EXAMINING MATHEMATICS COACHING

CONFIRMATORY FACTOR ANALYSIS OF THE ITC-COP SUBSCALE

OCTOBER 2011

FUNDING BY THE NATIONAL SCIENCE FOUNDATION
DISCOVERY RESEARCH K-12 PROGRAM (DR K-12),
AWARD NO. 0918326

CONFIRMATORY FACTOR ANALYSIS OF THE ITC-COP SUBSCALE

DAVID YOPP, ELIZABETH A. BURROUGHS, JOHN T. SUTTON,
DAN JESSE, AND MARK GREENWOOD

OCTOBER 2011

ACKNOWLEDGMENTS

This report is funded through a grant from The National Science Foundation,
Discovery Research K-12 Program (DR K-12), Award No. 0918326.
Any opinions, findings, and conclusions or recommendations expressed in this material
are those of the author(s) and do not necessarily reflect the views of The National Science Foundation.

Suggested citation: Yopp, D., Burroughs, E. A., Sutton, J. T., Jesse, D., & Greenwood, M. (2011). *Confirmatory Factor Analysis of the ITC-COP Subscale.* Bozeman: Montana State University and Denver, CO: RMC Research Corporation.

For questions about this report, please contact the Examining Mathematics Coaching Project
at (877) 572-5032 or email: *emc@math.montana.edu*

TABLE OF CONTENTS

TABLE OF EXHIBITS

INTRODUCTION

OVERVIEW

The Examining Mathematics Coaching (EMC) Project is a research and development effort that examines the effects of knowledge for coaching embedded in an innovative, previously developed coaching model applied to a population of K-8 teachers in diverse settings. EMC addresses the National Science Foundation (NSF) DRK-12 Proposal Solicitation challenge: How can the ability of teachers to provide Science, Technology, Engineering, and Mathematics (STEM) education be enhanced? The STEM discipline addressed is mathematics, and the audience addressed is school-based mathematics coaches along with the teachers they coach. The research sites include rural, urban, and suburban school districts, along with districts whose student populations are predominantly Native American.

BACKGROUND

The EMC Project is conducting research on knowledge that contributes to successful coaching in two domains: coaching knowledge and mathematics content knowledge. The influence of these knowledge domains on both coaches and teachers is examined (1) by investigating correlations between assessments of coach and teacher knowledge and practice in each domain, and (2) by investigating causal effects of targeted professional development for coaches. The impact of coaches' knowledge is measured through the lens of teacher change in the domains of content knowledge (focusing on number and operations), reform- and standards-based practice, attitudes and beliefs, mathematics teacher efficacy, and perceptions of coach effectiveness. Research findings are used to develop, modify, and apply tools to assist schools and STEM professional developers in areas of coaching such as selection, training, and assessment of impact.

The National Mathematics Advisory Panel (2008) reported that school districts across the country are using mathematics specialists, including coaches, to improve instruction in elementary school systems. They also note that there is little research supporting the effectiveness of mathematics specialists or, for that matter, the cost-effectiveness of using specialists. Despite the lack of supporting evidence, many schools are turning to coaching as a school-based effort to increase teacher effectiveness and student achievement. At present, a comprehensive understanding of the effectiveness of coaching does not exist, even though the components of coaching involve considerable cost and logistical effort for schools. Moreover, there is no common vocabulary to describe the full scope of coaching in all its forms; the particulars appear to be highly situational and not necessarily based on mutual agreement about what constitutes best practice for coach selection, training, and implementation. The question of what types of knowledge and skills coaches need to be effective has not been sufficiently addressed in education research.

DATA ANALYSIS METHODS

RESEARCH QUESTIONS

Hypothesis: The EMC Project posits that the effectiveness of a mathematics classroom coach is linked to several domains of knowledge. We posit that coaching knowledge and mathematics content knowledge contribute significantly to a coach's effectiveness as measured by the positive impact on teacher practice, attitudes, and beliefs. To test this hypothesis, this project is designed to address the following research questions:

1. To what extent does a coach's depth of knowledge in two primary domains (coaching knowledge and mathematics content knowledge) influence coaching effectiveness?
2. To what extent does professional development targeting these two knowledge domains improve coaching effectiveness?
3. To what extent are the effects of targeted professional development on coaching effectiveness explained by increases in coaching knowledge and mathematics content knowledge?

Specifically, the project is looking to address the following coaching and teacher outcomes:

COACHING OUTCOMES

1. To what extent does participation in EMC increase a coach's coaching knowledge?
2. To what extent does participation in EMC increase a coach's mathematics content knowledge?
3. What is the relationship between coaching knowledge and mathematics content knowledge?
4. Are there differences in levels of mathematics content knowledge, coaching knowledge, and coaching skills between the two treatment groups?
5. What factors, such as variables associated with program delivery, coaching characteristics, or teacher characteristics, influence coaching knowledge?
6. What factors, such as variables associated with program delivery, coaching characteristics, or teacher characteristics, influence a coach's mathematics content knowledge?

