Valid Measurement Without Factorial Invariance: A Longitudinal Example

Michael C. Edwards[1] R. J. Wirth[2]

[1] Department of Psychology
The Ohio State University

[2] Division of General Internal Medicine
University of Washington

June 18, 2010

Today’s Talk

- Measurement in longitudinal contexts
- Differential item functioning
- Differential dimensionality
- Constructive non-invariance

Theory

“Usually a well-developed theory contains one or more formal models which give concrete structure to the general concepts of the theory. These models may be viewed as explications of portions of the general theory. Such models, in turn, are connected systematically with directly observable phenomena. The function of such models is to permit the logical deduction of general and specific relationships that have not been empirically demonstrated but that may be demonstrable.”

Theory

“Usually a well-developed theory contains one or more formal models which give concrete structure to the general concepts of the theory. These models may be viewed as explications of portions of the general theory. Such models, in turn, are connected systematically with directly observable phenomena. The function of such models is to permit the logical deduction of general and specific relationships that have not been empirically demonstrated but that may be demonstrable.”

Lord & Novick, 1968, p. 15

Measurement Models

“Theoretical constructs are often related to the behavioral domain through observable variables by considering the latter as *measures* or *indicants* of the former. And conversely, theoretical constructs are often abstracted from given observable variables. We shall call an observable variables a *measure* of a theoretical construct if its expected value is presumed to increase monotonically with the construct.”

Lord & Novick, 1968, pp. 19-20

Classical Test Theory

One of the earliest and most widely used measurement models in psychology was true score theory, also commonly known as classical test theory (CTT).

The fundamental equation of true score theory is

$$x_i = \tau_i + e_i, \tag{1}$$

where x_i is the observed test score for person i, τ_i is the *true score* for person i, and e_i is the error.

Classical Test Theory

One of the earliest and most widely used measurement models in psychology was true score theory, also commonly known as classical test theory (CTT).

The fundamental equation of true score theory is

$$x_{ij} = \tau_{ij} + e_{ij}, \tag{2}$$

where x_{ij} is the observed test score for person i on exam j, τ_{ij} is the *true score* for person i on exam j, and e_{ij} is the error.

Item response theory

IRT is a collection of latent variable models that explain the process by which people respond to items in terms of item parameters and person parameters.

Item response theory

IRT is a collection of latent variable models that explain the process by which people respond to items in terms of item parameters and person parameters.

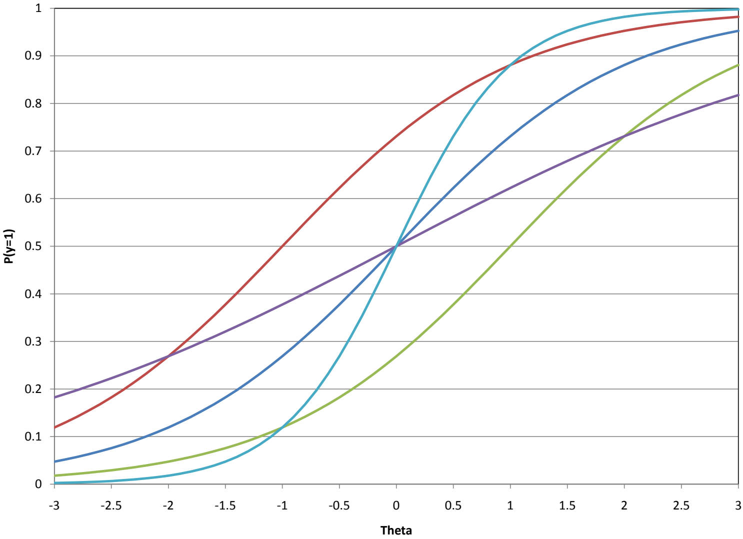
There are also many strong connections between IRT and other models like factor analysis or nonlinear mixed models.

The 2-Parameter Logistic Model

$$P(u_k = 1|\theta) = \frac{1}{1 + \text{exp}[a_k(\theta - b_k)]} \tag{3}$$

- u_k is the observed response to item k
- a_k is the slope for item k
- b_k is the location parameter for item k
- θ is the latent construct we're measuring

2PLM Trace Line



The 2-Parameter Logistic Model

$$P(u_{ik} = 1|\theta_i) = \frac{1}{1 + \text{exp}[a_k(\theta_i - b_k)]} \tag{4}$$

- u_{ik} is the observed response of person i to item k
- a_k is the slope for item k
- b_k is the location parameter for item k
- θ_i is the latent score of person i

Reliability and Validity

Reliability and validity are central concepts to the idea of measurement. As stated by Thissen & Wainer, 2001, p.11:

“...we can define *validity* as measuring the right thing, and *reliability* as measuring the thing right.”

