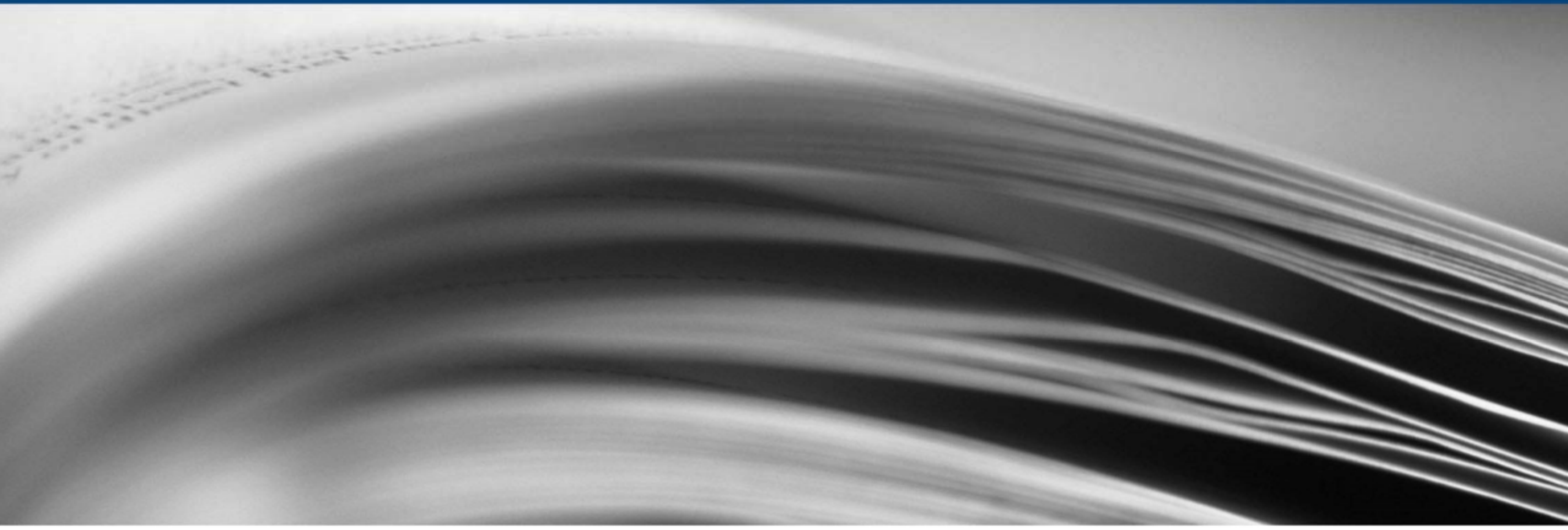


Examining Mathematic Coaching

# Analysis of the Coaching Knowledge Survey:
## Evidence for Validation and Next Steps

Prepared for:
**Montana State University**
Department of Mathematical Sciences
Wilson 2-299 B
Bozeman, MT 59717-2400

Prepared by:
**RMC Research Corporation**
633 17th Street, Suite 2100
Denver, CO  80202

March 2014

# EXAMINING MATHEMATICS COACHING: ANALYSIS OF THE COACHING KNOWLEDGE SURVEY EVIDENCE FOR VALIDATION AND NEXT STEPS

PREPARED FOR:

**David Yopp, PI**
**Beth Burroughs, Co-PI**

**Montana State University**
Department of Mathematical Sciences
Wilson 2-299 B
Bozeman, MT 59717-2400

PREPARED BY:

**RMC Research Corporation**
633 17th Street, Suite 2100
Denver, CO  80202

**MARCH 2014**

## Acknowledgments

This report was prepared for Montana State University

For questions about this report, please contact Dr. John T. Sutton
at 1-800-922-3636, or e-mail at sutton@rmcdenver.com.

RMC Research Corporation is an Equal Employment Opportunity
and Affirmative Action Employer and a Drug-Free Workplace.

