Structural Equation Modelling: Testing for the Factorial Validity, Replication and Measurement Invariance of Students' Views on Mathematics

Contributors: Emmanuel Adu-tutu Bofah & Markku S. Hannula
Pub. Date: 2014
Product: SAGE Research Methods Cases
Methods: Confirmatory factor analysis, Structural equation modelling, Measurement
Disciplines: Education, Psychology, Sociology
Access Date: November 24, 2020
Academic Level: Intermediate Undergraduate, Advanced Undergraduate, Postgraduate
Publishing Company: SAGE Publications, Ltd.
City: London
Online ISBN: 9781473950498
DOI: https://dx.doi.org/10.4135/978144627305014529518

Abstract

In this study, we provide a detailed account of processes involved in applying structural equation modelling to validate a survey instrument – the students' view of mathematics instrument – in a new cultural setting. First, we tested the factorial validity of the instruments in Ghana for 12th-grade students (N = 2034, M = 18.49, standard deviation = 1.25; 58.2% girls). Second, in the event of model misfit, we proposed and tested an alternate factorial structure. Third, we cross-validated the new structure with an independent sample from the Ghanaian data set. Fourth, we evaluated the factorial invariance across students' gender. Initial reliability estimates and confirmatory factor analysis indicated that the data set does not fit the hypothesized model (seven-factors). Subsequent exploratory factor analysis indicated a four-factor structure for the data set. The study has important implications for studies using structural equation modelling to validate survey instruments and shows the methodological challenges associated with the importation of a Western survey instrument into a different cultural environment.

Learning Outcomes

At the end of this case study, you should

- Be able to illustrate why importation of survey instruments (e.g. math self-concept) developed in one cultural setting into a new cultural setting is problematic regardless of their high reliabilities in the original settings
- Understand the process of constructs validation using structural equation modelling to propose and statistically test an alternative factorial structure in the event of model misfit
- Understand the process of cross-validation across a second independent sample
- Understand the process involved in testing the psychometric properties of a construct

Importing Western instruments to non-Western settings have been problematic because of the low validity and reliability problems associated with such items in the non-Western settings. For a review, see Van de Vijver (2000). In addition, there has been concern about bias in Western cross-cultural research. Van de Vijver argues that the bias is reflected in the methods used and the theoretical orientations adopted. This study reports on the adaptation of one such instrument, the ‘View of Mathematics’ (VOM) scale (see Hannula, Kaasila, Laine, & Pehkonen (2005) and Roesken, Hannula, & Pehkonen (2011) for more detailed studies and a historical perspective). Bofah and Hannula have argued that the backgrounds of students in Ghana differ from Finland in many respects (e.g. school types, educational resources, and disparity between and within schools). Finland is a Nordic welfare state and a member of the Organisation for Economic Co-operation and Development (OECD), whereas Ghana is a sub-Sahara African country. Finland has an excellent educational system that has achieved a remarkable result in the recent Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA). Ghana did not participate

in PISA but has been performing poorly in TIMSS (e.g. Mullis, Martin, Foy, & Arora, 2012).

In this study, we first provide a systematic description of the processes involved in construct validation with structural equation modelling (SEM) (exploratory factor analysis (EFA) and confirmatory factor analysis (CFA)) to analyse the structure of the VOM construct for 12th-grade students in Ghana. Second, we report the empirical findings of the structure of VOM in the Ghanaian context.

The Study

Four research questions that give support for construct validity and reliability were examined. First, to make any meaningful comparison with the hypothesized model, we computed the reliabilities (Cronbach's alpha (α)) of the constructs for the Ghanaian data set. Second, we validated the data set using a more robust approach with SEM, specifically using CFA. Third, EFA was used to determine the factor structure of the Ghanaian data set. Fourth, multigroup CFA was used to confirm the derived constructs with the validated sample, as well as between students' gender. For cross-validation purposes, we split the sample into two sub-samples.

