A Quantitative Inquiry: How Do We Really Measure Change in Children's Development?

Jennifer E.V. Lloyd, Ph.D.
Human Early Learning Partnership
University of British Columbia

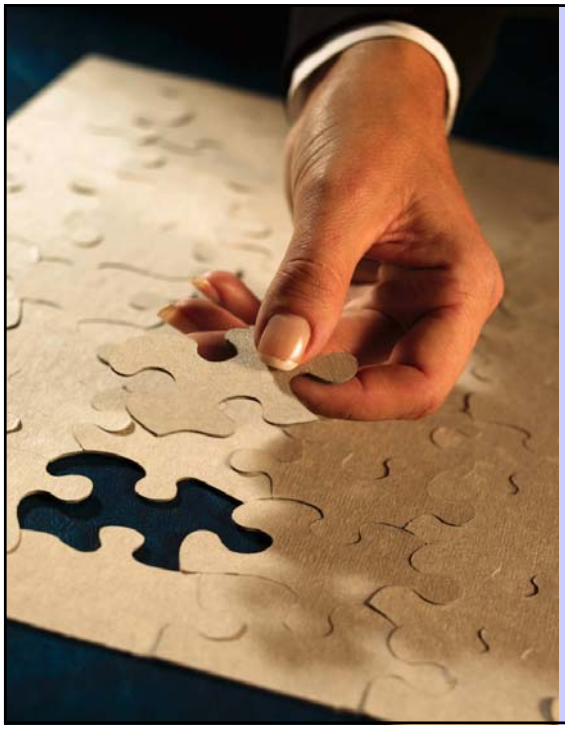
"The Languages of Assessment: It's All About Children and Families"
University of British Columbia
Thursday, May 29, 2008

Special Thanks

- Dr. Henry Braun
- Dr. Richard Carpiano
- Dr. Cody Ding
- Dr. Anita Hubley
- Alex Mann
- Dr. Jack Monpas-Huber
- Mari Pighini, Ph.D. candidate
- Dr. William Schafer
- Dr. Kim Schonert-Reichl
- Dr. Linda Siegel
- Dr. Bruno Zumbo (especially!)

- Social Sciences and Humanities Research Council (SSHRC)
- Human Early Learning Partnership (HELP)

Human Early Learning Partnership HELP



Setting the Context

One Oft Overlooked Assumption

"…one set of individuals being measured more than once on the same dependent variable"

Strict psychometric sense?
(Willett et al.)

Commensurable constructs ok?

Disciplines in which this Problem is Faced

Education:
- Achievement testing
- Cognitive development

Psychology:
- Aggression
- Sexual behaviour

Nursing:
- Dementia
- Parenting

Business:
- Expatriate research
- Organisational change

Importance of the Topic

- 'Non-solution' to the problem hardly satisfactory.
- Vast growth in use of longitudinal assessment.
- Study of change/growth necessary and important.
- Too few strategies handle the motivating problem, or are designed only for time-variable measures that can be linked *

* I'll explain more about this in a minute…

Three Scenarios

Easiest to achieve commensurability?

Research Scenario	Me…
Exact same measure acro… waves	Content, wording, response categories and response formats remain constant across waves
Linkable time-variable measures	Content, wording, response categories and/or response formats vary to some degree acr…
Non-linkable time-variable measures	C… ategories and/or response formats vary completely across waves, and there are no anchor (common) items whatsoever

Hardest to achieve commensurability?

How Can Time-Variable Measures Vary?

1. **Item stems change**
2. **Response formats change**
3. **Response categories change**
4. **Content changes**
5. **Testing context changes**
6. **Items don't function the same way over time**
   - Example: 5 x 7 = 35

Math skill? **Memory?**

Current State of Measurement Training

- **Recent decline in graduate training in statistics in North American universities' departments of psychology, and increasing trend towards "doctoral-level psychologists with little knowledge of psychometrics who nevertheless [are] engaged in psychological assessments"** (Merenda, 2003, p. 212).
- **Applied researchers are increasingly lacking advanced or expert psychometric training or knowledge** (Aiken et al., 2008).
- **Danger that the increasing availability and user-friendliness of complex statistical software packages "substitute for clear statistical thinking and model development"** (Singer, 1998, p. 350).

Human Early Learning Partnership HELP

A Foundational Issue:

Unpacking Commensurability

Unpacking "Commensurability"

- A primary assumption underlying growth modelling analyses is that one is measuring the same or commensurable construct across all waves of the study.
- This assumption necessitates that the researcher be satisfied that the *same primary dimension or latent variable is driving the respondents' responses across waves.*
- Ideally, the latent variable that drives test-takers' responses is a representation of the construct of interest.