# Table of Contents

# TABLE OF EXHIBITS

# Analysis of the Coaching Knowledge Survey: Evidence for Validation and Next Steps

## (Jesse, Sutton and Linick, 2014)

## Summary

This report describes a number of approaches that were taken in 2013-2014 to validate the 39-item Coaching Knowledge Survey (CKS). Three data sets are to be utilized in this effort: Sample 1 (*N* = 252) includes a convenience sample of pilot test respondents pooled with respondents that were EMC coaches before the project began; Sample 2 consists of repeated measures of CKS and other coaching effectiveness measures aggregated at the coach level ranging from about 40 to 50 participants; and Sample 3 (*N* = 191) includes only the original convenience sample of pilot test respondents. These three data sets were used in a variety of analyses designed for validating information for the CKS. A confirmatory factor analysis (CFA) of Sample 1 data revealed issues with the 39-item version of the CKS, so the results were used to identify a subset of plausible items for potential validation. Multidimensional scaling (MDS) was used to further explore the data set, and suggested multidimensionality of the 39 items. Analyses of Sample 2 results, which were longitudinal and linked to other independent perceptual and behavioral data, revealed that the CKS was negatively correlated to teacher scores on the Teacher Survey and on classroom observations conducted by trained observers, and led to the revisiting of the structure of the CKS. Separate analyses of teacher survey perceptual data and classroom observation data was conducted to identify a subset of 6 high performing "supercoaches", coaches who were able to achieve substantial progress with their mentees on these two measures across the course of the EMC project. Some aspects of predictive, convergent, divergent and concurrent reliability were addressed by using these data sets. High performing results from the supercoaches on the CKS were examined to identify items that discriminated between the two groups (supercoaches and others), discriminant analysis was conducted on CKS items to determine which items predicted membership in training groups, and Spearman correlations were calculated between CKS items, teacher survey results, and classroom observation results (see Appendix).

Following the identification of threats to the validity issues with the CKS, exploratory factor analysis (EFA) was conducted with Sample 3 data, which explored the possibility of multiple factors within the 39 items measured in the CKS (see Appendix). Problematic issues emerged when forcing multiple factors, such as reverse and multiple loadings. It was decided that the development of a one factor solution was the most appropriate and conservative option. Using the one factor option, 20 candidate items with sufficient loadings (>.40) on one factor were selected for use in cognitive interviews. Rasch analysis, a type of Item Response Theory (IRT) analysis, revealed that convenience sample respondents tended to agree on the 20 items identified by the EFA (ranging from 65% agreement to 96% agreement). Descriptive analyses were conducted on the 20 items identified by the EFA and IRT analyses, and results from these

analyses are presented in Appendix F. Three items were dropped from further qualitative analyses because more than 90% of respondents agreed when responding to the item. Information about these 17 items will be collected from a sample of the supercoaches through cognitive interviews.

## Rationale

In a previous investigation, EMC explored the following research question: To what extent does a coach's depth of knowledge in two primary domains (coaching knowledge and mathematics content knowledge) influence coaching effectiveness? While the answer was complex, it resulted in a 39 item survey of coaching knowledge grounded in the theory of several prominent researchers (e.g., Knight, 2007; West & Staub, 2003; Costa & Garmston, 2002). Structural Equation Modeling (SEM) results indicated that the relationship between latent coaching effectiveness and the CKS was negative and statistically significant. That is, the CKS measure did not predict coaching effectiveness as expected.  In fact, higher scores on the CKS were related to lower coaching effectiveness. Therefore, the structure of the CKS was explored in some detail to discover the nature of this unpredicted relationship.

In addition to the CKS, a number of other instruments were administered to coaches and the teachers they coached over a four-year period. Teacher survey measures were completed throughout the project. Teachers were also observed by trained data collectors using structured protocols. These measures have all been linked in a longitudinal data set, with teachers nested within coaches, and coaches nested within training cohorts. Coaches were randomly assigned to training cohorts.

