Using Multi-group Confirmatory Factor Analysis to Evaluate Cross-Cultural Research: Identifying and Understanding Non-Invariance

Gavin T L Brown
The University of Auckland
Lois R. Harris
Central Queensland University
Chrissie O'Quin
O'Quin Consulting, LLC
Kenneth E. Lane
Southeastern Louisiana University

Correspondence should be addressed to Associate Professor Gavin T L Brown, School of Learning, Development & Professional Practice, Faculty of Education, The University of Auckland, Private Bag 92019, Auckland, NEW ZEALAND or by email to gt.brown@auckland.ac.nz.

Acknowledgements: The authors would like to acknowledge support from The University of Auckland Faculty Research Development Fund grant (#3609273/9215) to the Measuring Teachers' Assessment Practices project and, especially, thank Dr. Jennifer Harnett for her help with the collection of the New Zealand data. An earlier version of this paper was presented at the 2011 annual meeting of the American Educational Research Association in New Orleans, LA.

Citation: 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.*

Abstract

Multi-group group confirmatory factor analysis (MGCFA) allows researchers to determine whether a research inventory elicits similar response patterns across samples. If statistical equivalence in responding is found, then scale score comparisons become possible and samples can be said to be from the same population. This paper illustrates the use of MGCFA by examining survey results relating to practicing teachers' conceptions of feedback in two very different jurisdictions (Louisiana, USA, *n*=308; New Zealand, *n*=518), highlighting challenges which can occur when conducting this kind of cross-cultural research. As the two contexts had very different policies and practices around educational assessment, it was considered possible that a common research inventory may elicit non-equivalent responding, leading to non-invariance. Independent models for each group and a joint model for all participants were tested for invariance using MGCFA and all were inadmissible for one of the two groups. Inspection of joint model differences in item loadings, scale reliabilities, scale inter-correlations established the extent of non-invariance. This paper discusses the implications of non-invariance within this particular study and identifies difficulties in using an inventory in cross-cultural settings. It also provides suggestions about how to increase the likelihood that a common factor structure can be recovered.

USING MULTI-GROUP CONFIRMATORY FACTOR ANALYSIS TO EVALUATE CROSS-CULTURAL RESEARCH: IDENTIFYING AND UNDERSTANDING NON-INVARIANCE

While calls for increased cross-cultural and international research are not new (e.g., Crossley and Broadfoot, 1992), modern ICTs are now making international collaboration more feasible, opening up new ways to learn from other educational systems. One powerful analytical approach which is potentially very useful for cross-cultural research is multi-group confirmatory factor analysis (MGCFA) (Tran, 2009); MGCFA allows researchers to evaluate the responses of different groups of participants to a common self-report inventory or questionnaire. Invariance means that differences in the key statistical properties of group responses to a test or questionnaire are so small that they are attributable to chance rather than to group characteristics (Vandenberg and Lance, 2000). The statistical properties of interest, drawing on the theory that item responses are caused by latent traits and can be modelled as linear regressions (Borsboom, 2005), include the starting value or intercept and a strength of association or slope of the regression which explains the response behaviour. It is rare that intercepts and slopes will be identical between groups, but these differences may be within chance and thus pose no obstacle to further substantive analysis. If, however, groups have statistically equivalent responding to a set of items, this does not mean they have the same level of endorsement to the items making up each latent trait. Thus, invariance means that the relationship of the latent trait to the items differs only by chance for all groups and this permits comparison of mean scores between the groups.

While invariance indicates that the same statistical and theoretical processes explain group response patterns (although groups may have different attitudes or performance), lack of invariance demonstrates that the research tool (e.g., self-report questionnaire or knowledge test) triggers systematically different responding in each group, making score comparisons invalid since differing response mechanisms underlie group answers. When testing for invariance during MGCFA, researchers must establish whether the latent trait has a different strength (i.e., regression weights) in explaining how the groups respond to each item. Additionally, as rating scales range from very negative to very positive, few people have an opinion or attitude that starts at zero on that scale; the starting point is called the intercept. Hence, researchers need to determine if the groups have different intercept values when presented with an item or statement. If the regression weights and intercepts used to model responses are statistically equivalent between groups (i.e., differences can be attributed to chance), then we can conclude that the groups have been sampled from the same overall population and compare their scores (Wu, Li, and Zumbo, 2007). For example, high school and primary school teachers in New Zealand were invariant in their responding to the *Teacher Conceptions of Assessment* inventory (Brown, 2011), meaning that although they differed in average levels of agreement with certain factors, the slopes and intercepts of the items within the factors differed only by chance, allowing their scale scores to be compared.

Because invariance is required if cross-cultural comparisons are to be made in studies using the same instrument, it is important to understand the different analytical techniques and processes which can be used to recover factor structures which may be invariant. Additionally, it is necessary to understand the diverse threats to invariance so researchers can minimise the likelihood of generating data sets where invariance will not be found. This paper first provides an overview to multi-group confirmatory factor analysis (MGCFA) procedures. It then illustrates MGCFA with a two-group, cross-cultural study of teacher responses to a *Teacher Conceptions of Feedback* (Harris and Brown, 2008) inventory, exploring techniques used to test for invariance, discussing implications when invariance cannot be established, and providing recommendations for cross-cultural researchers wanting to use research tools that have been developed in different contexts.

Confirmatory Factor Analysis

Factor analysis presumes that there are latent traits (e.g., underlying abilities or attitudes) whose existence can be inferred from inter-correlations within a pool of items with similar content (Kline, 1994). The regression from the latent trait to each item is used to indicate how much the factor explains participant responses. For example, intelligence is a latent trait used to explain the directly measured performance of students on tests (Deary, 2001); self-belief that 'I can do mathematics' explains answers to the Mathematics Self-Description Questionnaire (Marsh, Smith, and Barnes, 1983).

Factor analysis extends conventional correlational analysis (i.e., responses are inter-correlated) by presuming latent traits are causes or predictors of manifest responses to items or statements. However, factor models only provide hypothetical causal mechanisms which require corroboration in independent, experimental studies. Nonetheless, factor models which fit the data well are attempts to explain how or why groups have the opinions, attitudes, beliefs, or abilities they demonstrate.

Exploratory factor analysis solutions are weak, since items are allowed to load on all factors. Confirmatory factor analysis (CFA), in contrast, is a sophisticated causal-correlational technique to detect and evaluate the quality of a theoretically-informed model relative to the data set of responses. CFA explicitly specifies in advance the proposed paths among factors and items, normally limiting each item to only one factor and setting the loading to all other factors at zero (Byrne, 2001; Hoyle, 1995; Klem, 2000). Unlike correlational or regression analyses, CFA determines the estimates of all parameters (i.e., regressions from factors to items, the intercept of items at the factor, the covariance of factors, and the unexplained variances or residuals in the model) simultaneously, and provides statistical tests that reveal how close the model is to the data set (Klem, 2000).

In accordance with the principle of simple structure, most CFA models are designed so that items have zero factor loadings on all factors except the one for which they were designed. Within each factor, one item is fixed to a seed value of 1 to establish the 'scale' of the latent factor, while all other items are freely estimated. Usually, the residuals for all items are zero correlated with all other residuals in accordance with the assumption that unexplained variance is randomly distributed.

Not surprisingly, given the number of different parameters being estimated simultaneously (e.g., regressions, intercepts, covariances, and residuals), CFA requires large sample sizes, usually > 500 (Chou and Bentler, 1995). By chance, even samples as large as 400 will still produce improper, and thus inadmissible, solutions about 2% of the time (Boomsma and Hoogland, 2001). Improper solutions include those where the covariance matrix among correlated factors is not positive definite (usually indicating that too many factors have been specified) or ultra-Heywood cases that generate negative error variance (i.e., more than 100% of variance is explained by the regression).

Modifications to a CFA model are normally made since the theoretical specification rarely matches the data. Much CFA modeling begins with inter-correlated factors on the assumption that participant responses to the various aspects or dimensions of a phenomenon correlate with each other. To resolve inadmissible non-positive definite covariance matrix problems, factors can be merged (esp. if the inter-correlation between factors approaches unity) or a hierarchical structure, in which a common superordinate construct that explains the shared variance between factors, is introduced. When negative error variance is detected, it can be fixed to a small positive value (i.e., .005), if twice the standard error is greater than the observed value (Chen, Bollen, Paxton, Curran, and Kirby, 2001). Otherwise, the offending latent factor needs to be removed and its items regressed onto a highly correlated factor. Other modifications include removing items with weak loadings (i.e., β<.40) on their intended factors or items with strong modification indices to other factors. These strategies ensure simple structure and low residual values; thus good fit to the data is obtained.