TEACHER OUTCOMES

1. What is the relationship between mathematics teacher efficacy (MTE) and teacher mathematics content knowledge?
2. To what extent does EMC coaching influence teacher effectiveness through predictor variables measuring teacher characteristics?
3. Are there differences in teacher effectiveness and teacher characteristics between the two treatment groups?

Based on the research questions, the project has six specific hypotheses that will be tested:

- Hypothesis 1: Higher ratings in coaches' coaching knowledge will result in greater positive changes in teacher indicators of coaching effectiveness.

- Hypothesis 2: Higher ratings in coaches' mathematics content knowledge will result in greater positive changes in teacher indicators of coaching effectiveness.

- Hypothesis 3: Coaches' coaching knowledge and mathematics content knowledge will have positive, non-overlapping relationships to coaching effectiveness.

- Hypothesis 4: Coaching effectiveness will be higher for coaches who have received targeted professional development.

- Hypothesis 5: Coaches' coaching knowledge or mathematics content knowledge that was targeted by professional development will be higher than that of coaches who have not received professional development in that domain.

- Hypothesis 6: Targeted professional development will influence coaching effectiveness through increases in coaches' coaching knowledge and mathematics content knowledge.

STATISTICAL TERMS AND SYMBOLS

Throughout this report there are a number of statistical terms and symbols used in describing the data analysis methods and reporting the data analysis findings in tables. Exhibit 1 presents these terms and their descriptions to assist the reader in understanding them as they are used in this report.

EXHIBIT 1. STATISTICAL TERMS AND SYMBOLS USED IN THIS REPORT

Term or Symbol	Description
Comparative Fit Index (CFI)	This incremental measure of fit is directly based on the non-centrality measure. Let $d = \chi^2 - df$ where *df* are the degrees of freedom of the model. The Comparative Fit Index or CFI equals $\frac{d(\text{Null Model}) - d(\text{Proposed Model})}{d(\text{Null Model})}$ If the index is greater than one, it is set at one and if less than zero, it is set to zero. As the name suggests, this statistic is a commonly used index of comparative or incremental fit.
Confirmatory Factor Analysis	CFA is used to test whether measures of a construct are consistent with a researcher's understanding of the nature of that construct (or factor). In

Term or Symbol	Description
(CFA cont'd)	contrast to exploratory factor analysis, where all loadings are free to vary, CFA allows for the explicit constraint of certain loadings to be zero. CFI values close to .95 or greater indicate goodness of fit. Hu & Bentler, 1999, in (Brown, 2006).
Correlational analysis	Correlational analysis is the use of statistical correlation to evaluate the strength of the relations between variables, such that systematic changes in the value of one variable are accompanied by systematic changes in the other.
Cronbach's alpha (α)	Cronbach's alpha (α) is a measure of the reliability or internal consistency of a composite measure or scale that is based on multiple survey items. Values range from 0 to 1.
Exploratory factor analysis (EFA)	Exploratory factor analysis (EFA) is generally used to discover the factor structure of a measure and to examine its internal reliability. EFA is often recommended when researchers have no hypotheses about the nature of the underlying factor structure of their measure. Exploratory factor analysis has three basic decision points: (1) deciding the number of factors, (2) choosing an extraction method, and (3) choosing a rotation method.
Internal reliability	The internal consistency of survey instruments is a measure of reliability of different survey items intended to measure the same characteristic.
Mean	The mean or average value is a measure of central tendency computed by adding a set of values and dividing the sum by the total number of values.
N	*N* is the total number in a sample.
Pearson *r*	The Pearson product-moment correlation coefficient (*r*) is a measure of the relationship between two variables (i.e., a measure of the tendency of the variables to increase or decrease together). Values range from -1 to +1. A correlation of +1 indicates perfect positive correlation (i.e., the two variables increase or decrease together). A correlation of -1 indicates perfect negative correlation (i.e., one variable decreases as the other increases, or vice versa).
Root Mean Square Error of Approximation (RMSEA)	This absolute measure of fit is based on the non-centrality parameter. Its computational formula is: $$\frac{\sqrt{(\chi^2 - df)}}{\sqrt{[df(N - 1)]}}$$ where N is the sample size and *df* are the degrees of freedom of the model. The measure is positively biased (i.e., tends to be too large) and the amount of the bias depends on smallness of sample size and df, primarily the latter. The RMSEA is currently the most popular measure of model fit and is now reported in virtually all papers that use CFA or SEM, and some refer to the measure as the "Ramsey." It is an index that has been categorized as a parsimony correction. RMSEA values that are close to .06 or below indicate reasonable model fit. Hu & Bentler, 1999, in (Brown, 2006).
SD	The standard deviation (*SD*) is a measure of how spread out a set of values is. Higher SD indicates greater variability in data across respondents.