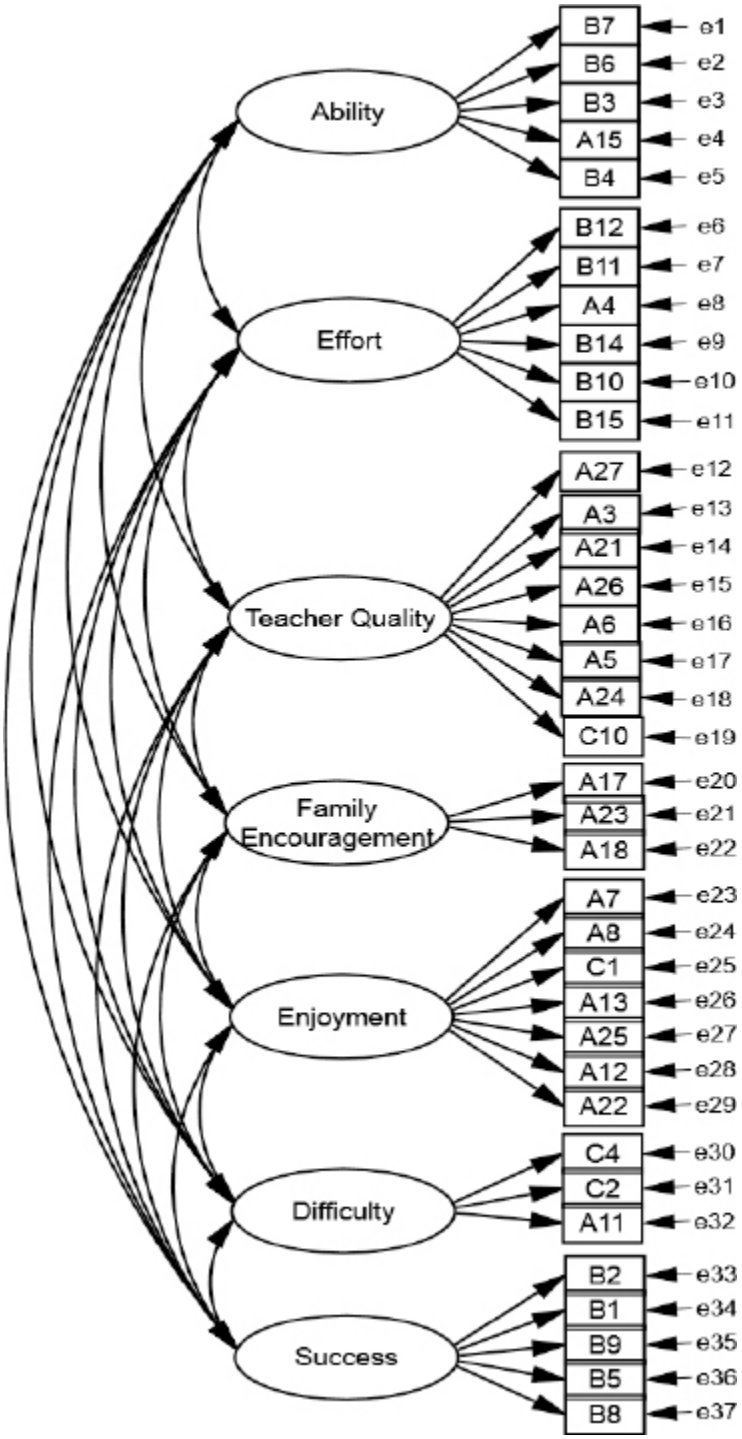
Reliability

Cronbach's α has for a long time been known to either underestimate or overestimate reliability (e.g. Geldhof, Preacher, & Zyphur, 2013; Novick & Lewis, 1967); therefore, a more robust reliability and composite reliability (ω) (see Raykov, 2012) used in conjunction with SEM will be estimated to complement the Cronbach's α estimates to the new factor structure of the Ghanaian data set. As the construct comes from a Western study, it would not be surprising to find lower reliability estimates.

Factor Structure

We hypothesize that VOM could be explained by seven factors: *ability*, *effort*, *success*, *teacher quality*, *family encouragement*, *difficulty of mathematics* and *enjoyment of mathematics* (see Figure 1). We will leave open the research question as to whether there is support for the theoretical model (seven-factor structure) identified in previous studies (e.g. Roesken et al.). However, we hypothesized a priori that (a) each item would have a non-zero loading on the VOM factor it was designed to measure but zero loadings on all the other factors, (b) the factors would be correlated and (c) all error terms associated with each item would be uncorrelated. We also expected a moderate-to-low correlation between the factors. For model identification purpose, the variances of the factors were fixed at value of one (Figure 1).

Figure 1. The seven-factor model of students' view of themselves as learners of mathematics as identified in Finland.



Measurement Invariance

In the multigroup invariance CFA, we expected support for the invariance of factor loading and factor variance and covariance (FVCV) (structural invariance) of the new proposed factor structure for the calibrated and validated sample as well for students' gender.

Methods

Most researches using VOM factors have been using principal component analyses (PCA) as the method for exploring the factor structure. The deficiencies associated with using PCA have been presented in the literature (e.g. Marsh et al., 2009; T. A. Schmitt, 2011), and thus, CFA was used to test for factorial validity. We cross-validated our new factorial structure with an independent sample from the Ghana data set.

Participants

The sample consisted of 2034 12th-grade Ghanaian students (mean age = 18.49 years, standard deviation (*SD*) = 1.25 years; 58.2% girls). Nine senior high schools were selected from urban and rural schools based on their rankings by the Ghana Education Service in their matriculation exams. The schools included single-sex, coed, private, religious and public schools. All the students were enrolled in a general mathematics class (core mathematics, 49.3%) and can opt for additional elective mathematics classes (elective mathematics, 50.7%). The students were enrolled in various academic disciplines: General Arts (33%), Business (19.2%), Science (29.1%) or Vocational Science (18.7%) courses. There were 63 different student classrooms with an average class size of 32 students.

Measures

The VOM instrument consists of 55 items. Items were assessed using a 5-point Likert scale. High reliabilities have been reported for the instrument. The Cronbach's α reliability in a study of Finnish upper secondary students was between 0.800 and 0.910 (Roesken, et al., 2011). See Hannula et al. (2005) and Hannula and Laakso (2011) for the historical development and other reported reliabilities.

Analyses

All analyses were done using Mplus 7.11 (Muthén & Muthén, 1998–2012). Analyses were based on the Mplus robust maximum likelihood estimator (MLR), with standard errors and a test of fit that were robust for non-normality and non-independence of observations (Muthén & Muthén, 1998–2012). In order to include all the observed data, missing data patterns were handled with the Mplus feature of full-information maximum likelihood (FIML). Prior to the analysis, we investigated the normality of each item (see Table 1). With guidelines of normality proposed by Curran, West, and Finch (1996), there were few non-normality items that supported the use of MLR.

Table 1. Descriptive statistics and normality indices.

Items	*M*	*SD*	Observed skewness	Observed kurtosis
A8	4.034	1.199	−1.113	0.276
A10	3.190	1.460	−0.124	−1.377
A11	2.837	1.417	0.166	−1.242
A12	3.080	1.518	−0.133	−1.426
A15	3.345	1.240	−0.383	−0.754
A19	3.589	1.387	−0.596	−0.920
A22	3.667	1.522	−0.664	−1.084
B3	3.509	1.587	−0.439	−1.454
B4	3.243	1.543	−0.127	−1.568
B8	3.440	1.292	−0.540	−0.800
B10	3.262	1.593	−0.179	−1.580
A25	3.262	1.384	−0.249	−1.156
A3	4.147	1.187	−1.312	0.670
A6	2.998	1.441	0.045	−1.352
A21	3.813	1.474	−0.830	−0.841
A24	3.327	1.501	−0.328	−1.324