As noted, Jesse et al. (2013) and Greenwood (2013) found specific negative relationships between the CKS and other measures of coaching effectiveness. There was a negative and significant relationship between the CKS and a latent coaching effectiveness measure that consisted of coach knowledge of content, coaching behaviors, teacher perceptions, knowledge, beliefs, and teacher behaviors as documented by formal observations using a structured protocol. Greenwood's (Aug 2013 draft) investigation showed a negative relationship between raw and decomposed CKS scores and teacher Mathematical Knowledge for Teaching (MKT) scores, while scores centered at the coach level across time indicted a modest positive relationship. These explorations were only partial considerations and do not address the effects after accounting for other variables in the model. It was further reported that coach-level average CKS values had a negative relationship to teacher MKT responses in the same model, and that higher coach average CKS scores were related to lower teacher MKT scores whereas the time-varying CKS scores are estimated to show increases in the teacher MKT scores. This suggested that increases in CKS over time were related to increases in teacher MKT scores with lower CKS scoring coaches being related to lower MKT scoring teachers.

This paper describes efforts to partially answer components of two primary research questions and several secondary research questions. The primary research questions addressed are:

Toward Validation of the CKS

RQ1: To what extent does a coach's depth of knowledge in two primary domains (coaching knowledge and mathematics content knowledge) influence coaching effectiveness?

RQ3: To what extent are the effects of targeted professional development on coaching effectiveness explained by increases in coaching knowledge and mathematics content knowledge?

Additionally, there was interest in identifying coaches who had teachers with particularly high growth on outcome measures across the project to anchor other efforts to produce validation evidence for the CKS. Specifically, this paper describes efforts to link coaching knowledge to two measures of coaching effectiveness at the teacher outcome level: perceptual teacher survey data and teacher behaviors documented during formal observations conducted by trained observers. Secondary research questions, derived from the primary research questions, include the following:

How are items on the CKS related to teacher outcomes?
Which items on the CKS predict which structured professional development experiences provided by the project coaches have had?
Which items on the original CKS can be used to constitute a one-dimensional scale for measuring coaching effectiveness?
What reliability evidence exists for the proposed revision of the CKS?
What evidence exists that the revised CKS demonstrates predictive, convergent, divergent, and concurrent validity?

## METHODOLOGY

In an effort to establish the validity of the CKS, predictive, convergent, and concurrent validity were considered in some detail. Following is a brief description of each.

***Predictive Validity.*** In theory, the CKS should be a predictor of coaching effectiveness. High correlations between CKS and other measures of coaching effectiveness would provide evidence that the measure has predictive validity. The EMC data set affords a unique opportunity to calculate these correlations with multiple scales, subscales, and individual items measured after initial CKS scores were obtained. Such an analysis framework would identify the ability of the CKS to predict later scores of similar measures of coaching effectiveness.

***Convergent and Divergent Validity.*** Many of the EMC measures have been collected at the same point in time, or relatively close points in time. This affords the opportunity to determine whether and how strong the relationship is between the CKS and other measures of coaching effectiveness at the same time. That is, teacher survey results should be positively correlated to the CKS, and classroom observation results should also be correlated with the CKS if it is a valid measure.  Other measures, which should not be influenced by what is measured by the CKS, should not be correlated with it at all in order to establish convergent and divergent validity.

***Concurrent Validity.*** It is possible to divide coaches into two groups: highly effective coaches, and typical coaches. This division, which was based on perceptual and behavioral data from teachers who were coached, was used to create the supercoach profile. This identification was used to determine whether other measures independent of the CKS would also distinguish supercoaches from other coaches in the study. The CKS was then used to determine whether it predicted membership in a highly-effective coach category. This was accomplished through discriminant analyses. Another form of concurrent validity can be established by determining which items predict professional development (PD) Group membership.