Term or Symbol	Description
Standardized Root Mean Square Residual (SRMR)	The SRMR is an absolute measure of fit and is defined as the standardized difference between the observed correlation and the predicted correlation. It is a positively biased measure and an absolute measure of fit. The bias is greater for small N and for low df studies. This measure tends to be smaller as sample size increases and as the number of parameters in the model increases. The SRMR is an absolute measure of fit, and a value of zero indicates perfect fit. The SRMR has no penalty for model complexity. It is a statistic that has been categorized as an absolute fit index. SRMR values that are close to .08 or below are considered to be evidence for goodness of fit. Hu & Bentler, 1999, in (Brown, 2006).

ITC-COP Subscale Selection

Inside the Classroom – Classroom Observation Protocol (ITC-COP)

The ITC-COP was developed by Horizon Research Inc. (2000) and was designed to measure the quality of an observed K-12 science or mathematics classroom lesson by examining the design, implementation, mathematics/science content, and culture of that lesson. Items on the ITC-COP are based on standards of quality mathematics and science instruction as outlined in The National Council of Teachers of Mathematics Standards and the National Science Education Standards. The data produced from the instrument were found to be reliable and valid, and the instrument is used to observe the teacher participants periodically throughout the study (Yopp et al., 2010). Factor analysis of the items related to frequency of responses suggests three underlying factors related to teaching skills, content knowledge, and collaboration. Changes in both the factor scores higher level synthesis and capsule ratings will be considered.

The ITC-COP measure of classroom behavior is being used to objectively document changes over time in classrooms of teachers who have been interacting with trained coaches. While the ITC-COP is a validated measure, it is somewhat general in nature relative to the goals of the project, so an effort was made to determine whether it was a viable tool for documenting project outcomes, and whether it might be modified to better address project needs. Exploratory factor analysis of initial results was conducted with these purposes in mind and strongly suggested that subsets of reflective items that are part of the tool provided valuable information about classroom activity. Since the tool had already been validated in other settings, there was an interest in determining whether use of all items used to reflect on classroom activities might also provide useful information. Therefore, internal reliability, correlational analyses and Confirmatory Factor Analysis (CFA) techniques were used to determine how the tool might best be utilized to document EMC Project outcomes.

METHODOLOGY

The ITC-COP was used by 12 trained observers in 196 classroom observations in the spring and fall of 2010 to collect pre-test or baseline information, and again in the spring of 2011 to collect follow-up or post-test information in 164 classrooms. The instrument is used to collect information about classroom context, design of instruction, implementation of instruction, delivery of mathematics content, and classroom culture. Additionally, observers provide capsule ratings of the entire lesson on a 5-point scale: "1" = ineffective instruction and "5" = exemplary instruction. With 5 exceptions, teachers were visited by the same observers in 2010 and 2011.

The instrument includes four aspects of the lesson: design, implementation, mathematics content, and classroom culture. Each aspect contains sets of "reflective" questions that are scored on a 7-point Likert-type scale: "1" = not at all, to "5" = to a great extent (with "6" = don't know and "7" = N/A). After rating classroom activity on each of the four aspects, observers generated synthesis ratings on a 5-point Likert-type scale: For example, "1" = Design of the lesson not at all reflective of best practices in mathematics/science education, to "5" = Design of the lesson extremely reflective of best practices in mathematics/science education. While the publishers of the instrumentation never intended the reflective items to be formally scored, they provided additional information about classroom activity that was of interest to EMC. Therefore, the utility of using the information provided by these subscale measures designed to create a frame of mind for making synthesis ratings was statistically explored.

Internal subscale reliabilities were calculated using Cronbach's alpha. Pearson correlation coefficients were calculated between all subscales for both a pre-test and post-test administration of the tool, and CFA was conducted with raw data using LISREL 8.80. Goodness of fit was evaluated by using three commonly used indices explained in the results: Standardized Root Mean Square Residual (SRMR), the Comparative Fit Index (CFI), and the Root Mean Square Error of Approximation (RMSEA). The first item in each subscale was fixed and initially set to 1.00, as is conventional in CFA.

The purpose of the analyses reported here is to inform decisions about how best to use the data from the ITC-COP to document changes in classrooms of teachers who have been coached by EMC participants. Capsule ratings and synthesis ratings are already being used, but a decision needs to be made about whether full subscales measuring four aspects of the lesson should be used, versus a more streamlined, empirically derived set of three subscales resulting from exploratory factor analysis conducted with pre-test or baseline data.