For cross-validation purposes, the data set was randomly split into two, with one-half of the sample (*N* = 1017) assigned as the calibration sample and the other half (*N* = 1017) as the validation sample. Data were analysed in three stages. First, CFA procedures were conducted to investigate whether the theoretical dimension illustrated in Figure 1 fits the Ghanaian 12th-graders' sample. This stage is the confirmatory analysis part. Second, because the model did not fit the data, we proceeded by employing EFA to determine an alternative factor structure that would fit the data set. We further use confirmatory factor to confirm the new factor structure to identify item parameters that may contribute to model misfit. Third, the invariance of VOM was tested across the calibration and validation sample by (a) freely estimating the item loadings on both samples (configural invariance), (b) constraining the factor loading equal for both the calibrated and validated samples (metric invariance) and (c) examining the common characteristics of individuals by testing the invariance of FVCV in both samples (structural invariance). Similar invariance procedures were estimated for the students' gender.

Model Goodness-of-Fit

The models were compared using chi-square difference testing with the Satorra–Bentler scaled (SBS) chi-square test statistic (SBS$\Delta\chi^2$−MLRχ^2), global fit indices such as the comparative fit index (CFI), Tucker–Lewis index (TLI), root mean square error of approximation (RMSEA), Bayesian information criteria (BIC), sample-size-adjusted BIC (SSBIC), Akaike information criteria (AIC) and SBS chi-square test statistics. The chi-square test is sensitive to large sample sizes, as such, so we also interpreted invariance from a practical/approximate perspective. We examined the change in CFI and RMSEA (ΔCFI, ΔRMSEA). If the decrease in model fit for the more restrictive model is less than or equal to 0.010 for CFI, or less than 0.015 for RMSEA, then there is reasonable support for the more restrictive model (Chen, 2007; Cheung & Rensvold, 2002). For TLI and CFI values > 0.90 (with >0.95 being ideal), and RMSEA values < 0.08 (with <0.05 being ideal) are acceptable (Brown, 2006), and for AIC, BIC and SSBIC, the model with the smallest value information criterion is preferred. When evaluating the worth of individual parameters, statistical significance Mplus z-values, goodness-of-fit based on the residual values, modifications indices (MIs) and model meaningfulness were also taken into account.

Cronbach's α and composite reliability (ω) (Raykov, 2012, equation 28.18) used in conjunction with SEM were estimated. Composite reliability (ω) takes into account the computed factor loadings and produces more precise estimates of reliability than those provided by α. Composite reliability is interpreted in the same way as Cronbach's α. Generally, ω values of 0.600 to 0.700 are acceptable in exploratory research (Hair, Black, Babin, & Anderson, 2010).

Measurement Invariance

Measurement invariance is measuring the equivalence of a construct in two or more groups (e.g. gender). Measurement invariance testing begins with a baseline model usually called the *configural model* where all

parameters in the model are freely estimated across groups. If the model fits the *configural model*, then one can assume that the same variables define each factor across groups. The next model to test is the metric or weak invariance model whereby the factor loadings are constrained to be equal across groups. When there is support for both the *configural* and *metric invariance* models, one can then impose constraints on FVCV to test for structural invariance. Non-invariance structural models suggest that the associations between the underlying factors vary across groups.

Results

Several stages of analysis and their outcomes are reported in this section. First, to make a meaningful comparison with the reliabilities of the hypothesized model from previous research, Cronbach's α reliabilities were computed; second, CFA is used to test the validity of the constructs. Third, since the hypothesized model did not fit the Ghanaian data set, EFA was used to determine the factor structure of the Ghanaian data set. Finally, we cross-validated the data set using multigroup structural equation approach.

Stage 1: Computing Cronbach's α for the Hypothesized Scales

The a priori model depicted in Figure 1 stem from early research by Roesken et al., (2011) which identified seven factors (*ability*, *effort*, *teacher quality*, *family encouragement*, *enjoyment of mathematics*, *difficulty of mathematics* and *success*) for VOM. The alpha coefficients were acceptable for mathematics *ability* (0.863) and *enjoyment* (0.764), and the others below the acceptable threshold (*effort* = 0.538, *teacher quality* = 0.190, *family encouragement* = 0.623, *difficulty* = 0.565 and *success* = 0.661). To test for the unidimensionality of the constructs, we applied CFA to test the whole model that was stage 2 (CFA – stage 1; EFA – stage 2).