### MEASURES USED

***Coaching Knowledge Survey*** (CKS). To measure coaching knowledge, a 40-item CKS, (later reduced to 39 items) grounded in the theoretical research was created. Two different scoring versions of this survey exist: a 7-point scale version, and a version in which 7-point scale items were converted to a “conforming” metric. A value of “0” meant the item “did not conform” to the coaching literature base, and a “1” indicated that the item response did conform to the literature. A “percent conforming” measure was created for individual conformity of answers to theoretical positions about coaching. This is the measure used in this study.

***Coaching Skills Inventory*** (CSI). The CSI, originally developed by Yopp (2008) and modified for EMC, is intended to measure a coach’s perspective on her or his own level of effectiveness or confidence with various coaching responsibilities. The data produced from the instrument are reliable and valid (Yopp et al., 2010).To measure coaching skills, a 24-item survey using a 5 point scale measures teacher reports of their perceptions about coach/teacher relationships, coaching skills, mathematics content, mathematics-specific pedagogy and general pedagogy. A series of other questions elicit information about educator background and practices, including participation in other mathematics and coaching professional development activities.

***Teacher Survey*** (TS)***.*** To measure teacher attitudes and dispositions around a number of constructs, a teacher survey was implemented to those coached by project participants. While the measure is multidimensional, a TS score was collected across the course of the project.

***Inside the Classroom Observation Protocol*** (ITC-COP). The ITC-COP is a widely used instrument suitable for documenting teacher behaviors in mathematics classrooms. It was used in this study to formally observe teachers each year of the project. Observers were trained and re-established validity of observations through follow-up trainings throughout the course of the project.

***Mathematical Knowledge for Teaching*** (MKT). All participants are asked to complete the MKT Survey of Content Knowledge for Teaching Mathematics (Hill & Ball, 2004). The instrument is designed to assess each teacher's level of mathematical and pedagogical content knowledge. The instrument has been used extensively in research studies and the data produced have been shown to be reliable and valid.

## DATA

Two different sets of data exist for the validation of the CKS: Sample 1 (item development samples); and Sample 2 (study participants). Sample 1 was created before the project began and consists of 252 responses. Sample 2, which consists of the same items given to all coaches over time, utilized almost all of the same questions. Sample 1 was used in CFA and in calculation of reliabilities. Sample 2 was used to calculate relationships with other measures. Sample 1 data were collected from a convenience sample of experienced coaches and coaching experts across the United States before the project began. A total of 252 respondents completed the pilot of the CKS, including 61 EMC project coaches and 191 other participants.

The Sample 2 data collection procedure for this project is complex and proceeded in multiple phases. Exhibit 1 identifies the timeframe for data collection. As noted, pretesting occurred in the winter, spring, and summer of 2010. All of the "A" values represent pretests for teachers and their coaches. The pattern follows with the "B" administrations as the first posttests, the "C" administrations as the second posttests, and the "D" administrations, which have just been completed in the spring of 2013, as the third posttests. Coaches have also completed the "E" administration of the instrumentation.

Sample 2 data were collected in the context of a larger study being conducted by EMC. It consists of two different cohorts of coaches, who experience two different interventions in randomized order. The first major intervention occurred in the summer of 2010 for one of two

PD groups. PD Group 1 was a group of coaches randomly assigned to the first cohort. Similarly, PD Group 2 was randomly assigned to the second cohort. Mathematics content PD was provided to PD Group 1 in the summer of 2010, PD Group 2 received coaching PD in the summer of 2011, Group 1 was trained in coaching PD in the summer of 2012, and to complete the cycle of training for Group 2 they were provided PD in mathematics content in the summer of 2013. It follows, then that using A as pretests, B as posttests for PD Group 1, and C as posttest for PD Group 2 is a reasonable approach to address all primary research questions. Using D as a posttest for both groups follows as a next logical step, as did using E for a posttest for both groups follows as a next logical step to test for PD effects. Complete data sets were obtained from 53 PD Group 1 participants and 334 PD Group 2 participants; although, a number of other coaches participated in the project.

