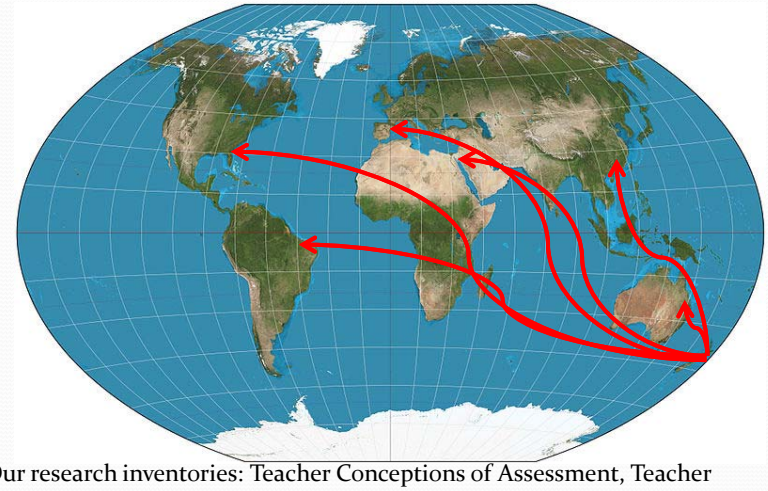
Cross-cultural research in education: Illuminating meaning by taking context into account

Gavin T. L. Brown
The University of Auckland

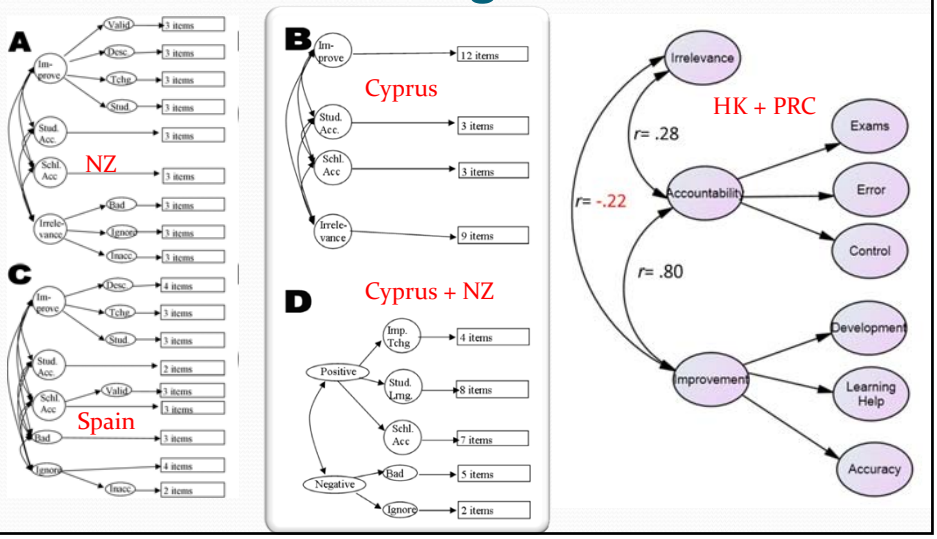
Presentation to the Department of Curriculum & Instruction, Education University of Hong Kong, August 2016

Research methods migrate



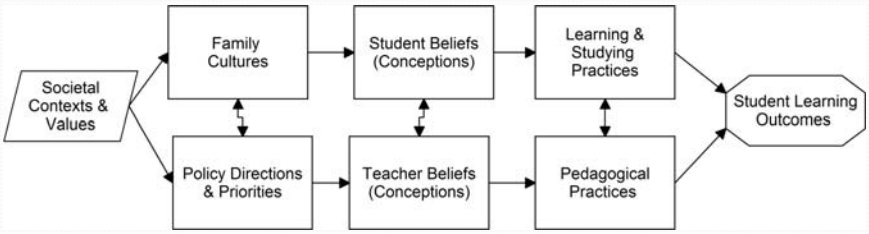
Our research inventories: Teacher Conceptions of Assessment, Teacher Conceptions of Feedback; Student Conceptions of Assessment

Teacher Conceptions of Assessment different but wrong?