Stage 2: Test for Factorial Validity – Confirmatory Factor Analyses

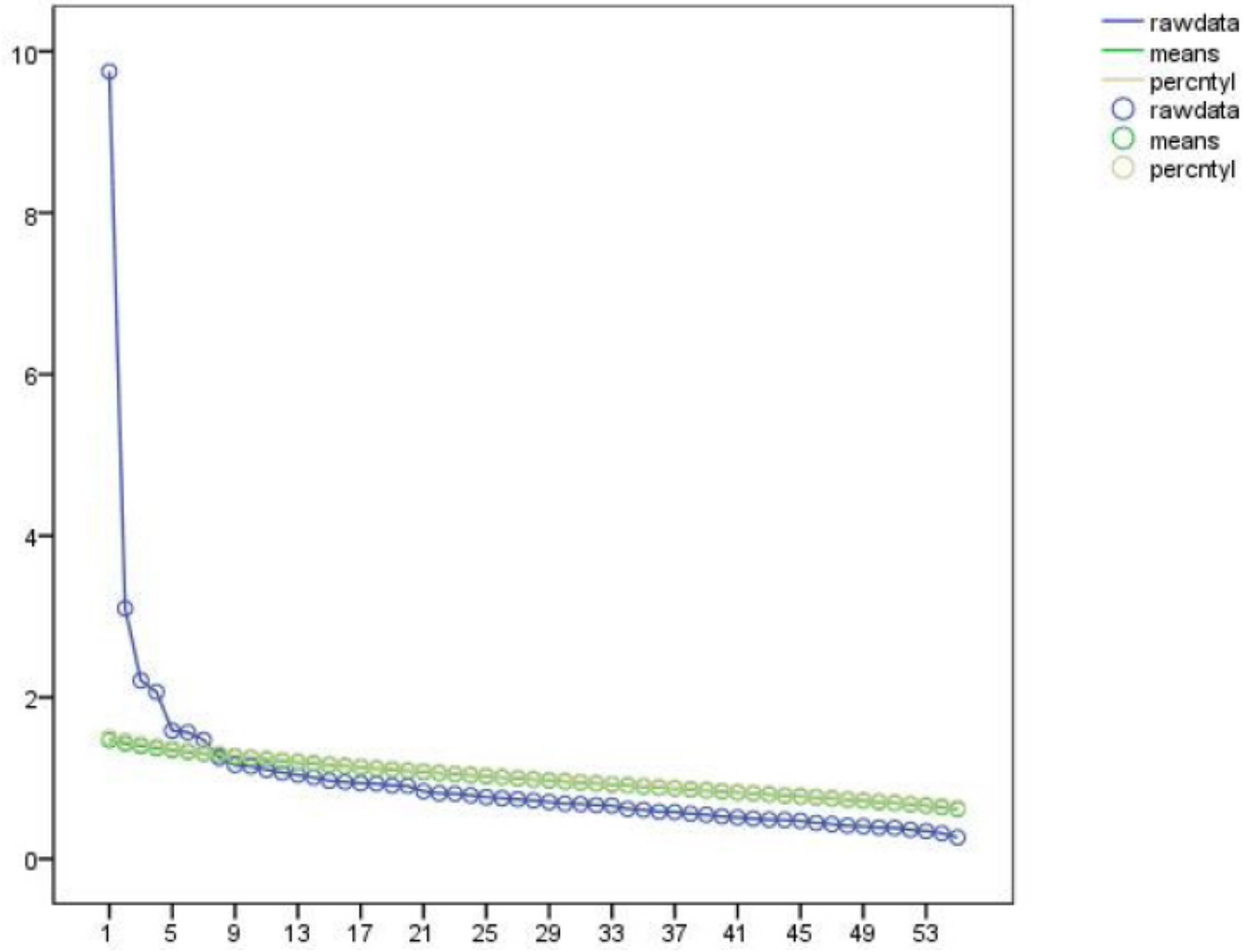
Fit indices for the seven-factor hypothesized model suggested rejecting the model (MLRχ^2(608) = 1922.993; CFI = 0.843, TLI = 0.828, RMSEA = 0.046). Correlations were very high between the *ability* and *difficulty* factors (r = 0.853), *ability* and *enjoyment* (r = 0.847) and between *difficulty* and *enjoyment* (r = 0.871), which was a possible sign of multicollinearity and suggested that the factor structures were not statistically distinguishable.

Stage 3: Exploratory Factor Analysis

After rejecting the a priori model, the data set was reanalysed using EFA to find out whether (a) the Ghanaian data set could be described by more or fewer factors, and (b) whether the same pattern of loadings would fit the validation sample. In previous research into VOM, factors correlate as such Geomin (oblique) rotation was used as the rotation procedure to get a cleaner simple factor structure that is similar to CFA (Schmitt, 2011). The results from the parallel analysis indicated a seven-factor solution (see Figure 2). Parallel analysis

is similar to the scree test, where one also plots the eigenvalues derived from a completely random set of data involving the same number of items and research participants. The point at which the eigenvalues for the actual data drop below the eigenvalues for the random data indicates the optimum number of factors (see Hayton, Allen, & Scarpello, 2004; Russell, 2002).

Figure 2. Eigenvalues from factor analysis and parallel analysis.



We estimated four-, five-, six- and seven-EFA factor solutions for the data set. We included a four-factor solution because of the high correlation that was identified early between three of the factors, presupposing those factors were measuring the same dimensions. Items with very low R^2, loadings of less than 0.300 and high cross-loadings were deleted. When the EFA was re-run, no item loaded on the seventh factor and analysis was continued with four-, five- and six-factor models. Again, items with a low R^2 and loadings of less than 0.300 or high cross-loadings were deleted and the EFA was re-run. Only 2 items loaded on the sixth factor and we ignored the sixth factor and continued the analysis with four- and five-factor solutions. There was a very high correlation between two of the constructs (*rs* = 0.849), an indication that the two constructs in the five-factor solution were statistically identical. The EFA for a four-factor structure was acceptable as the final model because it gave the most interpretable and reliable factor structure.

The Four-Factor Structure

We named the four factors *self-confidence*, *self-concept*, *family encouragement* and *teacher quality*. The a priori hypothesized model had 37 items, whereas 29[1] out of 55 items passed the threshold for inclusion in the analysis. In comparison with the original hypothesized model, the *self-concept* factor includes all 5 *ability* items, 5 out of 7 *enjoyment* items, 2 out of 3 *difficulty* items, 1 *success* item, 1 *effort* item and 2 new items (item A10: *My eagerness to study mathematics is seasonal* and item A19: *Mathematics has been a clear and precise subject to study*) making a total of 16 items. The *self-confidence* factor included 3 items from the *success* factor (B9, B2, B1) and 1 item from the *effort* factor (B15). Two items (A5, C10) failed to surpass the threshold value on the *teacher quality* factor. All items on the *family encouragement* factor surpassed the threshold for inclusion. All factor loadings were statistically significant ($p < .001$). With these findings, we can conclude that the VOM structure of the Ghanaian data set is empirically and theoretically different. Factor determinacy correlation between the estimated factor score and a factor was 0.958 for *self-concept*, 0.878 for *teacher quality*, 0.862 for *self-confidence* and 0.798 for *family encouragement*.