ANALYSES

Two separate subscales were created from these observations conducted by trained staff. The first set of subscales was empirically derived from the administration of the pre-test and analyzed using exploratory factor analysis (Yopp et al., 2011). The second set of subscales, which make use of all items, was derived by using a set of cognitive framing questions designed to create a mental set for making valid synthesis ratings of classroom activity.

In order to make a determination about whether it would be useful to employ four subscales derived directly from the ITC-COP measure using their four categories (design, implementation, content, and classroom culture)—as opposed to three categories (mathematics content knowledge, student-centered classroom culture, and student collaborative classroom culture) that employed most of the same items—first internal reliabilities for each subscale in the two groups of items were calculated. Exhibit 2 contrasts the psychometric properties of the two groups of subscales. These analyses were conducted for pre-test and post-test measures. For the most part, all subscale measures are characterized by high reliability, particularly on the post-test measures.

EXHIBIT 2. PSYCHOMETRIC PROPERTIES OF TWO SUBSCALE SETS CREATED FROM ITC-COP OBSERVATIONS

	2010 Pre-test *N* = 196				2011 Post-test *N* = 164		
	N	Mean	Standard Deviation (*SD*)	Cronbach's Alpha	Mean	*SD*	Cronbach's Alpha
Design	8	2.86	.788	.908	3.15	.829	.921
Implementation	6	2.89	.816	.911	3.24	.830	.898
Mathematics Content	9	2.91	.736	.897	3.15	.775	.914
Classroom Culture	6	2.87	.859	.921	3.19	.903	.916
Mathematics Content Knowledge	9	3.12	.752	.910	3.40	.723	.907
Classroom Culture – Student Centeredness	6	2.93	.894	.938	3.25	.896	.894
Classroom Culture – Student Collaboration	3	2.78	.906	.764	3.03	.966	.883

Next, Pearson correlation coefficients were calculated for each of the indicators derived from the ITC-COP. Exhibits 3 and 4 display the results of these analyses for 2010 and 2011, respectively, and reveal that all of the indicators are highly correlated.

EXHIBIT 3. INTERCORRELATIONS OF ITC-COP PRE-TEST INDICATORS: 2010 (N = 196)

	Capsule Rating	Design	Design Synthesis	Implementation	Implementation Synthesis	Mathematics Content	Content Synthesis	Classroom Culture	Culture Synthesis	Mathematics Content Knowledge	Classroom Culture – Student Centeredness	Classroom Culture – Student Collaboration
Capsule Rating	1.000											
Design	.783***	1.000										
Design Synthesis	.780***	.906***	1.000									
Implementation	.862***	.869***	.819***	1.000								
Implementation Synthesis	.837***	.809***	.790***	.919***	1.000							
Mathematics Content	.812***	.840***	.780***	.888***	.806***	1.000						
Content Synthesis	.752***	.770***	.709***	.789***	.762***	.881***	1.000					
Classroom Culture	.828***	.850***	.786***	.894***	.822***	.828***	.718***	1.000				
Culture Synthesis	.807***	.783***	.739***	.842***	.797***	.764***	.680***	.926***	1.000			
Mathematics Content Knowledge	.785***	.876***	.806***	.892***	.794***	.943***	.832***	.793***	.742***	1.000		
Classroom Culture – Student Centeredness	.836***	.836***	.774***	.919***	.845***	.836***	.731***	.982***	.925***	.809***	1.000	
Classroom Culture – Student Collaboration	.630***	.819***	.741***	.692***	.625***	.663***	.581***	.808***	.722***	.636***	.725***	1.000
Standard Deviation	*1.585*	*.788*	*.864*	*.815*	*.905*	*.736*	*.812*	*.859*	*.928*	*.752*	*.894*	*.906*

*** $p < .001$, 2-tailed test. Measures of mathematics content knowledge, classroom culture–student centeredness, and classroom culture–student collaboration share some items with design, implementation, mathematics content, and classroom culture subscales.

EXHIBIT 4. INTERCORRELATIONS OF ITC-COP POST-TEST INDICATORS: 2011 (N = 164)