Hence, CFA is a more powerful method than relying on approaches such as: (a) Cronbach's alpha to validate a factor or scale reliability, (b) univariate regression analysis to determine the strength of prediction between two manifest (not latent) variables, or (c) bivariate correlations to investigate relationships between manifest (not latent) variables (Raykov and Marcoulides, 2007). Another advantage of CFA is that, in addition to providing indices of fit between the model and the data, it can provide an estimate of the amount of variance explained by the paths in the model.

There is considerable debate as to appropriate indices and cutoff values for determining the quality of fit for a model to its underlying data set (e.g., special issues in *Personality and Individual Differences* 2007 v.42 or *Structural Equation Modeling* 2000, v. 7). It is important to report multiple fit indices (Fan and Sivo, 2005; Hu and Bentler, 1999) in evaluating a model since not all fit indices are stable under different model conditions. Two levels of fit are generally discussed; 'acceptable' fit can be imputed if RMSEA is < .08, SRMR is ≈ .06, gamma hat and CFI are > .90, and χ^2/df is <3.80; while, 'good' fit can be imputed when RMSEA is < .05, SRMR <.06, gamma hat and CFI are > .95, and χ^2/df is <3.00 (Cheung and Rensvold, 2002; Fan and Sivo, 2007; Hoyle and Duvall, 2004; Marsh, Hau, and Wen, 2004). If the model is deemed to fit the data, then the model need not be rejected as an accurate simplification of the data, bearing in mind that without experimental evidence, causal claims embedded in a model are hypothetical.

Multi-group Confirmatory Factor Analysis

Within cross-cultural settings, it may be inappropriate to utilize an inventory developed elsewhere because the statistical model will not automatically be admissible or equivalent in another context. It is important to validate models with an independent sample (different to the one used to develop the initial model), thus overcoming any chance artefacts due to sampling (Hoyle, 1995). In cross-cultural research, it is much more likely that an instrument inventory will not elicit equivalent responding because of differences in factors like context, culture, and language. Thus, it is essential that the equivalence of a statistical model is investigated to determine whether the responses of different populations differ by more than chance.

The invariance of a model across subgroups can be tested using a multi-group confirmatory factor analysis (MGCFA) approach with nested model comparisons (Vandenberg and Lance, 2000). Measurement invariance is accepted if the difference in model parameters between groups is so small that the difference is attributable to chance (Hoyle and Smith, 1994; Wu, Li, and Zumbo, 2007). If the model is statistically invariant between groups, then it can be argued that any differences in factor scores are attributable to characteristics of the groups rather than to any deficiencies of the statistical model or inventory. Furthermore, invariance indicates that the two groups are drawn from equivalent populations (Wu, Li, and Zumbo, 2007), making comparisons appropriate. The greater the difference in context for each population, the less likelihood participants will respond in an equivalent fashion, suggesting cross-cultural adaptations of inventories have to be carefully examined before substantive claims are made.

In order to make mean score comparisons between groups, multiple levels of equivalence have to be demonstrated. First, the pattern of fixed and free factor loadings among and between factors and items has to be the same (i.e., configural invariance) for each group (Cheung and Rensvold, 2002; Vandenberg and Lance, 2000). RMSEA <.05 is a good indicator of configural invariance (Wu, Li, and Zumbo, 2007). Second, the regression weights from factors to items should vary only by chance; equivalent regression weights (i.e., metric invariance) are indicated if the change in CFI is small (i.e., ΔCFI ≤ .01) (Cheung and Rensvold, 2002). Third, the regression intercepts of items upon factors should vary only by chance; again equivalent intercepts (i.e., scalar invariance) is indicated if ΔCFI ≤ .01. Equivalence analysis stops if each subsequent equivalence test fails or if the model is shown to be improper for either group. Strictly, configural, metric, and scalar invariance are required to indicate invariance of measurement and permit group comparisons (Vandenberg and Lance, 2000).

Hence, cross-cultural researchers seeking to examine common constructs across contexts need to demonstrate that the inventory produces similar responding prior to any analysis of scale scores, especially when an inventory is translated or adapted for use in a new linguistic environment or used within a very difficult cultural or policy context.

Using MGCFA in cross-cultural research: An example of practice

To demonstrate the complex process of invariance testing with MGCFA, our research into teacher conceptions of feedback in two diverse policy contexts (New Zealand and Louisiana, United States of America) is used to illustrate challenges that can arise when using MGCFA. We exemplify techniques used to identify an invariant solution and discuss what conclusions about the data can and cannot be drawn when samples are not invariant.

Study background

The research instrument used to illustrate MCGFA in this paper focused on teacher conceptions of feedback, specifically examining what they believed to be the reasons for providing feedback to learners. Feedback was chosen because it (a) is “among the most critical influences on student learning” (Hattie and Timperley, 2007, p. 102), (b) can increase learner satisfaction and persistence (Kluger and DeNisi, 1996) and (c) can encourage students to adopt more productive learning strategies (Vollmeyer and Rheinberg, 2005). This study involves secondary analysis of previously reported studies by testing the separately developed models on contrasting populations.

The first two authors (Harris and Brown, 2008) jointly developed an instrument primarily based on Hattie and Timperley’s (2007) model of feedback levels (i.e., task, process, self-regulation, self) and Irving, Harris, and Peterson’s (2011) four purposes of feedback (i.e., irrelevance, improvement, reporting and compliance, and encouragement). While both of these models were originally based on empirical data, it was unknown if these constructs were systematically present within teachers’ thinking, nor how teachers might endorse or relate particular types and/or purposes. Hence, this instrument was designed to establish a theoretically-informed, statistical model of how teachers understand and endorse these constructs in different educational contexts.

As argued by Brown and Harris (2009), it is presumed that teacher beliefs are strongly influenced by policy priorities within jurisdictions; these in turn influence the translation of policy into practice in the classroom and school. Policy directions also reflect the priorities of a society or culture; hence, variations in culture, society, policy, and practices would be expected to lead to systematic variation in teachers’ beliefs. MGCFA can detect similarities of responding within and across cultures; for example, statistically equivalent responding was found for primary school teachers in New Zealand and Queensland for the Teacher Conceptions of Assessment inventory (Brown, 2006).

In terms of assessment and feedback practices, the largest tensions in teacher thinking seem to exist between improvement and accountability (Harris and Brown, 2009). Generally, teachers appear to endorse practices and policies that they believe will lead to improved learning. However, teachers are generally more negative about using assessment to evaluate teachers and schools. What remains unclear is the relationship teachers see between assessment and feedback; it was considered possible that teachers, especially those in high-stakes accountability contexts, might view feedback as a distinct process that was more personal and interactive between students and their teachers and designed for learning rather than accountability purposes. It was hypothesised that participating teachers in both contexts would endorse feedback as positive for learning. However, because different assessment policy frameworks were present, it was anticipated that teachers may have differing attitudes, especially towards the role of formal evaluation within feedback.

Cultural contexts of the study

The two locations used in this study had extremely different assessment contexts, with Louisiana reflecting a very high-stakes accountability-oriented framework and New Zealand having a much lower-stakes improvement-oriented framework.

Louisiana. At the time of this study, Louisiana schools were under tremendous pressure to improve academic achievement as evaluated through external testing and accountability measures (e.g., No Child Left Behind Act of 2001); this approach is often considered assessment *of* learning, rather than assessment *for* learning (Stiggins, 2002). Louisiana has the second highest rate of poverty in the United States at 18% (Council for a Better Louisiana, 2006), along with a high rate of minority students (52% minority students in Louisiana compared with 41% nationwide) (The Louisiana Department of Education, 2006).

When this study was conducted in 2009, Louisiana schools were struggling to meet performance benchmarks, with 76% of public schools remaining below Louisiana’s 10-year performance goal set for the end of the 2009-10 school year (Council for a Better Louisiana, 2009). Under policy at the time, schools not meeting targets could have their principal dismissed or transferred, the school converted to a charter school, or up-to-half of the teachers dismissed (Louisiana Educator, 2010), despite evidence that some of these changes (e.g., conversion to charter schools) have not been shown to improve student results (Lussier, 2010; Ravitch, 2013).