Post Hoc Confirmatory Factor Analysis

We used CFA to determine whether the new factor structure fits the data in comparison with the hypothesized model. We compared the models using AIC, BIC and SSBIC between models because they were not nested. To improve the VOM variates, we used the MIs to ascertain the items that best measures the four factors.

Model Comparisons

In Table 2, one can see a large improvement of fit between Model 1 (M1: the hypothesized model) and Model 2 (M2: four-factor model). A review of the MIs for ways to improve the model further showed the error covariance between items B6 and B7 had the highest MI, which were needed in the model to yield a drop in the χ^2 value. The high error covariances suggested that the two indicators covaried for reasons other than the shared influence of the latent factors.

Table 2. Model specifications for the post hoc confirmatory factor analysis.

Model	MLRχ^2	*df*	AIC	BIC	SSBIC	Description
M1	1922.993	608	112532.485	113182.534	112763.290	Seven-factor model (Roesken et al. 2011)
M2	1032.429	371	88576.779	89034.768	88739.391	Four-factor model
M3	925.446	344	85466.637	85909.852	85624.004	After deleting B6
M4	868.265	343	85401.063	85849.203	85560.179	Including error covariances of B3 and B4
M5	728.292	317	82347.127	82780.493	82500.997	After deleting B7
M6	670.710	292	79277.107	79695.699	79425.731	After deleting A7
M7	612.874	268	76220.373	76624.030	76363.591	After deleting C1 (Figure 2)

Model 3 was postulated and tested, which included the deleting of item B6. The rationale for doing this was guided by three important considerations. First, the MIs that represented the error covariance between item B6 and B7 (measure on self-*concept* factor) were higher, and indicated that including this parameter in the model would yield a significant drop in χ^2 (49.712). Second, a review of correlation matrices revealed a strong association between items B6 and B7 (r = 0.703). This information, coupled with evidence of exceptionally similar item contents suggests some item redundancy due to content overlap. This, in turn, supported the justification for the elimination of 1 of the 2 items. Item B6 was chosen for deletion based on its large residual variance value. In Table 2, the difference between Model 2 and Model 3 reflected a further improvement in the fit based on the information criteria.

By reviewing the MI again, the error covariance between items B3 and B4 indicated that if included in the model, there would be a drop in χ^2 of 58.846. The correlation between items B4 and B3 (r = 0.651) was higher, and item contents did not overlap as was the case for B6 and B7. The error covariance between

items B4 and B3 was then included in model 4. The information criteria indicated a better fit for Model 4. A further MI review indicated that the error covariance between items B7 and B3 was very high, indicating that if included in the model, χ^2 would drop by 35.680. Model 5 was postulated, which included the deletion of item B7. Furthermore, MIs indicated a higher value that represented error covariances between items A7 and A22 for Model 6, and between items C1 and A22 for Model 7. Models 6 and 7 were postulated after the deletion of item A7 for Model 6 and item C1 for Model 7.

For Models 5, 6, and 7, the rationale for deleting B7 (Model 5), A7 (Model 6) and C1 (Model 7), respectively, were guided by the MI value that represented the error covariance between the items involved (i.e. all on the self-*concept* factors). Second, a review of correlation matrices revealed a strong association between items B3 and B7 (r = 0.727), items A7 and A22 (r = 0.712) and items C1 and A22 (r = 0.723). This information, coupled with evidence of exceptionally similar item contents, suggests some item redundancy due to content overlap. These findings supported the justification for the elimination of one of the paired error covariance items. For Model 5 item, B7 was chosen for deletion, Model 6 item A7 was chosen, and for Model 7 item C1 was chosen for deletion due to their respective large residual variance values. In Table 2, there was a consistent improvement in Models 5, 6 and 7. Factor determinacy was 0.939 for *self-* concept, 0.862 for *self-confidence*, 0.878 for *teacher quality* and 0.791 for *family encouragement*.