Context changes results….

- Policies, cultures, histories, and societies differ



- So does a research inventory automatically work?
- Careful preparation plus statistical validation using multiple group confirmatory factor analysis invariance testing.
- Context may explain why differences exist.

Adapted for context

- Language checking
  - Translate-back translate
  - Functional equivalence
- Terminology adjusted

Validity Checks

- Following procedures outlined by Gable & Wolf (1993)
  - Do items clearly belong to the intended categories?
  - Are categories understood and accepted by samples from population of interest?
- If not, then revise categories, items, and/or language used
  - Gable, R. K., & Wolf, M. B. (1993). *Instrument development in the affective domain: Measuring attitudes and values in corporate and school settings.* (2nd ed.). Boston, MA: Kluwer Academic Publishers.

Instructions for Category Checking

- **A. Rating Tasks**
- Please indicate the category that each statement best fits by circling the appropriate numeral.
- Please indicate how strongly you feel about your placement of the statement into the category by circling the appropriate number as follows:
- 3 no question about it
- 2 strongly
- 1 not very sure
- 0 absolutely not
-
- **評分指引**
  - 請把每一項陳述歸類，並圈出最合適的類別代號。
  - 請按照所歸入的類別，指出你有多肯定該陳述是屬於該類別，
- 3 毫無疑問
- 2 相當肯定
- 1 不太肯定
- 0 肯定不是

Sample Category Selection Task

Statements of Conceptions 概念陳述	Which categories? 屬於哪一個類別？						How sure are you? 你有多肯定？			
1. My classmates and peers are better at assessments than I am. 我的同學和同輩在評估的表現上都比我好。	I	II	III	IV	V	VI	3	2	1	0
2. Assessment controls too much what and how students learn. 評估過度限制學生的學習內容和方式。	I	II	III	IV	V	VI	3	2	1	0
3. I am embarrassed to let others know how well I am doing academically. 我不好意思讓別人知道我在學習上做得有多好。	I	II	III	IV	V	VI	3	2	1	0
4. Assessment results are filed & ignored. 評估結果會被存檔而後置之不理。	I	II	III	IV	V	VI	3	2	1	0

- Use odd number of raters (3, 5, 7)
- Accept success if:
  - Majority select same category as intended with an average confidence >2.00

Sample: Category assignment

I Negative				
	Items	No. of agree	Confidence Score	Conflicting Categories
AGREE	7. When being assessed, I feel alone and abandoned. 評估時，我感到孤獨及被遺棄。	4	2.75	
	49. Assessments cause anxiety, fear, nervousness, and pressure. 評估帶來焦慮、恐懼、緊張和壓力。	4	2.5	
	26. Assessment interferes with my learning. 評估干擾我的學習。	3	2.67	
	61. Assessment is value-less. 評估沒有價值。	3	2.33	
	23. The importance of assessment is over-rated. 評估的重要性被過高估計。	3	2	
	14. Assessments cannot tell me how well I have achieved. 評估不能反映我的成就。	3	1.33	

Sample: Language equivalence

	Item	No. of Raters	Mean Score
HIGH	A1. Assessment is unfair to students. A1. 評估對學生不公平。	5	3.8
	A10. Assessment forces students to learn in a way against their beliefs. A10. 評估迫使學生用有違自己信念的方法學習。	5	3.8
	A16. Assessments cannot tell me how well I have achieved. A16. 評估不能反映我的成就。	5	3.8
	A22. Assessments cause anxiety, fear, nervousness, and pressure. A22. 評估帶來焦慮、恐懼、緊張和壓力。	5	3.8
	A23. Assessments control what and how students learn. A23. 評估控制學生學習的方式及內容。	5	3.8

....