Hence, in all school settings, teachers and school administrators were under tremendous pressure to produce “results” identified through standardized measures. While there is evidence that these data do influence the writing of school improvement plans, the extent of genuine school improvement resulting from these plans varies considerably, as do educator perspectives about the effectiveness of using these data as feedback (Schildkamp and Visscher, 2010). While there was no existing work on the specific feedback practices of Louisiana elementary and high school teachers, it was hypothesised that informative feedback to students, designed along the lines Hattie and Timperley (2007) identified, would (a) raise performance scores on these high-stakes tests (e.g., indicating probable grade on end-of-year external examination) and (b) be valued by teachers.

New Zealand. New Zealand has a somewhat unique assessment environment, which should be conducive to strong teacher endorsement of formative feedback (Crooks, 2010). The New Zealand Ministry of Education (2010, p. 5) stated that “We have a deliberate focus on the use of professional teacher judgment underpinned by assessment *for* learning principles rather than a narrow testing regime”. The national assessment policy prior to Year 11 (students nominally 15 years old) emphasizes voluntary, school-based assessment for the purposes of raising achievement and improving instruction relative to the outcomes and objectives specified in the national curriculum (Crooks, 2010). School evaluations are carried out by the Education Review Office (ERO) which uses triennial site visits and school self-evaluations to ensure school quality. The ERO does not require that schools demonstrate effectiveness with any one assessment method, but rather allows them to select from a range of assessment methods to show schooling effectiveness.

The curriculum is child centered, non-prescriptive, holistic, and integrated, with specified outcomes and objectives across multiple levels. At the time of this study, there was no compulsory, state mandated assessment regime prior to Year 11, though since 2010 New Zealand elementary schools are expected to report student performance against National Standards in literacy and numeracy. Hence, all assessment practices in 2009 (at the time of this study) were voluntary and low stakes, making it possible for teachers to implement a range of feedback practices without concerns related to externally-mandated testing or grading. Hence, tests and examinations in New Zealand are evaluative for students (especially in the final years of schooling); whereas, standardized tests function, for schools, as improvement-oriented assessments. Research about New Zealand teacher feedback practices suggest that because of this low-stakes assessment environment and the country's assessment *for* learning agenda, teachers provide students with diverse feedback opportunities, including self- and peer-feedback (Harris and Brown, 2013; Brown, Harris, and Harnett, 2012; Cowie, 2009; Cooper and Cowie, 2010).

Data Collection

This study used two cross-sectional, self-administered self-report surveys to compare the beliefs of two samples of teachers (i.e., Louisiana and New Zealand).

Instrument. The *Teacher Conceptions of Feedback* (TCoF) inventory (Harris and Brown, 2008) was used in this study. Developed in New Zealand, the TCoF focused on teacher perceptions of why feedback is given (i.e., purposes), the format of feedback (i.e., type), and adaptive aspects of feedback (i.e., student involvement and timeliness). The TCoF had 71 items related to ten feedback constructs; a more detailed theoretical rationale is available in Brown, Harris, and Harnett (2012). The first four factors related to Irving, Harris, and Peterson's (2011) four purposes of feedback (i.e., irrelevance, improvement, reporting and compliance, and encouragement). The next four factors were related to Hattie and Timperley's (2007) four feedback types (i.e., task, process, self-regulation, and self). The final two factors were based on adaptive aspects, including self and peer feedback and feedback's timing. Sample items for each factor include:

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.

Respondents used a six-point, positively-packed agreement rating scale known to generate discrimination in contexts of social desirability (Brown, 2004). Responses were coded: strongly disagree=1, mostly disagree=2, slightly agree=3, moderately agree=4, mostly agree=5, and strongly agree=6.

Louisiana Participants and Data Collection. The study was carried out in one urban school district (n=42,742 students in 88 schools) with a large number of students at-risk of academic failure. 76% of students were classed as 'in poverty', 11% were disabled according to Exceptional Students Service Guidelines, and 89% were of minority ethnicity, with 62 different ethnic groups represented. The district's average test scores indicated unsatisfactory achievement levels, with 15 schools being in danger of state takeover because of unsatisfactory school performance scores across multiple years. Two schools were in danger of United States Department of Education disciplinary sanctions because of unsatisfactory subgroup scores (Louisiana Department of Education, 2008). Four schools had already been taken over by the state, two of which were middle schools.

The questionnaire was distributed to 818 middle school teachers (i.e., teachers instructing students in grades 6-8), leading to 308 completed surveys (38% response rate). Of those who responded, 78% self-reported as regular middle school teachers, while 13% were academic alternative middle school teachers. The balance described their role as either a discipline alternative teacher (i.e., supervisor of students serving in-school suspensions as an alternative to out-of-school suspension) or administrative staff. 80% of participants were female; nearly half (47%) were Caucasian, 40% were African American, 5% Asian, 2% Hispanic, and 5% Other.

The survey questionnaires were administered on-line using a commercial electronic survey collection tool. This format guaranteed respondent anonymity and was viable thanks to the ubiquity of computer usage within these schools. At the request of the district Assistant Superintendent, principals directed teachers to participate in the survey as a part of that month's teacher in-service held at individual schools. Teachers had one week to complete the survey. Reminders were sent out at four different times during the week.

New Zealand Participants and Data Collection. This study surveyed a representative national sample of New Zealand teachers. New Zealand is a multicultural and multiracial society; the 2013 census showed that 64% of the population identified as New Zealand European/Pākeha[1], 14% as indigenous Māori people of New Zealand, 7% as people from various Pacific Island nations (Pasifika), and 11% as Asian ethnicities. In 2006, median annual personal income for those 15 and over were highest for New Zealand Europeans and others ($31,200), with relatively lower values for ethnic minority groups (i.e., Māori $20,900; Pasifika $20,500; Asian $14,500). There are approximately 760,000 children and 52,000 teachers in just over 2,500 schools covering elementary and high school education (Dench, 2010). Academic achievement is relatively high; only two of the 30 OECD countries outperformed New Zealand 15 year olds on the 2006 PISA reading literacy tests and 3 of 30 countries on PISA mathematical literacy tests (Dench, 2010). Almost half (44%) of students earn university entrance after 13 years of schooling, and over half of all 15-29 year olds hold a recognized qualification from higher education providers (Dench, 2010).

In total, 1492 printed teacher surveys were delivered to 457 elementary and high schools randomly selected according to a stratified representative frame using size, region, and socio-economic strata. When forms were returned blank, they were sent out again to a school with a similar stratification. School principals distributed the questionnaires to volunteer Grades 5-10 teachers who were either generalist teachers or taught the subjects of

[1] Pākeha is the indigenous Māori word for white people.

mathematics or English; questionnaires were returned to the research team in postage paid envelopes. Valid responses were received from 518 teachers, constituting a 35% return rate. Of the valid responses, 72% were female (*n*=374) and 82% were of New Zealand European ethnicity (*n*=422). These proportions are consistent with the 2004 Teacher Census (New Zealand Ministry of Education, 2005) which had 80% of respondents as New Zealand European/Pākeha; 82% of elementary and 58% of high school teachers were female. Just over three-quarters had taught for six or more years, with 56% having taught more than 10 years. Approximately half (52%) described themselves as a teacher with no additional responsibilities (e.g., department head, dean, director, manager, or subject specialist).

Independent Model Development.

This section summarises the factor analytic procedures followed in developing the two independent models (i.e., Louisiana and New Zealand), highlighting key challenges that need to be addressed when using this technique. Since the TCoF inventory had been developed on a conceptual design of ten factors, analysis began first with CFA testing in order to recover the ten factors among the participant responses. The models were modified (e.g., items dropped or paths changed) to maximize replication of the ten original factors.

All the analyses reported in the primary reports were re-analysed for this paper with AMOS v20 (IBM, 2011) using maximum likelihood estimation of Pearson product moment correlations, which is defensible for ordinal rating scales of five or more response categories (Finney and DiStefano, 2006). An additional benefit of using maximum likelihood estimation is that it handles robustly moderate deviation from univariate normality (Curran, West, and Finch, 1996). While excessive kurtosis does not prevent analysis, it does result in reduced power to reject wrong models (Foldnes, Olsson, and Foss, 2012).

Multivariate normality is evaluated by inspection of Mardia's Mahalanobis d^2 values, with outliers being 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). However, deletion of outlying participants, while permitting analysis within assumptions of the method, should not be automatic; because within large samples, legitimate extreme cases will be included in the sampling frame (Osborne and Overbay, 2004). It makes sense to evaluate a model both with and without the outliers to determine whether deletion makes a difference to fit quality; a statistically significant difference in the Akaike Information Criterion (AIC) can be used to identify superior fit (Burnham and Anderson, 2004).