Validity

"Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment."

Messick, 1993, p.13

A slightly new take on types of validity

There is a rich literature on validity and there is already a fairly well established taxonomy. For the purposes of today's talk, I would like to propose two new categories of validity:

1. Statistical Validity
2. Substantive Validity

Statistical Validity

The basic idea of statistical validity is that whatever model we are using allows us to obtain accurate representations of the constructs we are interested in.

I consider this different from substantive validity, which I'm defining as the provision of evidence that the result of said model is really what we think it is. I think substantive validity is really just an umbrella term under which most of the commonly known forms of validity (e.g., content, construct, discriminant, etc.) fall. I'll say more about this later.

Statistical Validity in Longitudinal Settings

Although the statistical validity issue isn't limited to longitudinal settings, it is one of the more obvious places where problems could arise.

To model any construct over time, one must have scores that represent that same construct over time. This seemingly simple task can become tremendously complicated when time is added to the mix.

Statistical Validity in Longitudinal Settings

One potential threat to statistical validity is the need to change items as a construct develops over time. As someone ages it is common for the best indicators of a particular construct to change.

Statistical Validity in Longitudinal Settings

One potential threat to statistical validity is the need to change items as a construct develops over time. As someone ages it is common for the best indicators of a particular construct to change.

Consider delinquency as an example. It is fairly easy to imagine that different sets of items may be used to measure delinquency for children who are not yet in school, children who are in a school setting, and adults who are out of school.

Statistical Validity in Longitudinal Settings

One potential threat to statistical validity is the need to change items as a construct develops over time. As someone ages it is common for the best indicators of a particular construct to change.

Consider delinquency as an example. It is fairly easy to imagine that different sets of items may be used to measure delinquency for children who are not yet in school, children who are in a school setting, and adults who are out of school.

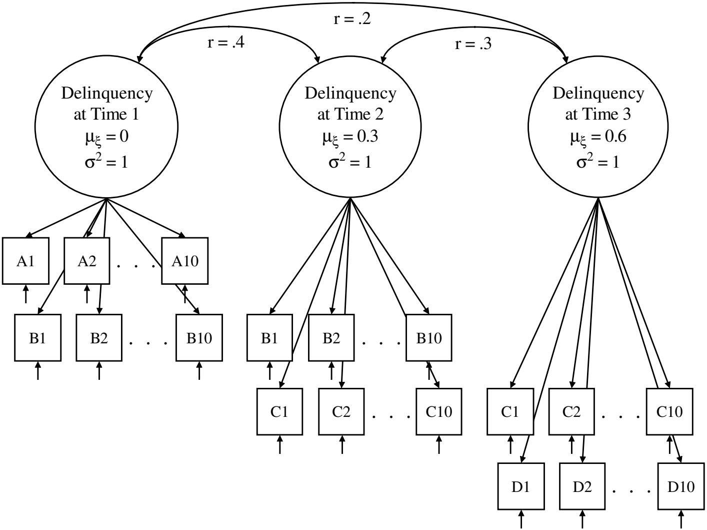
In CTT approaches unit weights are applied to create things like proportion scores. This is fine if all the items happen to be equally related to the construct and equally severe.

An example

Imagine that we would like to assess delinquency in children at ages 8, 12, and 16. We have four scales at our disposal to accomplish this task. Each is 10 items long and geared towards a particular age range:

- Scale A: 4-8
- Scale B: 8-12
- Scale C: 12-16
- Scale D: 16-20

An example


r = .2
r = .4
r = .3
Delinquency at Time 1
$\mu_\xi = 0$
$\sigma^2 = 1$
Delinquency at Time 2
$\mu_\xi = 0.3$
$\sigma^2 = 1$
Delinquency at Time 3
$\mu_\xi = 0.6$
$\sigma^2 = 1$
A1
A2
. . .
A10
B1
B2
. . .
B10
B1
B2
. . .
B10
C1
C2
. . .
C10
C1
C2
. . .
C10
D1
D2
. . .
D10

Results - Means

Generating Values	0.0	0.3	0.6

Results - Means

Generating Values	0.0	0.3	0.6
Mean Scores	1.99	2.17	2.19

Results - Means

Generating Values	0.0	0.3	0.6
Mean Scores	1.99	2.17	2.19
IRT Scores	-0.04	0.24	0.57

Results - Means

Generating Values	0.0	0.3	0.6
Mean Scores	1.99	2.17	2.19
IRT Scores	-0.04	0.24	0.57

As long as we do some planning, we are able to add and remove items and still maintain comparability of scores.