	Capsule Rating	Design	Design Synthesis	Implementation	Implementation Synthesis	Mathematics Content	Content Synthesis	Classroom Culture	Culture Synthesis	Mathematics Content Knowledge	Classroom Culture – Student Centeredness	Classroom Culture – Student Collaboration
Capsule Rating	1.000											
Design	.819***	1.000										
Design Synthesis	.815***	.906***	1.000									
Implementation	.874***	.856***	.814***	1.000								
Implementation Synthesis	.880***	.832***	.798***	.913***	1.000							
Mathematics Content	.875***	.819***	.776***	.848***	.796***	1.000						
Content Synthesis	.814***	.731***	.703***	.783***	.726***	.910***	1.000					
Classroom Culture	.833***	.845***	.764***	.872***	.826***	.813***	.718***	1.000				
Culture Synthesis	.835***	.787***	.721***	.823***	.796***	.790***	.710***	.927***	1.000			
Mathematics Content Knowledge	.856***	.865***	.801***	.868***	.804***	.952***	.857***	.801***	.759***	1.000		
Classroom Culture – Student Centeredness	.833***	.824***	.742***	.894***	.833***	.806***	.715***	.984***	.923***	.805***	1.000	
Classroom Culture – Student Collaboration	.711***	.862***	.789***	.745***	.722***	.697***	.601***	.850***	.774***	.690***	.789***	1.000
Standard Deviation	*1.665*	*.829*	*.922*	*.830*	*.924*	*.775*	*.847*	*.903*	*.939*	*.723*	*.896*	*.966*

*** $p < .001$, 2-tailed test. Measures of mathematics content knowledge, classroom culture–student centeredness, and classroom culture–student collaboration share some items with design, implementation, mathematics content, and classroom culture subscales.

CONFIRMATORY FACTOR ANALYSIS

In order to provide additional information about the viability of the two subscale creation methods being evaluated, CFA was conducted upon each. Exhibits 5 and 6 display the CFA results for the four subscales derived from the reflective questions from the pre-test and post-test observations. Exhibits 7 and 8 display CFA results from the empirically derived subscales grounded in exploratory factor analyses conducted upon pre-test data. Almost all of the loadings are relatively high and suggest that there is little difference between the two measurement approaches. It should be noted that Exhibit 6 directly reflects the results of the exploratory factor analysis reported earlier by EMC (Yopp et al., 2011) and should be interpreted with caution.

EXHIBIT 5. CONFIRMATORY FACTOR ANALYSIS RESULTS FOR REFLECTIVE ITEMS ON THE 2010 ITC-COP PRE-TEST (N = 196)