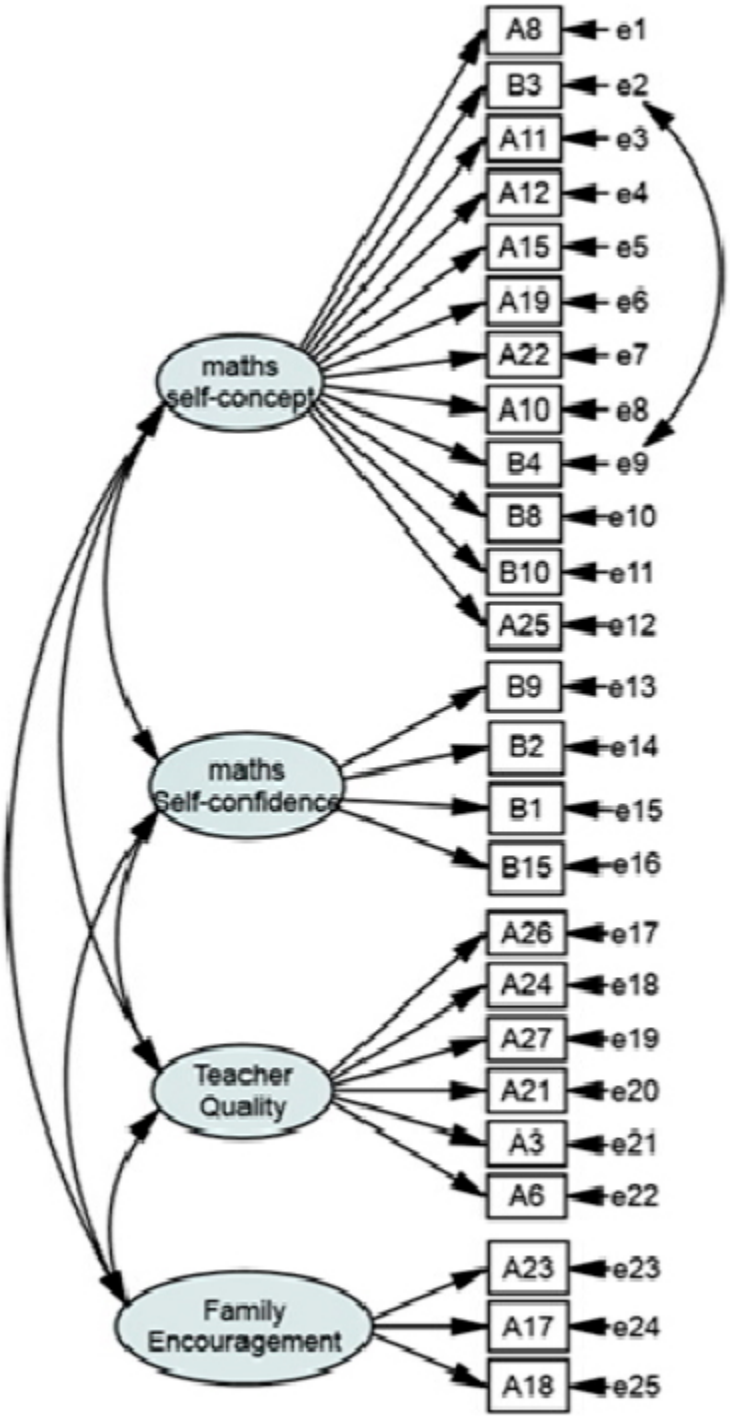
Although there were other mis-specified parameters, the strongest justification was included in the model and model parsimony guided by common sense was taken into account. In addition, the MIs were weaker than the mis-specified parameters that had been included. The remaining parameters were difficult to justify in terms of meaningfulness. The interpretation and the reason for the inclusion of the error covariance in the model have been discussed in most CFA literature (see, for example, Brown & Moore, 2012). The inclusion of the error covariance in the model is justified when these parameters represent non-random measurement errors due to *method effects*. Other possible causes have been discussed in the literature as one or a combination of the following: common assessment methods (e.g. questionnaires); reverse items, or similarly worded items, items that are presented sequentially, and items with high content overlap, items prone to differential susceptibility to other influences such as self-report items, demand characteristics, reading difficulties, item translation, acquiescence, and the format of the instrument or social desirability (Brown & Moore, 2012; Edelen & Reeve, 2007). The use of MIs has also been criticized, and others have argued that model specification that includes error covariances must be supported by substantive and/or empirical rationale (MacCallum, Roznowski, & Necowitz, 1992).

Possible Cause of the Error Covariance

Poor translation and content overlap were the main causes in all but one of the correlated errors. For the error covariance between items B3 and B4, the possible cause is that the items are presented sequentially in the survey and both measure students' weakness in mathematics (local dependency), on the same construct, and are negatively worded. The error covariance was highly significant (r = 0.444; p < .0001). To show that the inclusion of error covariance is justified and is not due to chance, the models with (MLRχ^2(268) = 612.874,

CFI = 0.932, TLI = 0.923, RMSEA = 0.036) and without (MLRχ^2(269) = 690.419, CFI = 0.916, TLI = 0.907, RMSEA = 0.039) the error covariance were compared. The fit indices and the SBS$\Delta\chi^2$–MLRχ^2 value of 61.314 and Δdf = 1 indicate that the model with the error covariance is significantly better than the model without the error covariance (the critical value for SBS$\Delta\chi^2$–MLRχ^2 is 3.84; α = 0.05, *df* = 1). Therefore, Model 7 (Figure 3) represents the VOM structure for the Ghanaian data. It is our baseline model for subsequent analyses related to cross-validation and multigroup invariance testing of the data.

Figure 3. Final hypothesized model and baseline model of the VOM constructs for the Ghanaian data set.



Stage 4: Cross-Validation Analyses

We cross-validated the new factor structure with a second data set. Cross-validation of the new factor structure was done using the multigroup approach by testing for invariance across the calibration and validation samples. The results indicated that the *configural model* (MG1) provided a good fit to the data, which indicated support for configural validity across the calibrated and validated samples (see Table 3). This good fit facilitated testing a more restrictive model. The *metric invariance* model (MG2) fit the data adequately, and it was not significantly different from the *configural model* ($SBS\Delta\chi^2 - MLR\chi^2$ = 8.917, Δdf = 21). In addition, the additional set of constraints did not lead to a meaningful drop in fit (ΔCFI, ΔRMSEA). Of substantial interest were the two specified residual covariances and the extent of their invariance across the calibration and validation samples. We considered it worthwhile for psychometric reasons and to remove any doubt of capitalizing on chance for their inclusion in the model (MacCallum et al., 1992). It was postulated that in Model MG3, the factor loadings, factor variances, factor covariances and the residual covariances were constrained to be equal. The overall goodness-of-fit indices and the ΔCFI and ΔRMSEA between Model MG3 versus Model MG2 across the sample supported structural invariance of our data. Comparison yielded a corrected χ^2 difference value that was not statistically significant ($SBS\Delta\chi^2 - MLR\chi^2$ = 9.493, Δdf = 11). Consistent with our study hypotheses, all correlations were in the low-to-modest range (r = 0.191−0.535) between the dimensions. Support for Model MG3 indicated the unidimensionality of the set of items. In addition, gender invariance (Model MG4 to MG6) was tested, and there was support for configural, metric and structural invariance, which gives further support for the validity and reliability of the constructs. Factor loadings and correlation are reported in Table 4.

Table 3. Summary of goodness-of-fit statistic for the cross-validation analysis and gender.