Human Early Learning Partnership HELP

Commensurability: A "Valid" Question

- Validity is arguably the most fundamental consideration in the evaluation of measures and their resultant scores.
- It refers to "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, 1989, p. 16).

Human Early Learning Partnership HELP

Commensurability: A "Valid" Question (cont'd)

- Validity is typically understood to be a property of ***measures***, but in fact it refers to the inferences one makes from the ***scores***.
- Regrettably, many researchers fail to report the psychometric properties of their measures' scores, often because they presume incorrectly that the measures' scores will be "as valid" as they were in previous administrations (Vacha-Haase et al., 1999).

Sources of Validity Evidence

1. Collect evidence based on test content.
2. Collect evidence based on relations with other variables.
3. Collect evidence based on internal structure.

(AERA, APA, & NCME, 1999)

Evidence #1: Test Content

- **Tables of specifications are test construction or "blueprint" documents that define the specific subdomains of the construct of interest, detail the specific subdomain to which each scale item belongs, and specify the proportion of scale items devoted to a specific subdomain.**
- **Longitudinal measures may generally be concluded to measure commensurable constructs across waves primarily when there is parity in the tables of specifications across waves.**

Human Early Learning Partnership HELP

Evidence #1: Test Content (cont'd)

Numeracy Subdomains	Proportion of Items Devoted to Each Subdomain	
	Grade 4 Version	Grade 7 Version
Number concepts and operations	35-45%	35-45%
Patterns and relations	20-30%	15-25%
Shape and space	20-30%	20-30%
Statistics and probability	5-15%	10-20%
Numeracy cumulative total	100%	100%

Human Early Learning Partnership HELP

Evidence #2: Relationships with Other Variables

- Relationship of test scores to variables external to the test.
- External variables may include "measures of some criteria that the test is expected to predict, as well as relationships to other tests hypothesized to measure the same constructs, and tests measuring related or different constructs" (AERA, APA, & NCME, 1999, p. 13).

Human Early Learning Partnership HELP

Evidence #2: Relationships with Other Variables (cont'd)

- Campbell and Fiske (1959) describe the ***multitrait-multimethod matrix (MTMM),*** a means by which to assess a measure's construct validity, or the extent to which the inferences from a test's scores accurately reflect the construct.
- Two subcategories: ***convergent validity*** which refers to the degree to which concepts that should be related theoretically are interrelated in reality, and ***discriminant validity*** which refers to the degree to which concepts that should not be related theoretically are, in fact, not interrelated in reality.
- It may be the case that longitudinal measures are commensurable if the pattern of wave-specific convergent and discriminant correlations are similar across all waves of one's study (i.e., the MTMM at Wave 1 is similar to the MTMM at Wave 2, and so forth).

Human Early Learning Partnership HELP

Evidence #2: Relationships with Other Variables (cont'd)

- It may also be possible to assess the commensurability of longitudinal measures using evidence of ***predictive validity***, which indicates how accurately test data can predict scores on a relevant criterion/comparison measure at a later time (AERA, APA, & NCME, 1999).
- If scores in a given wave are deemed to be ***strongly predictive**** of scores in a later wave/later waves, then it may be the case that the measures are sufficiently comparable for inclusion in growth modelling analysis.
- Examine the correlation between the scores of a test (e.g., Wave 1) and the scores of a criterion measure (e.g., Wave 2). If there is perfect test-retest reliability, the respondents' later scores will be completely predictable from their respective earlier scores.

Evidence #2: Relationships with Other Variables (cont'd)

# of vulnerabilities on the EDI (Kindergarten)	% Failing to meet expectations on the FSA (Grade 4)	% Not passing the FSA (Grade 4)
	Numeracy	
0	7.5	12.3
1	11.8	22.2
2-3	18.7	33.8
4-5	27.5	55.6
	Reading	
0	13.6	17.8
1	26.7	33.9
2-3	29.5	43.1
4-5	48.4	68.3

 Source: Hertzman (2006), McCain, Mustard, & Shanker (2007)

Evidence #2: Relationships with Other Variables (cont'd)

- ***Caution***: There is a trend for respondents' scores to (e.g.) improve steadily with time as a result of practise, education, or simply maturation.
- Furthermore, there have been frequent criticisms about the so-called "inherent" unreliability of two-wave comparisons (Zumbo, 1999).
- Therefore, one should use this strategy judiciously and be certain that the later scores are "correlated enough" with the earlier scores such that the possible existence of confounds do not cast doubt upon the test-retest reliability (Gregory, 2006).