Modification of each model stopped once acceptable fit quality was obtained across multiple fit indices. Throughout the process of developing a model, factor integrity, and interpretability were kept to the forefront.

Secondary Analyses: Invariance Testing of Previous Models.

While each data set had been published previously as a stand-alone study (O'Quin, 2009; Brown, Harris, and Harnett, 2012), the authors came together in 2010 to try to compare these two separate models using MGCFA in order to explore and reconcile the differences in the results.

Model Comparisons

The fit of each model, with and without outlier cases as determined by Mardia's Mahalanobis d^2 values, for each data set and for the joint analysis is provided in Table 1. Once outlier participants were removed, the subsequent analysis was checked for further multivariate outliers which exceeded the threshold, as described above; no further outliers were identified. Note, where models are inadmissible, the SRMR value cannot be computed and is shown as 'na'.

Table 1. Fit Statistics by Model and Data Source

			Fit Statistics						
Data Source and Model	*N*	# of items	χ^{2}♀	*df*	χ^2/df (*p*)	gamma hat	RMSEA (90%CI)	SRMR	AIC
Louisiana data									
1. 7 Hierarchical factors[†]	308	40	1758.12	733	2.40 (.12)	.86	.067 (.063-.072);	.080	1932.12
1a. 7 Hierarchical factors (no outliers)	298	40	1787.59	733	2.44 (.12)	.85	.070 (.066-.074)	.082	1961.59
1b. New Zealand 9 Hierarchical factors (with outliers)*	308	39	2048.20	694	2.95	.81	.080 (.076-.084)	na	2220.20
New Zealand data									
2. 9 Hierarchical factors	518	39	1700.44	694	2.45 (.12)	.91	.053 (.050-.056)	.062	1872.44
2a. 9 Hierarchical factors (no outliers)†	499	39	1670.22	694	2.41 (.12)	.91	.053 (.050-.056)	.062	1842.22
2b. Louisiana 7 Hierarchical factors (no outliers)*	499	40	2587.10	733	3.53 ()	.84	.071 (.068-.074)	na	2761.10
Joint Louisiana & New Zealand data									
3. 5 Inter-correlated factors	826	24	885.57	242	3.66 (.06)	.94	.057 (.053-.061)	.062	1001.57
3a. 5 Inter-correlated factors (no outliers) †	797	24	873.98	242	3.61	.94	.057 (.053-	.062	989.98

			Fit Statistics						
Data Source and Model	*N*	# of items	χ^{2}♀	*df*	χ^2/df (*p*)	gamma hat	RMSEA (90%CI)	SRMR	AIC
					(.06)		.061)		
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	1486.43
3c. 5 Inter-correlated factors as 2-group MGCFA (no outliers)* †	LA=298, NZ=499	48	1199.49	484	2.48 (.12)	.96	.043 (.040-.046)	na	1431.49

Note. RMSEA=root mean square error of approximation; SRMR=standardised root mean residual; AIC=Akaike Information Criterion; LA=Louisiana; NZ=New Zealand; ♀= all models have *p*<.001; *=model inadmissible; na=not estimable due to model inadmissibility; †=model with statistically significant better AIC fit than paired alternative.

Louisiana Model. The Louisiana variables were all univariate normal and multivariate normality Mahalanobis distance was exceeded by just 10 (3.25%) participants. The Louisiana CFA model (O'Quin, 2009) identified seven hierarchically-structured factors (Figure 1) in which teachers most strongly endorsed the notions that feedback encouraged improved learning, feedback helped students develop learning strategies, and feedback was organized and planned. In contrast, they gave weak agreement to the notion that students were expected to independently generate feedback and that it was irrelevant to students. Student preference for grading was part of irrelevance, while giving grades was part of professional requirements. Fit was marginal to acceptable and was better with outliers retained.

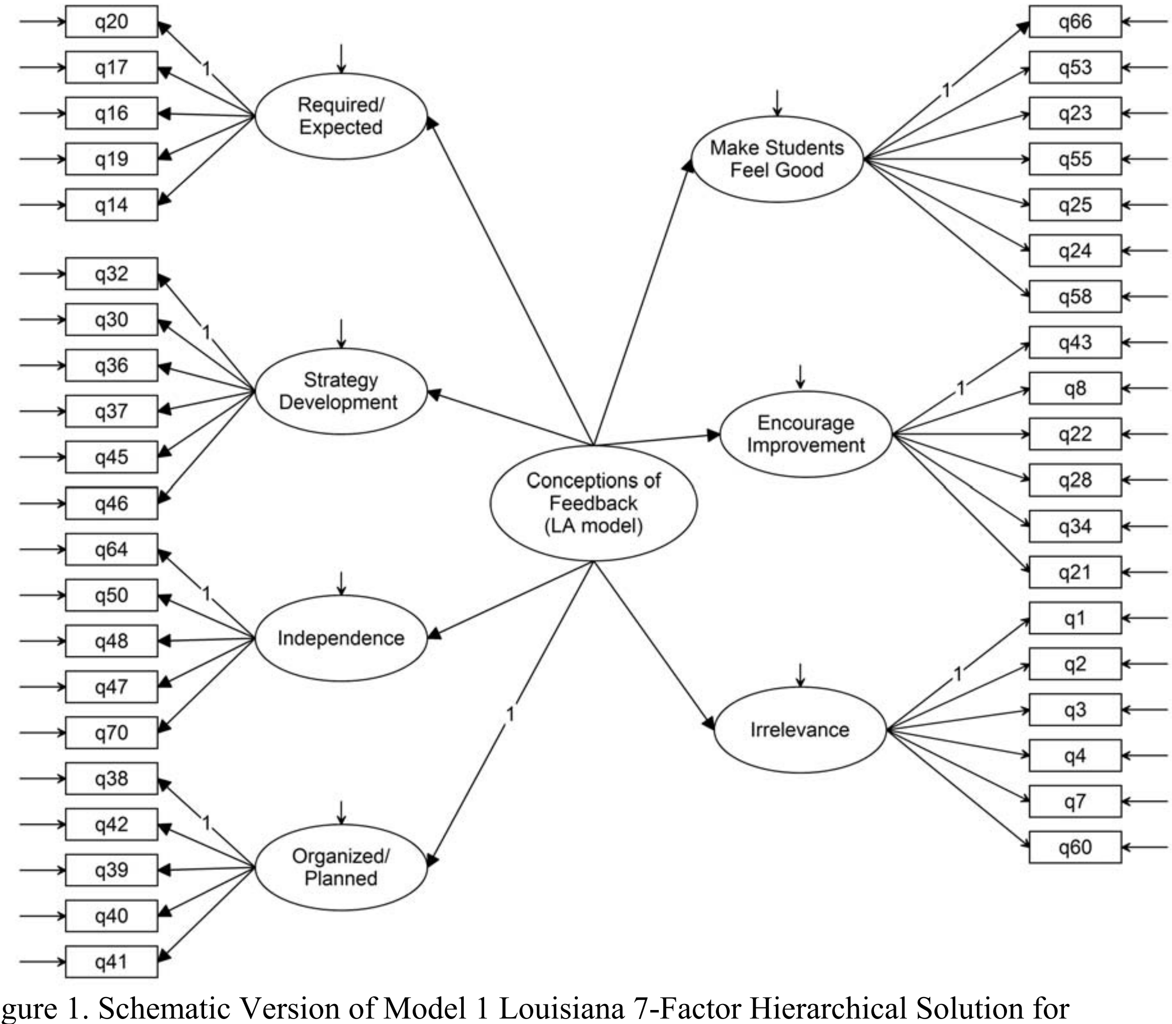

Figure 1. Schematic Version of Model 1 Louisiana 7-Factor Hierarchical Solution for Teacher Conceptions of Feedback
Note. Seed value for each factor shown as 1; error values removed for simplicity.

New Zealand Model. The New Zealand variables were all, but one, univariate normal. The outlier item (i.e., I avoid putting grades on student work as part of feedback) had 81% strongly disagree responses and was removed from the factor analysis. Multivariate normality Mahalanobis distance was exceeded by just 19 (3.67%) cases. The New Zealand CFA model (Brown, Harris, and Harnett, 2012) found nine 1st-order factors (i.e., eight of the original 10 were as designed, with Encouragement and Self items merging into a common ninth factor), all of which were predicted by a higher-order factor, thus creating a hierarchical model just as Model 1 in Louisiana (Figure 2). Fit was acceptable and was better with outlier participants removed. The model was found to be invariant between elementary and high school teachers in the sample.