Model	MLRχ^2	*df*	S	RMSEA	ΔRMSEA	CFI	ΔCFI	TLI	Model comparison
Cross-validation analyses									
MG1	1239.657	536	1.168	0.036	–	0.932	–	0.923	–
MG2	1243.273	557	1.174	0.035	–0.001	0.933	0.001	0.928	M2 vs. M1
MG3	1251.145	568	1.176	0.034	–0.001	0.934	0.001	0.930	M3 vs. M2
Gender measurement invariance									
MG4	1283.281	536	1.171	0.037	–	0.927	–	0.918	–
MG5	1300.825	557	1.176	0.036	–0.001	0.927	0.000	0.922	M5 vs M4
MG6	1319.907	568	1.179	0.036	0.000	0.927	0.000	0.923	M6 vs M5

Table 4. Factor structure relating the VOM items–gender invariance.

Item	Factor loadings	*SE*	Item wording
Mathematics self-concept factor (1)			
A8	0.591	0.028	Doing calculations has been enjoyable
RB3	1.119	0.027	Mathematics is my weakest subject
RA11	0.858	0.027	Mathematics has been difficult in upper secondary school
A12	1.120	0.024	Mathematics is my favorite subject
A15	0.851	0.025	I have made it well in mathematics
A19	0.877	0.029	Mathematics has been a clear and precise subject to study
RA22	1.052	0.028	Mathematics has been the most boring part of my study
RA10	0.687	0.032	My eagerness to study mathematics is seasonal
RB4	1.058	0.027	Mathematics is difficult for me
B8	0.635	0.029	I can handle advanced mathematics tasks
RB10	0.943	0.030	I have a wrong attitude about mathematics
A25	0.518	0.033	I enjoy pondering over mathematics tasks
Mathematics self-confidence factor (2)			
B9	0.575	0.038	I know that I can do well in mathematics
B2	0.531	0.036	I can get a good grade in mathematics.

Reliability of the New View of Mathematics Scales

The Cronbach's α reliability of two of the constructs (*family encouragement*, *self-confidence*) were below the acceptable threshold. This may be because of the brevity of the constructs, as Cronbach's α is positively related to the number of items on a construct. Two of the constructs reliabilities (both Cronbach's α and composite reliability) were within the acceptable thresholds of 0.700 (self-*concept*, α = 0.872, 95% confidence interval (CI) [0.864, 0.879]; ω = 0.868, 95% CI [0.859, 0.877]; *teacher quality*: α = 0.706, 95% CI [0.684, 0.727]); ω = 0.716, 95% CI [0.696, 0.737]). The thresholds limit may decrease to 0.600 (*self-confidence*: α = 0.690, 95% CI [0.654, 0.726]; ω = 0.697, 95% CI [0.659, 0.736]; *family encouragement*: α = 0.619, 95% CI [0.552, 0.687]; ω = 621, 95% CI [0.587, 0.654]) (Hair et al., 2010).

The Cronbach's α *values* slightly underestimated the *teacher quality*, *self-confidence*, *family encouragement* constructs and overestimated the *self-concept* construct. Constructs with error covariance were overestimated and constructs without covariance were underestimated. The lower reliabilities are an indication of substantial measurement error and/or no true individual differences in the data set. This may affec**t** the validity of interpretations based on manifest scale scores, weaken statistical power, and effect sizes (Schmitt, 1996). This is an indication that statistical methods that take into account measurement errors were best for the data set.

Summary

The study illustrates the approach and methodological challenges in construct validation. We demonstrated these by using both reliability estimates and SEM. We first computed Cronbach's α to make a meaningful comparison with the reliabilities of the hypothesized model. The Cronbach's αs were far below the acceptable limit. This is in support of many other studies that indicate that the reliabilities of imported constructs to different cultural settings is problematic irrespective of higher reliabilities in the original settings. A stable four-factor model was obtained through subsequent EFA and CFA. As with Cronbach's α, it underestimated the reliabilities when there was no error term and overestimated the reliability when there was an error term in the new model. Moreover, because α is sensitive to the number of items in a scale, it underestimated the *family encouragement* and the *self-confidence* constructs.

Using SEM, we were able to detect measurement error and bias, while also understanding the disparities. We could not affirm the seven-factor hypothesized model. A possible reason is the dramatic cultural difference between these two countries. Moreover, the more robust approach could also be a possible explanation. In this study, we think the best factor structure for the Ghanaian sample was identified and a more reliable conclusion and interpretations can be made compared to the results from previous studies. Being able to validate the factor structure with an independent samples from the same data set and for students gender, we can conclude that (a) there is strong empirical support for a new four-factor structure, (b) the same variables define each factor across all sub-samples and (c) all the latent variables have the same relationship within

the sample. In addition, there was support for students in single-sex and coeducational schools (interested readers can consult our article Bofah & Hannula, submitted). This also increases its value as an assessment instrument.

The mean comparisons with any sub-sample within the data set (e.g. gender) can be interpreted as representing the underlying mean differences. Our hypothesis that the error term for the item variables was not corrected was not met, but throughout this case, we have proved the need to include the error term in the model. The outcome of this case has shown the methodological challenges and theoretical issues associated with the importation of Western instruments into non-Western settings.

Concept Review

In this case, you have met the following concepts. Explain briefly how you understood these in the context of this case:

- configural invariance
- metric invariance or factorial invariance
- structural invariance
- factor covariance invariance
- factor variance invariance
- error covariance
- PCA
- factor analysis – EFA and CFA
- calibration and validation samples
- Cronbach's α and composite reliabilities
- method effects

Note

1. In the latter section, 4 items were deleted because of content overlap detected during the confirmatory factor analysis procedure.

Exercise and Discussion Questions

1. List the various steps in validating an instrument in a new cultural setting. Comment on the rationale behind the procedure.
2. Outline the process in testing a measurement instrument for cross-group or cross-validating equivalence. What is your personal opinion about the process?
3. What are the possible consequences of measurement error?

4. Discuss the possible problems that could emerge from item content overlap, local dependency and oppositely worded items in survey instruments.

Further Reading

Bofah, E. A., & **Hannula, M.** (in press). Studying the factorial structure of Ghanaian twelfth-grade students' views on mathematics. In Edited by: **B.Pepin** & **B.Roesken** (Eds.), *From beliefs to dynamic affect systems in mathematics education: Exploring a mosaic of relationships and interactions*. Switzerland: Springer.