### Exhibit 1. Timeline for EMC Data Collection

| | 2010 | | | | 2011 | | | | 2012 | | | | 2013 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Win | Spr | Sum | Fall | Win | Spr | Sum | Fall | Win | Spr | Sum | Fall | Spr | Sum | Fall |
| *Math Content: PD 1* | | | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 |
| *Coaching: PD 2* | | | | | | | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 |
| *Coaching: PD 1* | | | | | | | | | | | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 |
| *Math Content: PD 2* | | | | | | | | | | | | | | 🟩 | 🟩 |
| Coach MKT | A1 | A2 | | B | | | | C | | | | D | 🟦 | 🟦 | E |
| Coach CKS | | | A | B | | | | C | | | | D | 🟦 | 🟦 | E |
| Coach Coaching Skills Inventory (CSI) | A1 | A2 | | B | | | | C | | | | D | 🟦 | 🟦 | E |
| Intensity | | A | | | | B | | | | C | | | D | | |
| Coach Outside PD | | | | | | B | | | | C | | | D | | |
| Teacher MKT | A1 | A2 | | | B | | | | | C | | | D | | |
| Teacher Survey | A1 | A2 | | | B | | | | | C | | | D | | |
| ITC-COP Observation | | A | | | B | | | | | C | | | D | | |

To simplify an explanation of the analyses that follow, subsets of variables have been identified as predictors and criteria. The CKS items were used to predict teacher survey scale scores; ITC-COP ratings from classroom observations, and professional development group. These predictors and criteria are listed in Exhibit 2.

### EXHIBIT 2. PREDICTORS AND CRITERIA USED TO DEVELOP VALIDITY EVIDENCE FOR THE CKS

| PREDICTORS | CRITERIA |
|---|---|
| **Coach CKS:** | **Teacher Attitudes and Perceptions change (Teacher Survey):** |
| Time 1: Items CKS1aR_A to CKS12R_A | |
| Time 2: Items CKS1aR_B to CKS12R_B | TSTotalA1 |
| Time 3: Items CKS1aR_C to CKS12R_C | TSTotalA2 |
| Time 4: Items CKS1aR_D to CKS12R_D | TSTotalB |
| Time 5: Items CKS1aR_E to CKS12R_E | TSTotalC |
| | TSTotalD |
| | **Coaching Skills Inventor:** |
| | CSITotalA1 |
| | CSITotalA2 |
| | CSITotalB |
| | CSITotalC |
| | CSITotalD |
| | CSItotalE |
| | **Outside Coaching PD** |
| | **Outside Math PD** |
| | **Coach Mathematical Knowledge for Teaching:** |
| | MKTIRTA1 |
| | MKTIRTA2 |
| | MKTIRTB |
| | MKTIRTC |
| | MKTIRTD |
| | MKTIRTE |
| | **Teacher Behavior (ITC-COP Classroom Observations):** |
| | ITCCap7ptA |
| | ITCCap7ptB |
| | ITCCap7ptC |
| | ITCCap7ptD |
| | **Last Known Coach PD:** |
| | 1=PD Group 1 |
| | 2=PD Group 2 |

*Note.* Variable names refer to SPSS file created in the Fall of 2013 from KNOX data.

Inputs on the left of Exhibit 2 are CKS items, and the criteria on the right are outcomes and groups.

## Analyses

Preliminary CFA using the 7-point scale items revealed that the model was not a good fit to the data. This was followed up with a CFA using the polychoric matrix approach for mixed data sets. Model fit indices, indicator loadings, factor correlations, multidimensional scaling, and reliability analyses were then examined to determine which candidate items should be retained after the trimming process.

Utilizing the suggested protocol for describing SEM analyses identified by Brown (2006), Schreiber, et al., (2006) and Cherasaro (2012), The following information was reported: Lisrel Version 8.80 Confirmatory Factor Analysis Results, including Chi Square statistics, Root Mean Square Error of Approximation (RMSEA), the Comparative Fit Index (CFI), the Standardized Root mean Square Residual (SRMR), the Non-Normed fit index (NNFI), Standardized factor loadings, and *t* values and *p* levels. Means, frequencies, standard deviations, scale reliabilities, and intercorrelations of items are also reported in the Appendix.