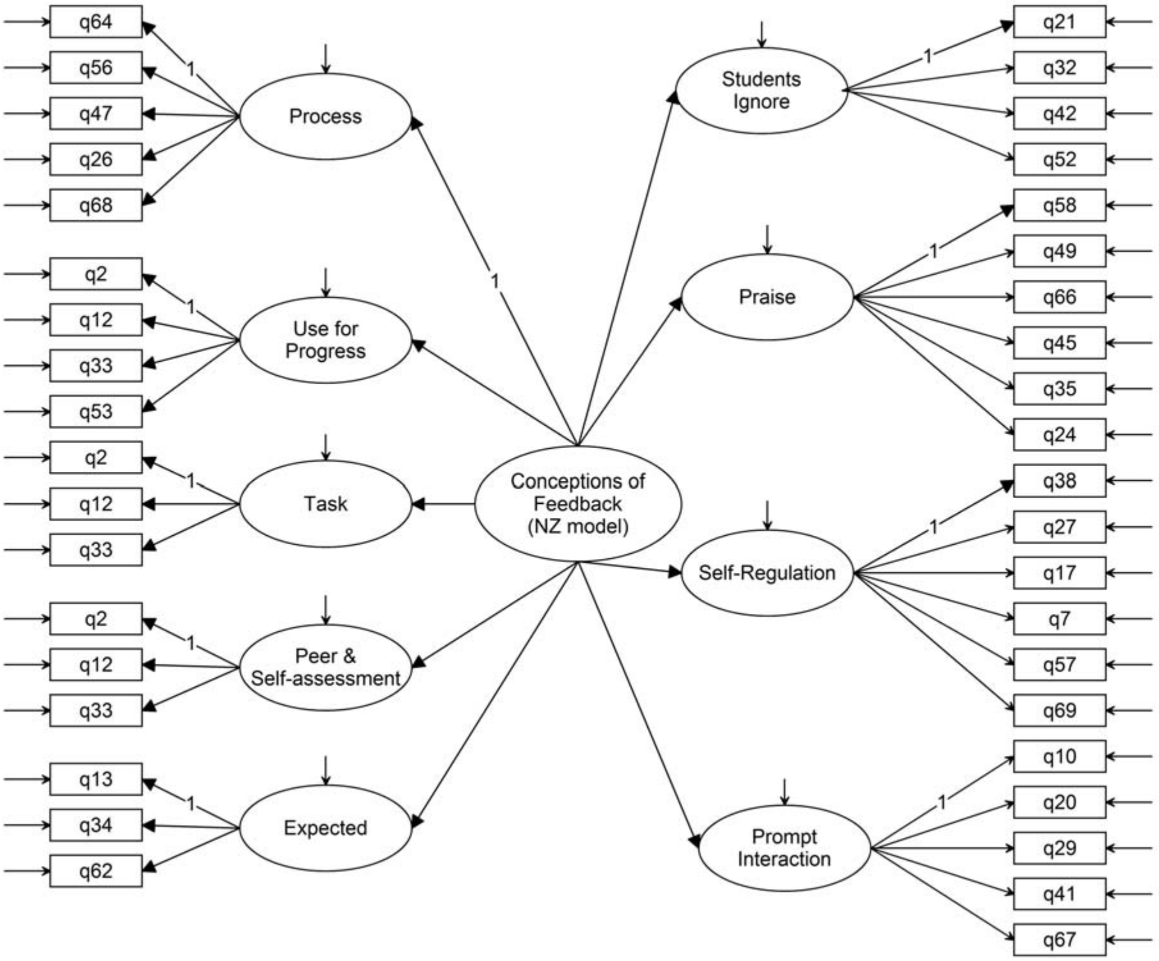

Figure 2. Schematic model of New Zealand 9 Factor, Hierarchical Solution for Teacher Conceptions of Feedback
Note. Seed value for each factor shown as 1; error values removed for simplicity.

Multigroup confirmatory factor analyses. While the same inventory had been used in both locations, quite different structural relations and factors had been found, raising doubts about the plausibility of model equivalence. The Louisiana model of seven factors (Model 2b in Table 1) was inadmissible for the New Zealand sample because of negative error variance on one factor and fit was in the reject range. Likewise, the New Zealand model of nine hierarchical factors (Model 1b in Table 1) was inadmissible for Louisiana teachers because of negative error variance on one factor and fit was unacceptable. Hence, neither of the two independently created analytic models worked for the other group, despite both having similar hierarchical structures.

Joint New Zealand-Louisiana Model

Because both of the theoretically derived, yet independent, models were not admissible for the other group, it was decided to use a purely exploratory and statistical approach to determining if a common model of responses to the TCoF could be identified. Discovering a common model might suggest that differences between groups were artificial results arising from treating the groups as independent. The different models prevent direct comparison between groups, an important goal for most researchers. Post facto separation of participants into their two groups after finding a common model would provide corroboration that the model was acceptable for each group. Inspection of fit indices (e.g., AIC) would determine if the joint model had better fit to the data than the independent sample models.

The recommended approach is to develop models separately for each group and then test each independent model with the other group; which is what testing of Models 1 and 2 on the alternate group achieved. However, it is not uncommon for a purely empirical and data-driven approach, such as developing a common joint model, to be taken when researchers seek to discover a basis for comparison. It could be, despite the lack of a theoretical basis for a common model, that this exploratory approach might discover a solution otherwise masked by treating participants differently according to their groups.

A further reason in this paper to develop the joint model was to illustrate the potentially illusory result of ignoring group differences. EFA followed by CFA is commonplace in educational research and this approach, when applied to demographically different groups, may produce spurious comparisons not warranted by in-depth investigation of model equivalence across contributing groups. Hence, especially for the illustration of this latter pitfall, we attempted a common factor analysis across all teachers while ignoring their group membership.

After merging the New Zealand and Louisiana data into a single data set, minus the Mahalanobis distance outliers, exploratory factor analysis suggested five factors which were confirmed in a restricted CFA and, after modification procedures, acceptable fit was found (Model 3a in Table 1 and Figure 3). Again, two cautions about this process are worth repeating; (a) the CFA is being run on the same data from which the exploratory result was generated and (b) further modifications were still required post-EFA to achieve adequate fit. While these processes are commonplace, the resulting models should be treated as exploratory rather than robust proof of the theoretical or conceptual model.

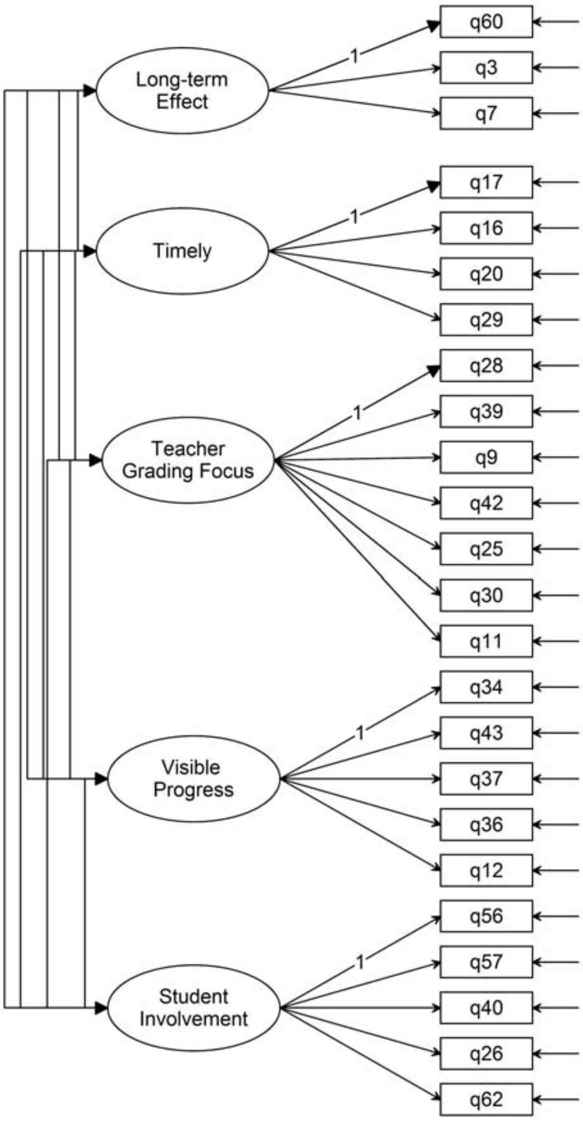

Figure 3. Schematic Model of Joint New Zealand and Louisiana 5 Inter-correlated Factor Solution for Teacher Conceptions of Feedback
Note. Seed value for each factor shown as 1; error values removed for simplicity.