Bofah, E. A., & **Hannula, M.** (submitted). Students' views on mathematics in single-sex and coed classrooms in Ghana. *International Journal of Mathematical Education in Science and Technology*.

Hannula, M. S., & **Laakso, J.** (**2011**, July). The structure of mathematics related beliefs, attitudes and motivation among Finnish grade 4 and grade 8 students. In Edited by: **B.Ubuz** (Ed.), *Proceedings of the 35th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 1), PME, Ankara, Turkey.

Hannula, M. S., **Kaasila, R.**, **Laine, A.**, & **Pehkonen, E.** (**2005**). The structure of student teacher's view of mathematics at the beginning of their studies. In Edited by: **M.Bosch** (Ed.), *Proceedings of the fourth congress of European research in mathematics education (CERME 4)* (pp. 205–214). Sant Feliu de Guíxols, Spain: Fundemi IQS – Universitat Ramon Llull.

Raykov, T. (**2012**). Scale construction and development using structural equation modeling. In Edited by: **R. H.Hoyle** (Ed.), *Handbook of structural equation modeling* (pp. 472–492). New York, NY: The Guilford.

Roesken, B., **Hannula, M. S.**, & **Pehkonen, E.** (**2011**). Dimensions of students' views of themselves as learners of mathematics. *ZDM*, 43, 497–506. doi: http://dx.doi.org/10.1007/s11858-011-0315-8http://dx.doi.org/10.1007/s11858-011-0315-8

Van de Vijver, F. J. R., & **Leung, K.** (**2000**). Methodological issues in psychological research on culture. *Journal of Cross-Cultural Psychology*, 31, 33–51. doi: http://dx.doi.org/10.1177/0022022100031001004http://dx.doi.org/10.1177/0022022100031001004

References

Brown, T. A. (**2006**). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Brown, T. A., & **Moore, M. T.** (**2012**). Confirmatory factor analysis. In Edited by: **R. H.Hoyle** (Ed.), *Handbook of structural equation modeling* (pp. 361–379). London, England: Guilford Press.

Chen, F. F. (**2007**). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural*

Equation Modeling, 14, 464–504. doi: http://dx.doi.org/10.1080/10705510701301834http://dx.doi.org/10.1080/10705510701301834

Cheung, G. W., & **Rensvold, R. B.** (**2002**). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233–255. doi: http://dx.doi.org/10.1207/S15328007SEM0902_5http://dx.doi.org/10.1207/S15328007SEM0902_5

Curran, P. J., **West, S. G.**, & **Finch, J. F.** (**1996**). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, 1, 16–29. doi: http://dx.doi.org/10.1037/1082-989X.1.1.16http://dx.doi.org/10.1037/1082-989X.1.1.16

Edelen, M. O., & **Reeve, B. B.** (**2007**). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, 16, 5–18. doi: http://dx.doi.org/10.1007/s11136-007-9198-0http://dx.doi.org/10.1007/s11136-007-9198-0

Geldhof, G. J., **Preacher, K. J.**, & **Zyphur, M. J.** (**2013**). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*. Advance online publication. doi: http://dx.doi.org/10.1037/a0032138

Hair, J. F., **Black, W. C.**, **Babin, B. J.**, & **Anderson, R. E.** (**2010**). *Multivariate data analysis*. Upper Saddle River, NJ: Prentice Hall.

Hayton, J. C., **Allen, D. G.**, & **Scarpello, V.** (**2004**). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, 7, 191–205. doi: http://dx.doi.org/10.1177/1094428104263675http://dx.doi.org/10.1177/1094428104263675

MacCallum, R. C., **Roznowski, M.**, & **Necowitz, L. B.** (**1992**). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, 111, 490–504. doi: http://dx.doi.org/10.1037/0033-2909.111.3.490http://dx.doi.org/10.1037/0033-2909.111.3.490

McLeod, D. B. (**1992**). Research on affect in mathematics education: A reconceptualization. In Edited by: **D. A.Grouws** (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 575–596). New York, NY: Macmillan Publishers.

Marsh, H. W., **Muthén, B. O.**, **Asparouhov, T.**, **Lüdtke, O.**, **Robitzsch, A.**, **Morin, A. J.**, & **Trautwein, U.** (**2009**). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal*, 16, 439–476. doi: http://dx.doi.org/10.1080/10705510903008220http://dx.doi.org/10.1080/10705510903008220

Mullis, I. V., **Martin, M. O.**, **Foy, P.**, & **Arora, A.** (**2012**). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College:

Muthén, L. K., & **Muthén, B. O.** (**1998–2012**). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén &

Muthén.

Novick, M. R., & **Lewis, C.** (**1967**). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 32(1), 1–13. doi: http://dx.doi.org/10.1007/BF02289400http://dx.doi.org/10.1007/BF02289400

Raykov, T. (**2012**). Scale construction and development using structural equation modeling. In Edited by: **R. H.Hoyle** (Ed.), *Handbook of structural equation modeling* (pp. 472–492). New York, NY: The Guilford.

Russell, D. W. (**2002**). In search of underlying dimensions: The use (and abuse) of factor analysis in personality and social psychology bulletin. *Personality and Social Psychology Bulletin*, 28, 1629–1646. doi: http://dx.doi.org/10.1177/014616702237645http://dx.doi.org/10.1177/014616702237645

Schmitt, N. (**1996**). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8 (4), 350–353. doi: http://dx.doi.org/10.1037/1040-3590.8.4.350http://dx.doi.org/10.1037/1040-3590.8.4.350

Schmitt, T. A. (**2011**). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment*, 29, 304–321. doi: http://dx.doi.org/10.1177/0734282911406653http://dx.doi.org/10.1177/0734282911406653