We haven't been doing this too much in psychology, but in education this forms the foundation of almost every high stakes test in existence.

Measurement Non-invariance (a.k.a. DIF)

Another threat to statistical validity is measurement non-invariance, or more conveniently, differential item functioning (DIF).

Measurement Non-invariance (a.k.a. DIF)

Another threat to statistical validity is measurement non-invariance, or more conveniently, differential item functioning (DIF).

DIF implies that an item's relation to the construct is changing over some level of covariate (e.g., gender, ethnicity). Age/time is ubiquitous in longitudinal modeling and therefor care must be taken to insure that items relate to the construct in the same manner at all ages/times.

Measurement Non-invariance (a.k.a. DIF)

Another threat to statistical validity is measurement non-invariance, or more conveniently, differential item functioning (DIF).

DIF implies that an item's relation to the construct is changing over some level of covariate (e.g., gender, ethnicity). Age/time is ubiquitous in longitudinal modeling and therefor care must be taken to insure that items relate to the construct in the same manner at all ages/times.

If DIF is present, but not modeled, then the latent scores contain both "true" variability due to differences among individuals in their levels of the latent construct as well as some sort of bias introduced by the fact that the same response means different things at different times.

Equating to the rescue (again)

From the model's standpoint, whether you add/remove items or change the parameters assigned to a particular item is irrelevant - the model sees them both the same way.

This means that equating allows us to model changing relationships between an item and the construct while still maintaining score comparability and statistical validity.

Statistical Validity continued...

DIF is one thing, but surely we can't obtain statistically valid scores when the dimensionality is changing over time!?!?

Statistical Validity continued...

DIF is one thing, but surely we can't obtain statistically valid scores when the dimensionality is changing over time!?!?

Why not?

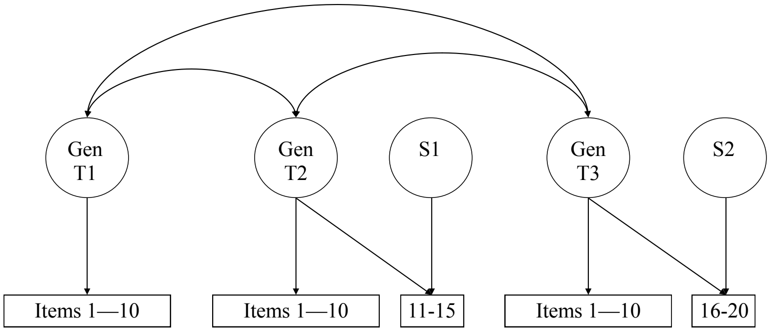
Differential Dimensionality

It seems quite possible that in some cases, the dimensionality that we're dealing with at any given time point is also changing.

For example, one could imagine that a set of items related to school-based behaviors may load on a "school" factor in addition to a delinquency factor.

If another set of items is added to track work-based behaviors, at later time points you may find a "work" factor in addition to delinquency.

An example



Gen T1
Gen T2
S1
Gen T3
S2
Items 1—10
Items 1—10
11-15
Items 1—10
16-20

The simulation design

We generated N=3,000 simulees from the above model using parameters culled from a literature review of published IRT parameters in psychological journals. The generating latent structure looked like this:

	G_{t1}	G_{t2}	G_{t3}	S_{t2}	S_{t3}
G_{t1}	1				
G_{t2}	0.44	1.2			
G_{t3}	0.24	0.39	1.4		
S_{t2}	0	0	0	1	
S_{t3}	0	0	0	0	1
μ	0	0.3	0.6	0	0

The (pre-)Results

	RMSE	
Parameter	N=3,000	N=300
Gen a	0.05	0.2
Spec a	0.05	0.31
b1	0.04	0.13
b2	0.03	0.12
b3	0.03	0.1
b4	0.04	0.13

The Results

Generating

	G_{t1}	G_{t2}	G_{t3}
G_{t1}	1		
G_{t2}	0.44	1.2	
G_{t3}	0.24	0.39	1.4
μ	0	0.3	0.6