Table 2 provides the factors, items, and standardized loadings for the joint solution, as well as loadings for the Louisiana and New Zealand data. Note from Table 2 that about half of the items in Model 3 are identical to items in Model 1 Louisiana or Model 2 New Zealand, and only one-quarter of items in Model 3 were in both Models 1 and 2. This indicates clearly that the purely empirical exploratory technique provided quite a different mix of items and factors.

MGCFA of the jointly developed model (Models 3b and 3c in Table 1) was inadmissible for the Louisiana group because the covariance matrix among the five correlated factors was not positive definite. Hence, we must conclude that even this jointly developed model did not apply equally to both groups. Furthermore, the substantial increase in AIC values also supports the conclusion that the two-group solution was worse-fitting than the one-group solution.

Analyses of Possible Causes of Non-Invariance

Although the Louisiana group was inadmissible for Models 3b and 3c (Table 1), scale statistics were examined to isolate aspects of the statistical model which may cause the non-invariance (Table 3). Mean scores were determined by simple averaging of all items loading onto each factor, while inter-correlations were standardized values reported in the CFA. While this type of analysis is not warranted since the joint model was inadmissible for one group, it is used to illustrate the nature of non-invariance. In a cross-cultural study that had not been guided by MGCFA, it is possible group comparisons might be inappropriately carried out because the joint model had acceptable fit.

Table 3 shows that, notwithstanding the overall acceptable fit of Model 3b, the scale alpha reliability estimates are more robust for the Louisiana group, with none of the scales meeting the conventional .70 threshold among the New Zealand teachers. There are large discrepancies in loading strength between the two groups, especially in factors Teacher Grading Focus and Long Term Effect (Table 2), clearly showing that these items do not function in a similar way across groups. The mean scores had large differences for three factors (Cohen's [1992] *d*>.60, Hattie, 2009). Likewise, the inter-correlations among the factors differed by more than chance for the two groups.

Hence, we conclude that MGCFA confirmed that both separate and joint approaches to developing produced non-equivalent results. This limits the substantive comparisons possible between teachers in the two different contexts, but allows for speculation about the substantial differences found between these populations.

Table 2. Model 3a 5-Factor Solution Standardised Loadings for Joint, Louisiana, and New Zealand data sources

	Item location in previous study		Factor Loadings		
Factor and Items	Louisiana Model	New Zealand Model	Joint data	Louisiana data	New Zealand data
Teacher Grading Focus					
Students quit trying if feedback points out faults in their work	na	na	.77	.77	.09
Students only pay attention to the grades or scores I give them	Irrelevance	na	.71	.75	.33
I seldom give written feedback because students throw it away	na	Irrelevance	.70	.51	-.04
I use positive comments to soften the blow of poor results	Feel Good Reward	na	.67	.68	.29
Teacher feedback is far more accurate than feedback from a student's peers	na	na	.64	.64	.69
Teachers are the most reliable source of feedback	na	na	.61	.71	.57
I tell my students whether their work is good or bad	na	na	.53	.52	.19
Visible Progress					
I can see progress in student work after I give feedback to students	na	Improvement	.67	.71	.61
My feedback reminds students of error correction strategies so they can fix their own mistakes	Help Learn	Self-Regulation	.67	.75	.61
My feedback is specific and tells students what to change their work	Help Learn	Task	.62	.75	.53
At my school, teachers are expected to give both spoken and written feedback to students	Professional Requirement	Accountability	.41	.56	.30
I aim to raise student performance with my detailed comments	na	na	.39	.37	.47
Student Participation & Involvement					
My students generate ideas about improving their learning independent of me	Independent	Self-Regulation	.74	.77	.66
Feedback helps students construct their own ideas about how	Help Learn	na	.66	.78	.64

	Item location in previous study		Factor Loadings		
Factor and Items	Louisiana Model	New Zealand Model	Joint data	Louisiana data	New Zealand data
to improve					
In feedback, I describe student work to stimulate discussion about how it could improve	Organised & Planned	Process	.63	.59	.64
Students are able to provide accurate and useful feedback to each other and themselves	na	PASA	.62	.62	.61
Feedback practices at my school are monitored by school leaders	Professional Requirement	Accountability	.44	.42	.33
Timeliness					
My feedback focuses on the procedures underpinning tasks rather than whether the work is correct or incorrect	na	na	.63	.64	.62
Feedback is about helping students evaluate their own work	na	Self-Regulation	.58	.46	.69
Students should not have to wait for feedback	na	Timeliness	.47	.48	.44
Feedback that takes more than a week to get to the student is useless	na	Timeliness	.40	.41	.40
Long Term Effect					
I encourage students to correct/revise their own work without my prompting	na	na	.64	.31	.55
I give feedback because my students and parents expect it	Professional Requirement	na	.44	.61	.13
Students use feedback even if they get it long after they have completed the task	Independent	na	.16	.51	.01

Note. na=not applicable in source model.

Table 3. Scale Statistics and Inter-Correlations for Teacher Conceptions of Feedback Inventory Joint Five-Factor Model 3b (no outliers) for Louisiana and New Zealand Teachers

	Scale Reliability (Cronbach α)		Scale *M* (*SD*)		Effect Size	Inter-correlations				
Factors	NZ	LA	NZ	LA	Cohen's *d*	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***	—

Note. NZ=New Zealand; LA=Louisiana; Cohen's *d* effect sizes are positive if LA>NZ; 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.

Discussion

This paper has explored data from a cross-cultural study using MGCFA to identify and examine challenges occurring when conducting such research, especially when invariance is not achieved. Despite a fundamentally null-result due to non-invariance, this study allows us to consider the importance of using the MGCFA approach. These results showed that (a) the TCoF items did not group in a similar way, (b) the items elicited very different levels of agreement, and (c) the factors had very different relationships to each other across cultural contexts, allowing the conclusion that the absence of invariance was not a product of statistical modeling. The MGCFA approach forces the substantive researcher to accept that teacher perceptions in this example differ in more than trivial ways across the contexts and that differing statistical models are necessary to explain participant responses.

This study illustrates why researchers must check that the statistical models of research tools actually apply to new sets of participants to 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. In this example, 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. Researchers are encouraged to use advanced techniques like MGCFA to validate scales and inventories, rather than simply rely on previously published values and studies. Additionally, researchers should consider developing instruments that have ecological validity for their own environment, rather than importing inventories or tests from other contexts. However, if existing inventories have been psychometrically examined and validated in their local context, adoption may be warranted, though testing of the instrument with their own data is still required.

Conventional advice on conducting CFA applies to this study and other cross-cultural analyses. Large samples (i.e., >400) are needed to ensure admissibility of models and accurate estimation of errors (Boomsma and Hoogland, 2001). It is entirely possible that the positive not definite covariance matrix which caused Models 2 and 3 to be inadmissible for the Louisiana group was an artefact of the relatively small sample size; larger samples are known to provide smaller standard errors and greater power to detect admissible solutions. Testing multiple plausible rival models, as was done with Models 1 to 3, also provides stronger evidence that a selected model is defensible (Thompson, 2000).

It is also important for researchers searching for a common model using two distinct data sets to be cautious about possible negative consequences arising from this pursuit; important constructs may be lost along the way. For example, the common model (Model 3) generated in this study only had 5 factors, whereas, the models developed in the two original contexts were more detailed (i.e., 9 factors in the New Zealand study and 7 factors in the Louisiana study). This illustrates that when trying to generate common models to bring together multiple data sets, researchers must pay close attention to what constructs may have been excluded to improve fit; just because a particular model has good fit statistics does not mean that it robustly illustrates (a) the data from all subgroups or (b) the richness of a conceptual theorisation about a construct. Comparison to previous research is greatly weakened if the statistical models become highly divergent through model modification. If a previous model is well-attested, such an approach should not be engaged in lightly.

It is useful here to consider what reliance on these statistical techniques does to the substantive nature of inquiry. While CFA relies on a theoretical framework, the pursuit of adequate fit often results in simplification of the complexity of participant responding (e.g., items drop from a factor or factors cannot be retained). This may suggest that the expert conceptualisation of researchers is not mirrored in the lay, implicit beliefs of participants. It may also mean that sampling of a broad population, rather than a narrow sample from within the population, reduces the probability of recovering participant responding. Furthermore, the purely empirical and statistical approach within exploratory factor analysis is even more likely to not reflect researchers' sophisticated conceptualisation of a phenomenon. At least the CFA approach is dependent on a theoretical framework, even if it does not always recover the complexity of expert views, as was found in the Louisiana and New Zealand studies.