	A7. I ignore or throw away my assessment results. A7. 我忽略我的評估結果。	3	2.6
	A18. Assessment indicates how good a student is. A18. 評估顯示學生的能力。	3	2.6
LOW	A12. Poor performance is irrelevant to me. A12. 評估表現欠佳與我無關。	2	2.8

Analysis of data

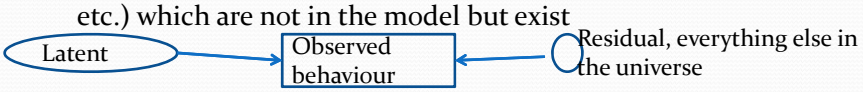
Looking for simplification

- MODEL = A theoretically informed simplification of the complexities of reality created to test or generate hypotheses

Modelling Self-report:

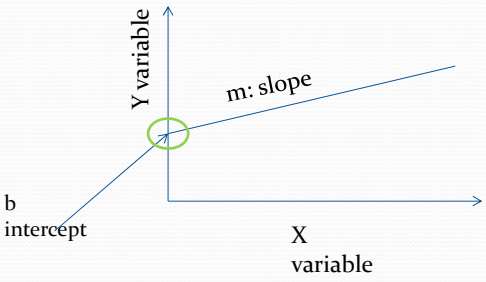
Latent trait theory

- Invisible traits explain responses & behaviours
  - But other things do too—random and systematic which we might not have data on….so these residuals influence responses
    - NB: ellipse=latent (not directly observed), rectangle=manifest
  - Example:
    - Intelligence (latent) explains how many answers (manifest) you get right on a test but there is influence from other things (e.g., breakfast, happiness, study effort, quality of teaching, etc.) which are not in the model but exist



Regression analysis

- This represents linear regressions
  - Increases in Latent (x) cause increases in Observed (y)
  - Slope is strength of association
  - Intercept is biased starting point
    - Some people tend to always start with more than others



Classic formula in high school

$y = mx + b$

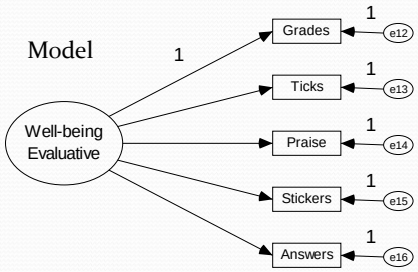
Multiple regressions

$y = b_0 + b_1X_1 + b_2X_2 \ldots\ldots$

(just keep adding an X for each new variable)

Multiple Indicators create a factor

- Latent trait explains responses
  - Responses are a sample of all possible responses
  - Everything else in the world influences responses also
- CFA are simplifications of reality of data
  - If fit well, then acceptable to work with aggregate values



Reality

	G	T	P	S	A
Grade	--				
Tick	0.75	--			
Praise	0.69	0.48	--		
Sticker	0.73	0.78	0.69	--	
Answer	0.55	0.51	0.73	0.37	--

Evaluating Model Fit

	Goodness of Fit		Badness of fit	
Decision	*p* of $\chi 2/df$	CFI gamma hat	RMSEA	SRMR*
Good	**>.05**	**>.95**	**<.05**	**<.06**
Acceptable	>.05	>.90	<.08	<.08
Marginal	>.01	.85-.89	<.10	
Reject	<.01	<.85	>.10	>.08

Note.
Report multiple indices but beware.....
CFI punishes **falsely** complex models (i.e., >3 factors)
RMSEA rewards **falsely** complex models with mis-specification

See Fan & Sivo, 2007

*AMOS only generates SRMR if NO missing data;
thus, important to clean up missing values prior to any analysis. Recommend expectation maximization (EM) procedure

Invariance testing

- Multi Group invariance testing indicates how well the same model fits 2 different groups
- If responses differ only by chance then the inventory works in the same way for both groups; they are drawn from one population
- If responses differ by more than chance than one set of factor scores cannot be used to compare groups
  - Different models and scores are needed

Testing for Invariance

- If fit indices change within chance as the equivalence constraint is imposed on the model, then that aspect of the model is invariant
  - Change in comparative fit index: ΔCFI <.01 indicates equivalence
- Equivalence is needed for
  - Configural (all paths identical)
  - Metric (all regression weights similar)
  - Scalar (all intercepts similar)
  - Each tested sequentially

Preparation: Estimation

- Maximum likelihood estimation of Pearson product moment correlations,
  - defensible for ordinal rating scales of five or more response categories (Finney & DiStefano, 2006).
  - Additional benefit: handles robustly moderate deviation from univariate normality (Curran, West, & Finch, 1996).
    - Esp. kurtosis up to 11.00
  - excessive kurtosis does not prevent analysis, it does result in reduced power to reject wrong models (Foldnes, Olsson, & Foss, 2012).