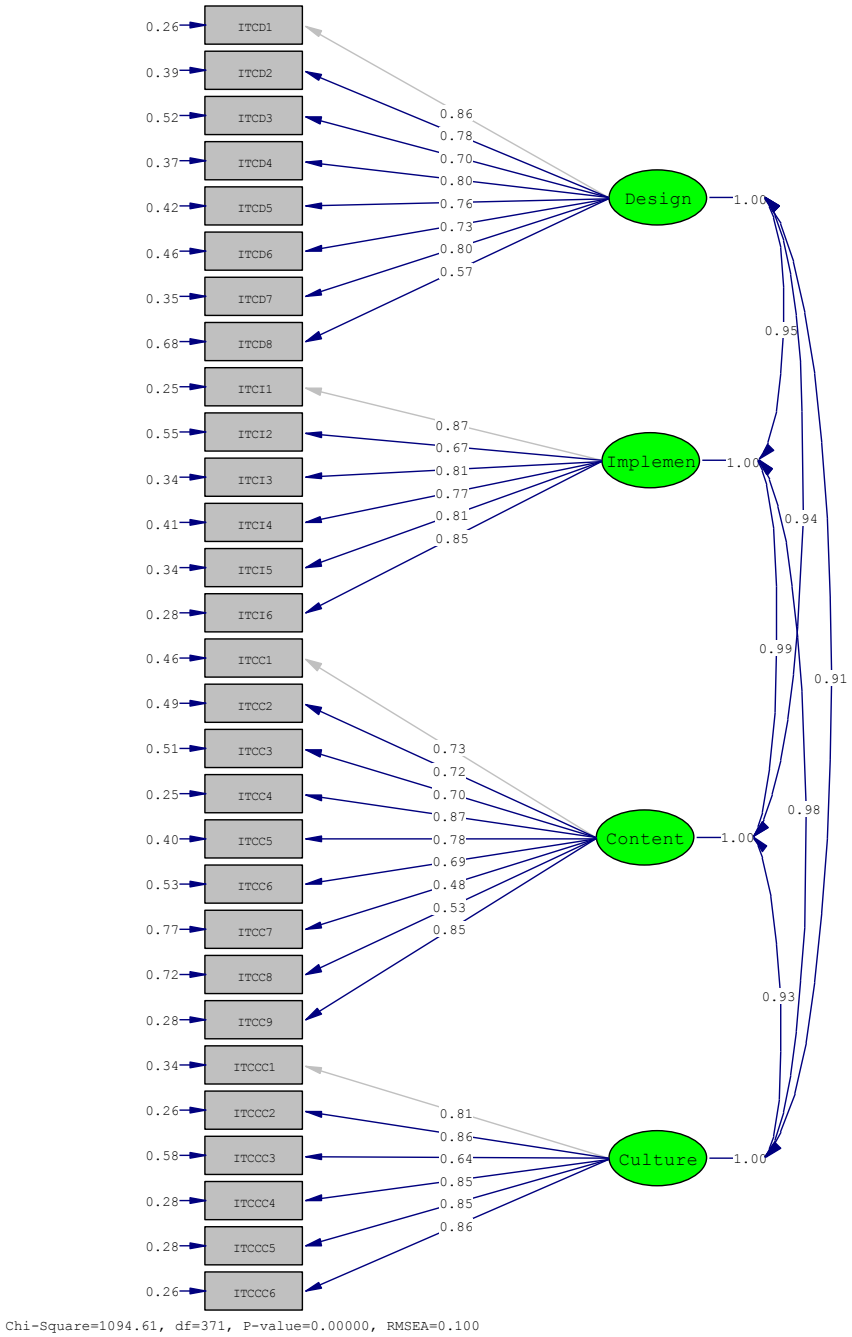
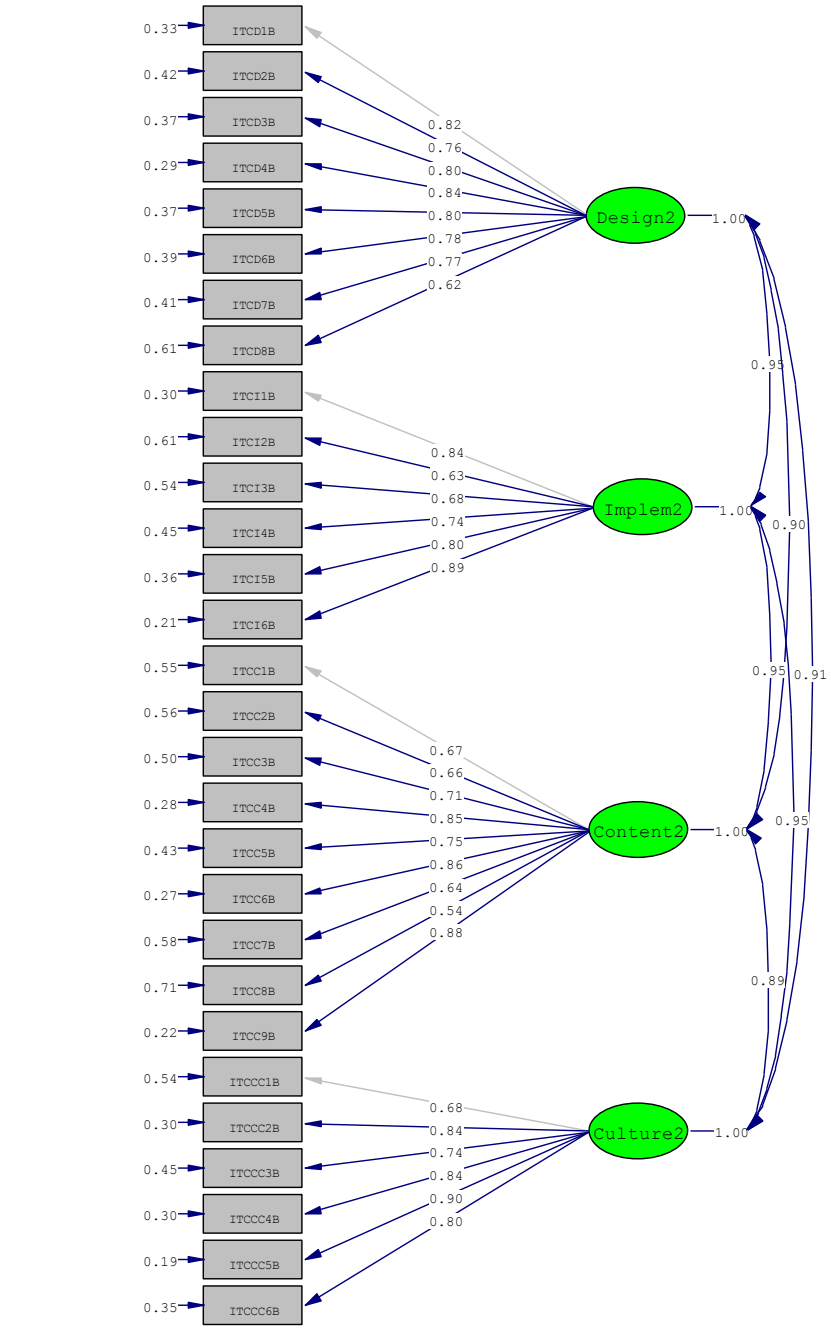
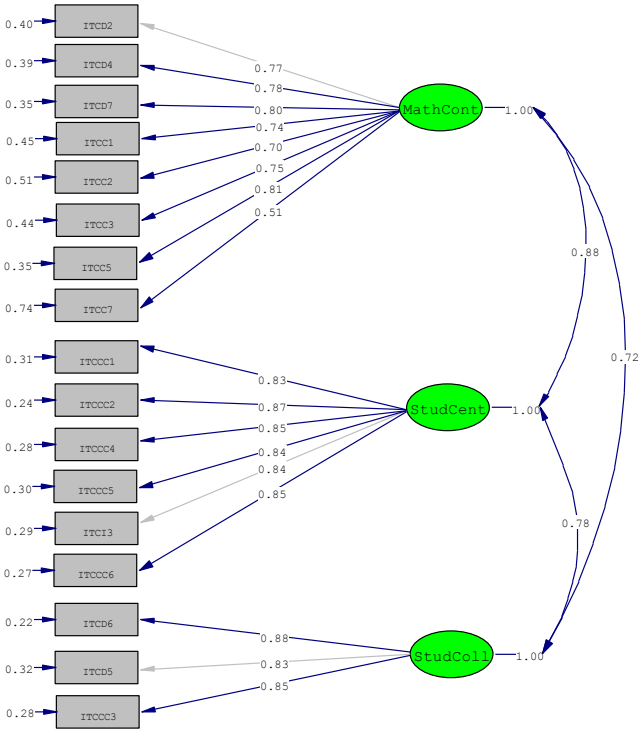