Human Early Learning Partnership HELP

Evidence #3: Internal Structure

- Internal structure analyses can indicate the degree to which the relationships among the items and components conform to a given construct.
- Typically, such analyses are designed to show whether or not particular items and measures function differently for different subgroups of respondents (AERA, APA, & NCME, 1999).
- The literature on ***invariance***, a term that typically refers to the degree to which measures function the same across subgroups, is often tricky to navigate. Therefore, let's unpack what is meant by measurement invariance, factorial invariance, and configural invariance.

Human Early Learning Partnership HELP

Evidence #3: Internal Structure (cont'd)

- ***Measurement invariance (MI):*** **achieved when measures function the same across subgroups, and is typically viewed as a requirement to conducting substantive cross-group comparisons (e.g., tests of group mean differences)** (Vandenberg & Lance, 2000).
- **MI holds if and only if the probability of an observed score (i.e., the score on the measure), given the true score and the group membership, is equal to the probability of that score given only the true score.**
- **If MI does not hold over two or more waves, differences in observed scores are not directly comparable** (Meade et al., 2005) **and are, hence, not appropriate for inclusion in growth modelling analyses.**

Evidence #3: Internal Structure (cont'd)

- ***Factorial invariance (FI):*** **The aforementioned definition of MI "fits nicely into the framework of factor analysis wherein a factor score (i.e., the score on the latent variable) can be seen as the proxy for a person's true score, and the items are the observed random variables"** (Wu et al., 2007, p. 3).
- **In this sense, a factor can be conceptualised as a type of latent variable.**
- **Whereas MI necessitates that the same latent variable is measured and is measured on the same metric so that cross-group factor scores are comparable, FI requires that the measurement model linking the observed indicators to the unobserved factor(s) be identical across subgroups** (Wu et al., 2007).

Evidence #3: Internal Structure (cont'd)

- ***Configural invariance (CI):*** **the minimum condition required for factorial invariance, is based on Thurstone's (1947) principle of simple structure** (Horn, McArdle, & Mason, 1983) **in which items or measures are structured such that they have non-zero factor loadings on one and only one factor.**
- **CI is achieved when there is evidence of equality of the number of select factor loadings and where the matrix of factor loadings in the different subgroups has the same pattern of zero and non-zero factor loadings (the pattern of zero and non-zero factor loadings defines the structure of the measure itself).**
- **Identical patterns across subgroups are thought to provide evidence that the measure taps the same construct across populations** (O'Sullivan, Scholderer, & Cowan, 2005).

Evidence #3: Internal Structure (cont'd)

- **Due to developments and innovations by Golembiewski, Billingsley, and Yeager (1976) and Meade, Lautenschlager, and Hecht (2005), among others, it may be possible to assess the commensurability of longitudinal measures from a general invariance perspective – a perspective that integrates aspects of MI, FI, and CI invariance.**
- **Meade et al. (2005) outline two methods for establishing invariance in longitudinal designs: Confirmatory Factor Analysis (CFA) and Item Response Theory (IRT).**

In Conclusion

The "Take Home" Message

Whether it be…

- ✓ initially identifying the construct of interest
- ✓ formulating the research design
- ✓ selecting study participants
- ✓ choosing particular longitudinal measures
- ✓ deciding on the number of waves and the timing between them
- ✓ ***assessing the commensurability of constructs over time***
- ✓ developing the appropriate statistical models or
- ✓ reporting results

… the entire empirical process of studying individual change necessitates ongoing integrated evaluative judgements on the part of the researcher about the inferences that will eventually be made from the scores (Messick, 1989).

Human Early Learning Partnership HELP

Forthcoming Publications

- **Lloyd, J.E.V., & Zumbo, B.D. (in press). The non-parametric difference score: A workable solution for analysing two-wave change when the measures themselves change across waves. *Journal of Modern Applied Statistical Methods.***

- **Lloyd, J.E.V., Zumbo, B.D., & Siegel, L.S. (in press). When measures change over time: A workable solution for analysing change and growth across multiple waves. *Journal of Educational Research & Policy Studies.***

Human Early Learning Partnership HELP

Thank You

Jennifer E.V. Lloyd, Ph.D.

jennifer.lloyd@ubc.ca
604-827-4456