Preparation: Multivariate Normality

- Evaluated by inspection of Mardia's Mahalanobis d^2 values,
  - outliers = participants who have d^2 greater than the χ^2 cutoff for p=.001 with *df* equal to the number of variables being analysed (Ullman, 2006).
  - deletion of outlying participants should not be automatic;
    - within large samples, legitimate extreme cases will be included in the sampling frame (Osborne & Overbay, 2004).
  - evaluate model with and without the outliers to determine whether deletion makes a difference to fit quality;
    - statistically significant difference in the Akaike Information Criterion (AIC) can be used to identify superior fit (Burnham & Anderson, 2004).
  - Check after removing outliers if model still has no outliers

Study 1

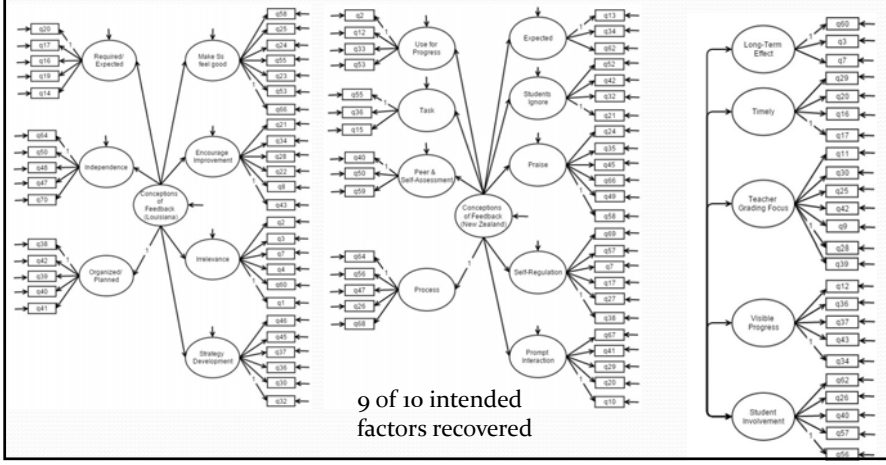
- Teacher Conceptions of Feedback self-report inventory
- New Zealand vs. Louisiana
  - Feedback purposes are feedback purposes, right?
  - But policies differ
    - Louisiana: high stakes use of assessment to evaluate schools
    - New Zealand: low stakes use of assessment to guide teaching and learning
  - So should purposes of feedback be identical?
  - If we want to compare groups, we need similar responding to the same stimuli (the TCoF)

TCoF inventory

- ***Purposes.***
  - *Irrelevance/Lacking Purpose*. (7 items) Feedback is pointless because students ignore my comments and directions.
  - *Improvement*. (6 items) Students use the feedback I give them to improve their work.
  - *Reporting and Compliance*. (7 items) I give feedback because my students and parents expect it.
  - *Encouragement*. (6 items) The point of feedback is to make students feel good about themselves.
- ***Types.***
  - *Task*. (7 items) My feedback tells students whether they have gotten the right answer or not.
  - *Process*. (9 items) My feedback focuses on the procedures underpinning tasks rather than whether the work is correct or incorrect.
  - *Self-Regulation*. (8 items) Good feedback reminds students that they already know how to check their own work.
  - *Self*. (8 items) Good feedback pays attention to student effort over accuracy.
- ***Other.***
  - *Peer and self-feedback*. (6 items) Students are able to provide accurate and useful feedback to each other and themselves.
  - *Timeliness of feedback*. (7 items) Delaying feedback helps students learn to fix things for themselves.

Models for each sample developed independently & together

Louisiana **New Zealand** **Joint Analysis**



9 of 10 intended factors recovered

Do they fit the other group?