Hence, researchers may find that relying only on large-scale self-report survey mechanisms would not necessarily capture the complexity of participants' thinking; qualitative and interpretive methods may be useful adjuncts to such a research endeavour. However, as survey research presupposes that the researcher's conceptualisation of a theoretical construct is equivalent to participant's own descriptions or meanings about the same construct, researchers are cautioned not to expect confirmation between survey and interview methods that have been analysed statistically and interpretively, respectively (Harris and Brown, 2010). Both qualitative and quantitative methods have their strengths and limitations and researchers have to interpret findings in light of these methodological challenges.

While this study generated three different plausible models, MGCFA confirmed that none could legitimately be applied to both groups of participants, most likely due to the many differences (e.g., cultural, political, demographic) between groups. The inability to identify a model which was actually comparable across the two groups could be related to differences in policy factors (e.g., educational evaluation and assessment policies, practices, and consequences), the nature and number of the teachers sampled, or possibly the method of administration. That a model for each group is obtainable with the same inventory suggests that the problem probably does not lie in the research method or inventory itself. Rather, the incompatible results are likely to reflect real world differences in how feedback is understood and used in these two jurisdictions. This is especially evident in how the teachers in the high-stakes school accountability environment (LA) conceived of grading in quite a different way to teachers in the formative assessment environment (NZ). The large differences in mean scores, inter-correlations, and model structures simply point to the two samples being drawn from two completely different populations, notwithstanding the shared characteristic of being practicing school teachers. Hence, further substantive theorisation is required to develop a more comprehensive understanding of teacher conceptions of feedback that can transcend cultural and policy boundaries. At this stage, it would appear that such models need to explicitly include policy pressures and priorities in the formulation of how teachers understand and use feedback.

However, while differing policy contexts may help explain differences in teacher thinking about feedback, it is not possible to rule out that the non-invariance was a result of the different sampling and administration methods employed. This issue reflects real world difficulties in collecting comparable cross-cultural data given researchers may need to adopt slightly different recruitment and collection methods to maximize responding within their environment. While both samples were comprised of primarily female middle grades teachers, participants were selected differently; the New Zealand sample was nationally representative, while the Louisiana sample was drawn from one district, in especially difficult circumstances vis a vis educational achievement. Furthermore, New Zealand has a relatively high achieving school population, compared to the high proportion 'at risk' within the Louisiana sample. While the grade levels being taught by respondents were overlapping (i.e., grades 6-8 for Louisiana teachers and grades 5-10 for New Zealand teachers), the New Zealand sample did contain a slightly wider range of grade levels. The New Zealand sample was also much less ethnically diverse, while both samples were predominantly female. The survey administration method was also noticeably different. Putting aside the issue of paper versus web response format which may be problematic, the issue of response bias as a consequence of the semi-compulsory approach adopted in Louisiana has to be considered. It should also be considered that, even if two independent samples are drawn from the same population, MGCFA should be conducted. Equivalence was found between primary and secondary teachers for the NZ Model 2 TCoF data (Brown, Harris, and Harnett, 2012). MGCFA would identify non-equivalence between samples if the original modeling had been deficient in some important way. Certainly, both models discussed in the current example remain tentative and require replication using independent and similar samples of participants before robust claims can be made. Nonetheless, the current study provides some divergent validation for the instrument; the models are different in accordance with the differences in context.

To eliminate the methodological confounds identified in this study, future research with a self-report inventory should (a) compare teachers working in similar types of school setting (e.g., mainstream public schools), (b) use identical media for data collection, (c) collect data under the same level of compulsion, and (d) ensure teachers work in parallel grade ranges. Taking such an approach would increase the probability of such research meeting the expectations that groups participating in a natural experiment should differ in only one important dimension (e.g., spatial distribution) (Murnane and Willett, 2011). While these guidelines may seem straight forward, given different cultural and social contexts, it can be potentially difficult to conduct research within two societies under similar protocols, but researchers should aim to keep data collection as similar as possible. In collaborations constrained by different policy protocols and with small budgets, there may be substantial differences in how studies are implemented; such procedural differences can impact invariance testing.

Despite its many challenges, cross-cultural research is clearly important for any research field. This paper demonstrates that MGCFA can help illuminate important differences between groups and prevent researchers engaging in substantive but indefensible claims. When invariance is not found, this paper has suggested additional techniques (e.g., examination of factor inter-correlations, factor to item loadings, sample sizes, evaluation of negative error variances, and restructuring of factor relationships) that might help provide clues as to the sources of these differences, leading to improved and valid future studies that will benefit education in diverse cultural contexts.

References

Boomsma, Anne, and Jeffrey J. Hoogland. 2001. "The robustness of LISREL modeling revisited." In *Structural equation modeling: Present and future*, edited by R. Cudeck, Stephen Du Toit and Dag Sorbom, 139-168. Lincolnwood, IL: Scientific Software International.

Borsboom, Denny. 2005. *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge, UK: Cambridge University Press.

Brown, G. T. L. 2004. "Measuring attitude with positively packed self-report ratings: Comparison of agreement and frequency scales." *Psychological Reports* 94 (3):1015-1024. doi: 10.2466/pr0.94.3.1015-1024.

Brown, Gavin T. L. 2006. "Teachers' conceptions of assessment: Validation of an abridged instrument." *Psychological Reports* 99 (1):166-170. doi: 10.2466/pr0.99.1.166-170.

Brown, G. T. L. 2011. "Teachers' conceptions of assessment: Comparing primary and secondary teachers in New Zealand." *Assessment Matters* 3:45-70.

Brown, G. T. L., and L. R. Harris. 2009. "Unintended consequences of using tests to improve learning: How improvement-oriented resources heighten conceptions of assessment as school accountability." *Journal of MultiDisciplinary Evaluation* 6 (12):68-91.

Brown, G. T. L., L. R. Harris, and J. Harnett. 2012. "Teacher beliefs about feedback within an Assessment for Learning environment: Endorsement of improved learning over student well-being." *Teaching and Teacher Education* 28 (7):968-978. doi: 10.1016/j.tate.2012.05.003.

Burnham, Kenneth P., and David R. Anderson. 2004. "Multimodel Inference: Understanding AIC and BIC in Model Selection." *Sociological Methods & Research* 33 (2):261-304. doi: 10.1177/0049124104268644.

Byrne, Barbara M. 2001. *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming*. Mahwah, NJ: LEA.

Chen, Feinian, Kenneth A. Bollen, Pamela Paxton, Patrick J. Curran, and James B. Kirby. 2001. "Improper solutions in structural equation models: Causes, consequences, and strategies." *Sociological Methods & Research* 29 (4):468-508.

Cheung, Gordon W., and Roger B. Rensvold. 2002. "Evaluating goodness-of-fit indexes for testing measurement invariance." *Structural Equation Modeling* 9 (2):233-255.

Choi, Chee-Cheong. 1999. "Public Examinations in Hong Kong." *Assessment in Education: Principles, Policy & Practice* 6 (3):405 - 417.

Chou, C.-P., and P. M. Bentler. 1995. "Estimates and tests in structural equation modeling." In *Structural equation modeling: Concepts, issues, and applications*, edited by R. H. Hoyle, 37-55. Thousand Oaks, CA: Sage.

Cohen, Jacob. 1992. "A power primer." *Psychological Bulletin* 112 (1):155-159.

Cooper, Beverley, and B. Cowie. 2010. "Collaborative research for assessment for learning." *Teaching & Teacher Education* 26:979-986. doi: 10.1016/j.tate.2009.10.040.

Council for a Better Louisiana. 2006. "Fighting poverty, building community." Accessed February 17, 2007 http://cabl.org/PDFs/Poverty_UPDATE_2006.pdf.

Council for a Better Louisiana. 2009. "2009 Louisiana Report Card on Major Education Indicators." Accessed July 9, 2010 http://www.cabl.org/pdfs/reportcard_09.pdf.

Cowie, B. 2009. "My teacher and my friends helped me learn: student perceptions and experiences of classroom assessment." In *Student perspectives on assessment: What students can tell us about assessment for learning*, edited by D. M. McInerney, Gavin T .L. Brown and G. A. D. Liem, 85-105. Charlotte, NC: Information Age Publishing.