EXHIBIT 6. CONFIRMATORY FACTOR ANALYSIS RESULTS FOR REFLECTIVE ITEMS ON THE 2011 ITC-COP POST-TEST (N = 164)



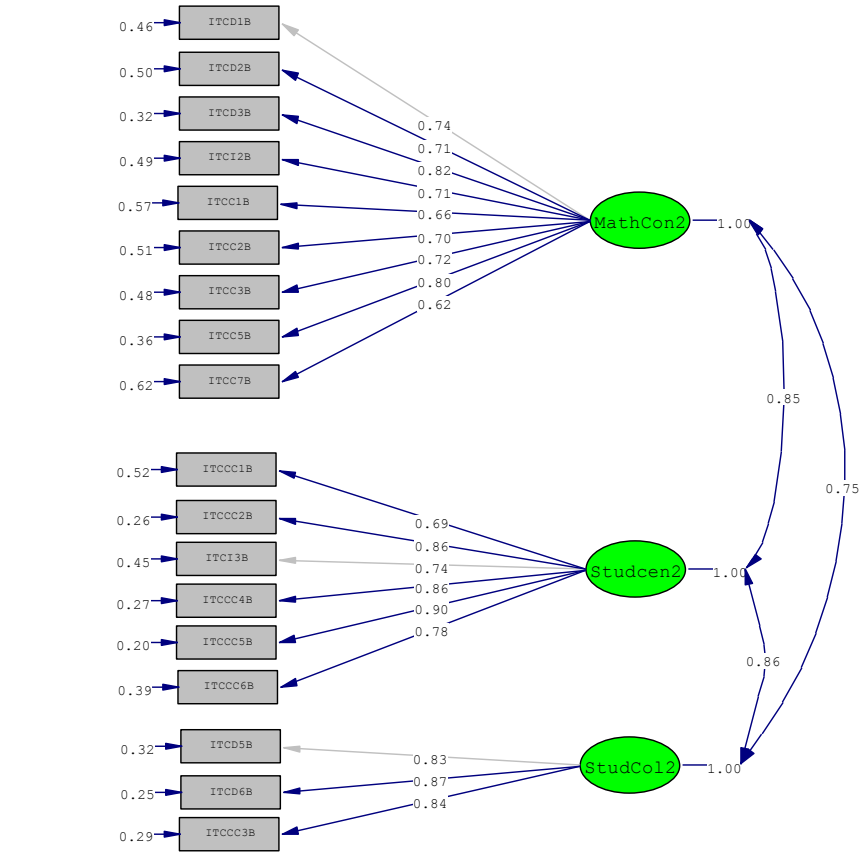
Chi-Square=1133.21, df=371, P-value=0.00000, RMSEA=0.113

EXHIBIT 7. CONFIRMATORY FACTOR ANALYSIS RESULTS FOR EMPIRICALLY DERIVED ITEMS ON THE 2010 ITC-COP PRE-TEST (N = 196)



Chi-Square=284.94, df=116, P-value=0.00000, RMSEA=0.086

EXHIBIT 8. CONFIRMATORY FACTOR ANALYSIS RESULTS FOR REFLECTIVE ITEMS ON THE 2011 ITC-COP POST-TEST (N = 164)



Chi-Square=436.53, df=132, P-value=0.00000, RMSEA=0.119

Modeled after the protocol of Brown (2006), three separate indices for goodness of model fit are presented: absolute fit, fit adjusted for model parsimony, and comparative or incremental fit. The SRMR was reported as an index of absolute fit. It is "… the average discrepancy between the correlations observed in the input matrix and the correlations predicted by the model" (p. 82). A value of "0" indicates a perfect fit, and a value close to .08 or below is considered to be a heuristic for interpretation indicating good fit. The index chosen for fit adjusting for model parsimony, the RMSEA, "… assesses the extent to which a model fits reasonably well in the population" (p. 83). An RMSEA value of 0 is a perfect fit, and a value close to .06 or below is considered to be a good fit. The CFI measures comparative or incremental fit. This index ranges from 0 to 1, and a CFI close to .95 or greater is considered to be a good fit.