N=3,000

	G_{t1}	G_{t2}	G_{t3}
G_{t1}	1		
G_{t2}	0.44	1.14	
G_{t3}	0.2	0.38	1.37
μ	0	0.31	0.59

N=300

	G_{t1}	G_{t2}	G_{t3}
G_{t1}	1		
G_{t2}	0.44	1.43	
G_{t3}	0.23	0.43	1.32
μ	0	0.23	0.52

The Results

Generating

	G_{t1}	G_{t2}	G_{t3}
G_{t1}	1		
G_{t2}	0.4	1	
G_{t3}	0.2	0.3	1
μ	0	0.3	0.6

N=3,000

	G_{t1}	G_{t2}	G_{t3}
G_{t1}	1		
G_{t2}	0.41	1	
G_{t3}	0.17	0.3	1
μ	0	0.31	0.59

N=300

	G_{t1}	G_{t2}	G_{t3}
G_{t1}	1		
G_{t2}	0.37	1	
G_{t3}	0.2	0.31	1
μ	0	0.23	0.52

Differential dimensionality and DIF

Just to be thorough, we simulated a situation where there is both differential dimensionality and DIF.

The Results

Generating

	G_{t1}	G_{t2}	G_{t3}
G_{t1}	1		
G_{t2}	0.44	1.2	
G_{t3}	0.24	0.39	1.4
μ	0	0.3	0.6

N=3,000

	G_{t1}	G_{t2}	G_{t3}
G_{t1}	1		
G_{t2}	0.46	1.17	
G_{t3}	0.29	0.39	1.41
μ	0	0.31	0.56

N=300

	G_{t1}	G_{t2}	G_{t3}
G_{t1}	1		
G_{t2}	0.36	1.11	
G_{t3}	0.3	0.35	1.43
μ	0	0.31	0.56

The Results

Generating

	G_{t1}	G_{t2}	G_{t3}
G_{t1}	1		
G_{t2}	0.4	1	
G_{t3}	0.2	0.3	1
μ	0	0.3	0.6

N=3,000

	G_{t1}	G_{t2}	G_{t3}
G_{t1}	1		
G_{t2}	0.42	1	
G_{t3}	0.24	0.3	1
μ	0	0.31	0.56

N=300

	G_{t1}	G_{t2}	G_{t3}
G_{t1}	1		
G_{t2}	0.34	1	
G_{t3}	0.25	0.28	1
μ	0	0.31	0.56

Substantive Validity

If you weren’t convinced before, hopefully you know believe that we can accommodate a wide variety of “differential-ness” in our measurement models and still obtain statistically valid scores. That’s only part of the challenge though. To be good scientists, we must also provide evidence that those scores are substantively valid as well.

If we have reason to suspect that DIF will occur in a particular situation, this provides an opportunity to generate some very compelling substantive validity evidence.

Constructive Non-invariance

We suspect that researchers, if they sat down and thought about it, would have some expectations about what *should* happen if they are measuring what they think they are measuring.

By using relevant theory it should be possible to make predictions about how the items will change in their relationship to the construct over time.

Constructive Non-invariance

In an IRT context, this could look something like:

- *Defying parents* should have a lower slope for 12 year olds than for 8 year olds
- *Defying parents* should have lower thresholds for 12 year olds than for 8 year olds
- *Hitting* should have a higher slope for 16 year olds than for 12 year olds
- *Hitting* should have higher thresholds for 16 year olds than for 12 year olds

Constructive Differential Dimensionality?

If this is true for DIF, why couldn’t it be true for differential dimensionality?

The example illustrated here is a fairly simple one, but we hope that by showing researchers that these kinds of models are possible it will generate some creative thinking about areas where there may be differential dimensionality.

Final Thoughts

Our statistical frameworks are very capable of supporting changing item sets, DIF, and differential dimensionality - even in the complex landscape of longitudinal data.

What remains a challenge is demonstrating that constructs are what we claim they are. This is heightened if the operationalization of a construct changes over time.

The End

Thanks.

`edwards.134@osu.edu`

References

Edwards, M. C., & Wirth, R. J. (2009). Measurement and the study of change. *Research in Human Development*, *6*, 74-96.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Messick, S. (1993). Validity. In R. L. Linn (Ed.), *Educational measurement* (p. 13-103). Phoenix, AZ: The Oryx Press.

Thissen, D., & Wainer, H. (2001). An overview of test scoring. In D. Thissen & H. Wainer (Eds.), *Test scoring* (p. 1-19). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.