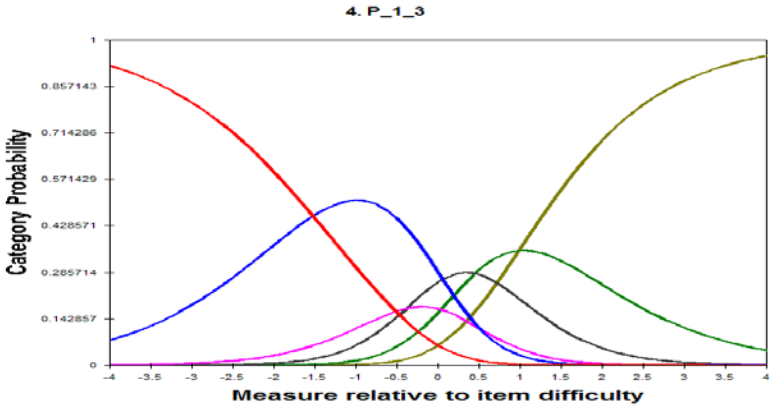
- Step Calibrations for Age Group

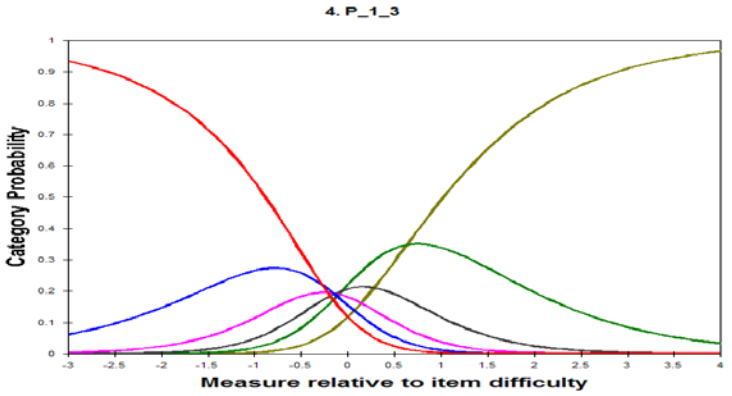
	Invariant Step Structure	Free Step Structure
Deviance (G^2)	15926	15827
Parameters	73	118
G^2_Δ		99
$Parameter_\Delta$		45
$G^2_\Delta/Parameter_\Delta$		2.2
p-value		< 0.001

MIRT

- Step Calibrations for Age Group

Step	Under 30	Over 30
1	NONE	NONE
2	-1.53	-.22
3	.49	-.10
4	-.39	-.08
5	.45	-.01
6	.98	.68





Results – CFA vs. MIRT

- Factor Structure
  - PSWCoP is a multidimensional measure
  - 4-factor model with unique item loadings is the best model when considering model fit and theory

Results – CFA vs. MIRT

- Measurement Invariance for Academic Standing
  - CFA
    - Configural invariance at a minimum
    - Evidence suggests metric invariance
  - MIRT
    - Evidence of DIF on 7 out of 12 items
      - "Neglible"
    - Evidence is ambiguous regarding step calibration invariance
- Decision
  - Further evaluation of step calibration between groups is needed
  - Any inferential test of group differences should be interpreted with caution

Results – CFA vs. MIRT

- Measurement Invariance for Age Group
  - CFA
    - Configural invariance at best
  - MIRT
    - No evidence of DIF
    - Evidence is ambiguous regarding step calibration invariance
- Decision
  - Evidence is ambiguous regarding measurement invariance
  - Further exploration needed
  - Inferential tests of group differences may not be appropriate

Discussion

- Both methods informative for establishing factor structure
- Both measures informative for testing different types of measurement invariance
- CFA may be more informative for evaluating measures at the factor level
- MIRT may be more informative for evaluating measures at the item level

References

Allen, M.J., & Yen, W. M. (2002*). Introduction to measurement theory.* Long Grove, IL: Waveland Press.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

Fan, X. (1998). Item response theory and classical test theory: An empirical consideration of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357-382.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.

Harvey, R.J., & Hammer, A.L. (1999). Item response theory. *The Counseling Psychologist, 27*(3), 353-383.

Horn, J.L., & McArdle, J.J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*(3), 117-144.

Jackson, D.L., Gillaspy, J.A, & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods, 14*(1), 6-23.

Kline, R.B. (2005). *Principles and practices of structural equation modeling* (2nd ed.). New York: Guilford Press.

Sun, J. (2005). Assessing goodness of fit in confirmatory factor analysis. *Measurement and Evaluation in Counseling and Development, 37*(4), 240-256.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity.* New York: Cambridge University Press.

Wenger, E., McDermott, R., & Snyder, W.M. (2002). *Cultivating communities of practice.* Boston: Harvard Business School Press.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, N.J.: Lawrence Erlbaum Associates.

Wright, B.D., & Masters, G.N. (1982). *Rating scale analysis: Rasch measurement.* Chicago: MESA Press.

Wu, M.L., Adams, R.J., & Wilson, M., & Haldane, S. (2008). ACER *Conquest* 2.0: Generalized item response modeling software [computer program]. Hawthorn, Australia: ACER.

Contact Information

Philip Osteen, PhD
University of Maryland
School of Social Work
525 W. Redwood St.
Baltimore, MD, 21201
410-706-3612
posteen@ssw.umaryland.edu