Data Source and Model	*N*	# of items	Fit Statistics $\chi^{2\varphi}$	*df*	χ^2/df (*p*)	gamma hat	RMSEA (90%CI)	SRMR
Louisiana model								
1. 7 Hierarchical factors†	308	40	1758.12	733	2.40 (.12)	.86	.067 (.063-.072);	.080
1b. New Zealand 9 Hierarchical factors*	308	39	2048.20	694	2.95 (.09)	.81	.080 (.076-.084)	na
New Zealand model								
2. 9 Hierarchical factors	518	39	1700.44	694	2.45 (.12)	.91	.053 (.050-.056)	.062
2b. Louisiana 7 Hierarchical factors*	499	40	2587.10	733	3.53 (.06)	.84	.071 (.068-.074)	na
Joint Louisiana & New Zealand data								
3. 5 Inter-correlated factors	826	24	885.57	242	3.66 (.06)	.94	.057 (.053-.061)	.062
3b. 5 Inter-correlated factors as 2-group MGCFA*	LA=308, NZ=518	48	1254.43	484	2.59 (.11)	.96	.044 (.041-.047)	na

$^{\varphi}$= all models have $p<.001$; *=model inadmissible; na=not estimable due to model inadmissibility; †=model with statistically significant better AIC fit than paired alternative.

NO! The model from one context did not fit the other, even when a model was created using responses of both groups at the same time!!!!

How are they different?

Factors	Scale Reliability (Cronbach α) NZ	LA	Scale *M* (*SD*) NZ	LA	Effect Size Cohen's *d*	Inter-correlations I	II	III	IV	V
I. Teacher grade focus	.47	.83	2.91 (.63)	4.56 (.84)	2.31	—	.99	-.34	.77	.92
II. Visible progress	.62	.76	4.67 (.70)	4.85 (.79)	.25	*.24***	—	-.31	.75	.85
III. Student participation & involvement	.69	.76	4.03 (.81)	4.63 (.87)	.72	*.25***	*.67***	—	.19	-.42
IV. Timeliness	.61	.56	4.27 (.86)	3.86 (.99)	-.45	*.04***	*.67**	*.74***	—	.61
V. Long term effect	.15	.45	3.74 (.79)	2.82 (.86)	-1.13	*-.17***	*.75***	*.58***	*.82***	—

Inter-correlations for NZ (n=499) below diagonal in italics, for LA (n=298) above diagonal; paired comparison of inter-correlations statistical significance *$p<.05$, **$p<.01$.

What's different in Model 3?
Reliabilities, Means, and Inter-correlations

The inventory simply does not mean the same thing to both groups despite same language and shared profession as teachers

Benefit of MGCFA

- In this case, MGCFA forces the researcher to accept that teacher responses to stimuli differ in more than trivial ways across the contexts and that different models and scores are needed.
- MGCFA helps researchers avoid making serious logical errors:
  - It is highly likely that the theoretical and conceptual framework of an externally developed research tool will be invalid in a dissimilar context.
  - Reliance on scale reliabilities for each factor would have led inappropriately to acceptance of the model for the Louisiana data, while reliance on the overall fit of the joint model (Model 3) would have led falsely to acceptance of the model as appropriate for both groups.