Crooks, T. J. 2010. "Classroom assessment in policy context (New Zealand)." In *The international encyclopedia of education*, edited by Barry McGraw, P. Peterson and Eva L. Baker, 443-448. Oxford, UK: Elsevier.

Crossley, M., and P. Broadfoot. 1992. "Comparative and International Research in Education: scope, problems and potential." *British Educational Research Journal* 18 (2):99-112. doi: 10.1080/0141192920180201.

Curran, P. J., S. G. West, and J. F. Finch. 1996. "The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis." *Psychological Methods* 1 (1):16-29.

Deary, I. J. 2001. *Intelligence: A Very Short Introduction*. Oxford, UK: OUP.

Dench, Olivia 2010. Education Statistics of New Zealand: 2009. Wellington, NZ: Ministry of Education.

Fan, Xitao, and Stephen A. Sivo. 2005. "Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited." *Structural Equation Modeling* 12 (3):343-367.

Fan, Xitao, and Stephen A. Sivo. 2007. "Sensitivity of fit indices to model misspecification and model types." *Multivariate Behavioral Research* 42 (3):509–529.

Finney, Sara J., and Christine DiStefano. 2006. "Non-normal and categorical data in structural equation modeling." In *Structural equation modeling: A second course*, edited by Gregory R. Hancock and Ralph D. Mueller, 269-314. Greenwich, CT: Information Age Publishing.

Foldnes, Njal , Ulf Henning Olsson, and Tron Foss. 2012. "The effect of kurtosis on the power of two test statistics in covariance structure analysis." *British Journal of Mathematical and Statistical Psychology* 65:1-18. doi: 10.1111/j.2044-8317.2010.02010.x.

Harris, Lois R., and G. T. Brown. 2008. Teachers' Conceptions of Feedback Inventory (Unpublished test). Auckland, NZ: University of Auckland.

Harris, Lois R., and Gavin T. L. Brown. 2009. "The complexity of teachers' conceptions of assessment: tensions between the needs of schools and students." *Assessment in Education: Principles, Policy & Practice* 16 (3):365-381. doi: 10.1080/09695940903319745.

Harris, L. R., and G. T. L. Brown. 2010. "Mixing interview and questionnaire methods: Practical problems in aligning data." *Practical Assessment Research & Evaluation* 15 (1):Available online: http://pareonline.net/pdf/v15n1.pdf.

Harris, L. R., and G. T. L. Brown. 2013. "Opportunities and obstacles to consider when using peer- and self-assessment to improve student learning: Case studies into teachers' implementation." *Teaching and Teacher Education* 36:101-111. doi: 10.1016/j.tate.2013.07.008.

Hattie, J. 2009. *Visible learning: A synthesis of meta-analyses in education*. London: Routledge.

Hattie, John, and H Timperley. 2007. "The power of feedback." *Review of Educational Research* 77 (1):81-112.

Hoyle, R. H. 1995. "The structural equation modeling approach: Basic concepts and fundamental issues." In *Structural equation modeling: Concepts, issues, and applications*, edited by R. H. Hoyle, 1-15. Thousand Oaks, CA: Sage.

Hoyle, R. H., and Jamieson L. Duvall. 2004. "Determining the number of factors in exploratory and confirmatory factor analysis." In *The SAGE Handbook of Quantitative Methodology for Social Sciences*, edited by David Kaplan, 301-315. Thousand Oaks, CA: Sage.

Hoyle, R. H., and G. T. Smith. 1994. "Formulating clinical research hypotheses as structural equation models - a conceptual overview." *Journal of Consulting and Clinical Psychology* 62 (3):429-440.

Hu, Li-Tze, and P. M. Bentler. 1999. "Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives." *Structural Equation Modeling* 6 (1):1-55.

Amos [computer program] version 20, Build 817. Amos Development Corporation, Meadville, PA.

Irving, S., Lois Harris, and Elizabeth Peterson. 2011. "'One assessment doesn't serve all the purposes' or does it? New Zealand teachers describe assessment and feedback." *Asia Pacific Education Review* 12 (3):413-426. doi: 10.1007/s12564-011-9145-1.

Klem, Laura. 2000. "Structural equation modeling." In *Reading and Understanding More Multivariate Statistics*, edited by Laurence G. Grimm and Paul R. Yarnold, 227-260. Washington, DC: APA.

Kline, Paul. 1994. *An easy guide to factor analysis*. London: Routledge.

Kluger, A. N., and A. DeNisi. 1996. "The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory." *Psychological Bulletin* 119 (2):254-284.

Louisiana Department of Education. 2006. "2004-2005 Louisiana state education progress report." Accessed October 18, 2006. http://www.louisianaschools.net.

Louisiana Department of Education. 2008, August. "Ten schools overcome unacceptable label: School performance scores show overall improvement." Accessed August 25, 2008. http://www.louisianaschools.net/lde/comm/pressrelease.aspx?PR=1123

Educator, Louisiana. 2010, May 18. "The title of education legislation "misleading"." Accessed July 9, 2010. http://louisianaeducator.blogspot.com/2010_05_16_archive.html.

Lussier, C. 2010, July 5. "Some charters fall short." Accessed July 9, 2010. http://www.2theadvocate.com/news/97781354.html?index=1andc=y.

Marsh, H W, Kit-Tai Hau, and Zhonglin Wen. 2004. "In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings." *Structural Equation Modeling* 11 (3):320-341.

Marsh, H. W., Ian D. Smith, and Jennifer Barnes. 1983. "Multi-trait multi-method analyses of the Self-Description Questionnaire: Student-teacher agreement on multidimensional ratings of student self-concept." *American Educational Research Journal* 20 (3):333-357.

Murnane, Richard J., and John B. Willett. 2011. *Methods matter: Improving causal inference in educational and social science research*. Oxford, UK: Oxford University Press.

New Zealand Ministry of Education. 2005. "Report on the Findings of the 2004 Teacher Census." Accessed December 10, 2010. http://www.educationcounts.govt.nz/publications/schooling/teacher_census

New Zealand Ministry of Education. 2010. "Ministry of Education Position Paper: Assessment [Schooling Sector]: Ko te Wharangi Takotoranga Arunga, a te Tauhuhu o te Matauranga, te matekitenga." Accessed November 20, 2010. http://www.minedu.govt.nz/theMinistry/PublicationsAndResources/AssessmentPositionPaper.aspx.

O'Quin, Christel Rogerie. 2009. "Feedback for students: What do teachers believe?" EdD in Educational Leadership unpublished dissertation, The Consortium of Southeastern Louisiana University and University of Louisiana Lafayette.

Osborne, Jason W., and Amy Overbay. 2004. "The power of outliers (and why researchers should always check for them)." *Practical Assessment, Research & Evaluation* 9 (6):http://PAREonline.net/getvn.asp?v=9&n=6.

Ravitch, Diane. 2013. *Reign of Error: The hoax of the privatization movement and the danger to America's public schools*. New York: A. E. Knopf.

Raykov, Tenko, and George A. Marcoulides. 2007. *A first course in structural equation modeling*. New York: Psychology Press.

Schildkamp, K., and A. Visscher. 2010. "The use of performance feedback in school improvement in Louisiana." *Teaching and Teacher Education* 26:1389-1403.

Stiggins, Richard J. 2002. "Assessment crisis: The absence of assessment for learning." *Phi Delta Kappan* 83 (10):758-65.

Thompson, Bruce. 2000. "Ten commandments of structural equation modeling." In *Reading and Understanding More Multivariate Statistics*, edited by Laurence G. Grimm and Paul R. Yarnold, 261-283. Washington, DC: APA.

Tran, Thanh V. 2009. *Developing cross-cultural measurement*. Oxford, UK: Oxford University Press.

Ullman, Jodie B. 2006. "Structural Equation Modeling: Reviewing the Basics and Moving Forward." *Journal of Personality Assessment* 87 (1):35-50.

Vandenberg, Robert J., and Charles E. Lance. 2000. "A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research." *Organizational Research Methods* 3 (4):4-70.

Vollmeyer, R., and F. Rheinberg. 2005. "A surprising effect of feedback on learning." *Learning and Instruction* 15 (6):589-602.

Wu, A. D., Z. Li, and B. D. Zumbo. 2007. "Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data." *Practical Assessment, Research & Evaluation* 12 (3):Available online: http://pareonline.net/getvn.asp?v=12&n=3.