Exhibit 9 displays the fit indices for each of the CFA models presented and reveals that all of the absolute fit indices are below the .08 criteria; all of the comparative or incremental fit indices are above the .95 criteria, but none of the fit adjusting for model parsimony results are below the .06 criteria. Two of the three criteria indicate adequate model fit. However, failure to meet the RMSEA criteria suggests that the models might not be adequate. The Pre-Test Empirical model is what might be called a "mediocre fit," but the other three models should be rejected using the RMSEA criteria (Brown, p. 87). Nonetheless, since sample sizes are somewhat small (N = 196 and 164, respectively), these results should not be cause for concern.

EXHIBIT 9. PSYCHOMETRIC PROPERTIES OF TWO SUBSCALE SETS CREATED FROM ITC-COP OBSERVATIONS

	Absolute Fit (SRMR)	Comparative or Incremental Fit (CFI)	Fit Adjusting for Model Parsimony (RMSEA)
Pre-Test Reflective	0.051	0.098	0.100
Post-Test Reflective	0.060	0.097	0.113
Pre-Test Empirical	0.050	0.098	0.086
Post-Test Empirical	0.065	0.096	0.119

CONCLUSIONS AND IMPLICATIONS

First, almost all of the subscales tested in these analyses were highly reliable, and that reliability increased for the post-test results. These findings suggest that observers are highly trained and that the instrumentation is viable for documenting classroom behaviors of interest to EMC. Second, almost all of the measures studied in this investigation are highly correlated. This may suggest that breaking them apart into subscale measures may be problematic on the surface, but may also reflect the reliable and valid nature of the measures being used. Third, CFA analyses strongly suggest that both models examined—the reflective model and the empirically derived model—are adequate to describe the data collected from numerous classroom observations. It is safe to assume that the factor structure has been confirmed for both the reflective and empirical subscale sets.

There are some limitations to this study that should be noted. This is a preliminary analysis, and other more detailed analyses are possible and perhaps desirable. Some of the data analyzed here was used to empirically derive subscales, so results should be interpreted with caution. Specifically, empirically derived subscale analyses for the pre-test use the same data to create the subscales, so results are not independent. More faith should be put into post-test analyses, but, again, the observers are the same and the population is a subset of the pre-test population. Sample sizes are small for CFA. A minimal heuristic is 5 observations per concept measured. In this case, that means that no more than 32 or 33 concepts should be measured with post-test data. Again, caution is warranted when interpreting results.

Since the results for the two approaches are so close, it is difficult to make a confident determination based on these analyses alone as to which approach should be used moving forward. Other considerations should be explored, such as whether the empirically derived subscales are a better theoretical match for project outcomes than the use of the reflective measures. All of this should be considered with the knowledge that the reflective items used on the ITC-COP were never intended to be analyzed in detail and were meant to create a mindset for establishing other ratings. Nonetheless, it was worthwhile to explore the possibility of tailoring the ITC-COP for this project.

All of the measures are highly correlated. Subscale measures also predict other outcomes on the instrumentation. CFA and reliability analyses suggest that either r approach has merit. Hopefully, the analyses reported here will contribute to the discussion about which measurement approaches are best for documenting project impact in the classroom.

REFERENCES

Brown, T. (2006). *Confirmatory Factor Analysis for Applied Research.* New York: The Guilford Press.

Examining Mathematics Coaching. (2010). *Instrumentation Codebook.* Bozeman, MT: EMC.

Horizon Research Inc. (2000). *Inside the classroom: Observation and analytic protocol.* Chapel Hill, NC: Author. Retrieved from http://www.horizon-research.com/insidetheclassroom/instruments/obs.php

Yopp, D., Burroughs, E. A., Sutton, J. T., Swackhamer, L. E., & Greenwood, M. (2010). *Construct reliability and validity of EMC instrumentation*. Denver, CO: RMC Research Corporation.

Yopp, D., Burroughs, E. A., Sutton, J. T., Swackhamer, L., and Greenwood, M. (2011). *Research Protocol: March 2011*. Bozeman, MT: Montana State University and Denver, CO: RMC Research Corporation.