Advances in MGCFA

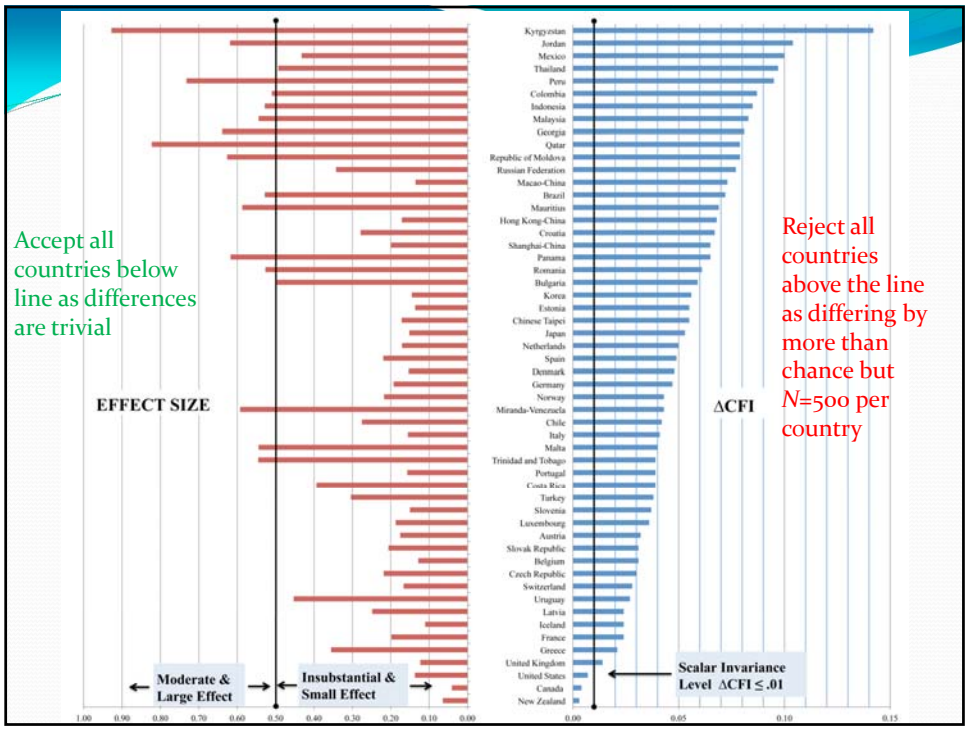
- Simultaneous examination of factor loadings and intercepts after establishing configural invariance
  - (a) item probability curves are influenced by both intercepts and slopes simultaneously,
  - (b) subsequent examination increases number of comparisons which may result in higher Type I error rates, and
  - (c) item non-invariance or non-equivalence of loadings and/or intercepts (or thresholds) is unimportant from a practical point of view.
- magnitude of measurement non-invariance effect size index (d_{MACS})
  - dMACS computer program (Nye & Drasgow, 2011).

dMACS: unidimensional

- effect size indices must be calculated separately for each latent factor.
  - Because group-level differences are integrated over the assumed normal distribution of the latent trait in the focal group (i.e., with a mean of F and a variance of F), the distributions will not necessarily be the same for different dimensions.
- Thus, the parameters used to estimate the effect size will not be the same for each latent factor, and effect sizes must be estimated separately for items loading on different factors.

Study 2: PISA Reading 2009 Booklet 11

- 28 reading literacy items = 1 factor
  - Multiple choice items were scored 0 or 1;
  - Polytomous items ranged from 0 to 2.
  - Reading processes measured were
    - Access and Retrieve (11 items),
    - Integrate and Interpret (11 items), and
    - Reflect and Evaluate (6 items).
  - Reading literacy used various text formats & types
- *N* = 32,704 from 55 countries
- Pairwise comparison: Australia vs. 54 countries



Hence

- Do NOT rely on previously published values and studies
  - Configural invariance and robust alpha values are not enough
- MGCFA needs to be run to establish if inventories or tests elicit similar admissible and similar responding
- But lack of invariance may not be fatal in and of itself
  - Check dMACS
- The whole point is to determine if comparisons can be made before proceeding to substantive discussion of results
- Consider developing instruments that have ecological validity for their own environment, rather than importing inventories or tests from other contexts.

Further Reading

- Study 1 available:
  - Brown, G. T. L., Harris, L. R., O'Quin, C. R., & Lane, K. (2015). Using multi-group confirmatory factor analysis to evaluate cross-cultural research: Identifying and understanding non-invariance. *International Journal of Research and Method in Education*. Advance online publication. doi: 10.1080/1743727X.2015.1070823
- Study 2 available:
  - Asil, M., & Brown, G. T. L. (2016). Comparing OECD PISA reading in English to other languages: Identifying potential sources of non-invariance. *International Journal of Testing, 16*(1), 71-93. doi: 10.1080/15305058.2015.1064431