Since each CKS item was dichotomized to indicate whether coach responses were confirming to the theoretical framework suggested by the literature, a polychoric correlation matrix was calculated using LISREL 8.8 in order to calculate internal reliability. Bonanomi, Ruscone and Osmetti (n.d.) identified a procedure for calculating ordinal alphas from polychoric correlations:

Ordinal Alpha = (k*Mean rpc)/[1+(k-1)* Mean rpc]

## Where k is the number of items, and Mean rpc is the Average of The Polychoric Correlations

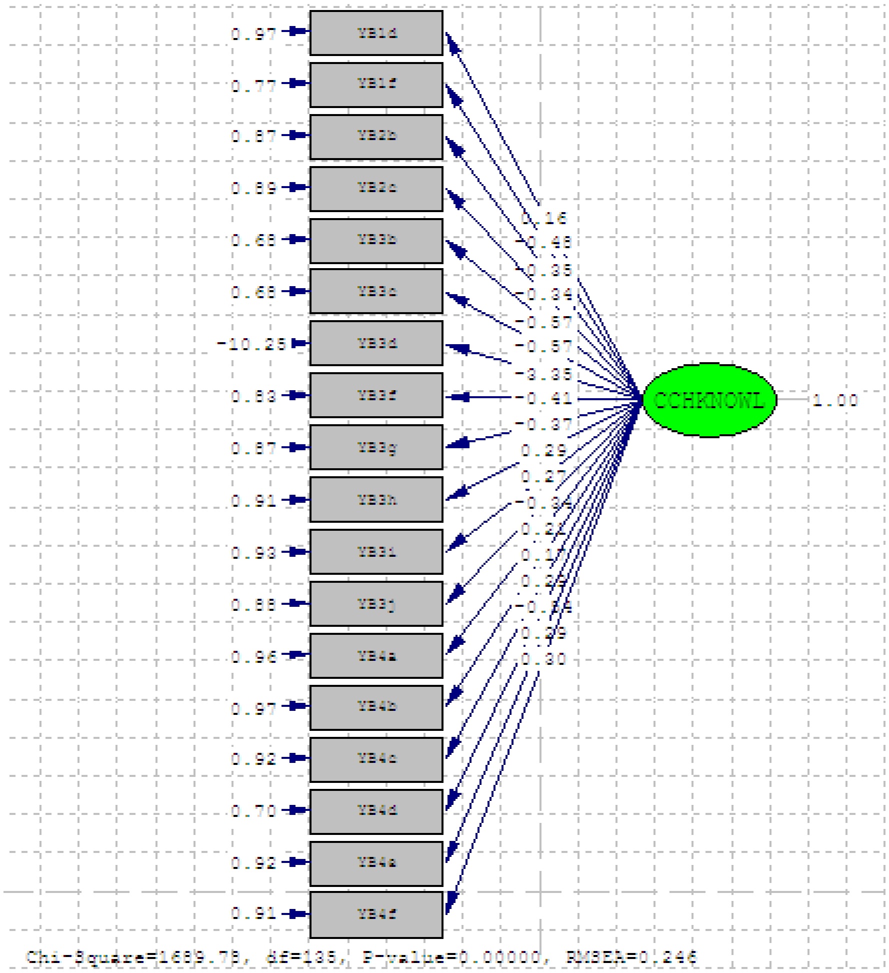
CKS items were correlated with different groups of criteria: TS scale measures, classroom observations conducted by trained observers, coach knowledge, and coach perceptions of skills. Since data were nominal, ordinal, and interval in nature, Spearman correlations were calculated. To be conservative, teacher and coach data were aggregated to the coach level to calculate correlations and to conduct discriminant analyses to determine which items predicted professional development group and supercoach status.

## RESULTS

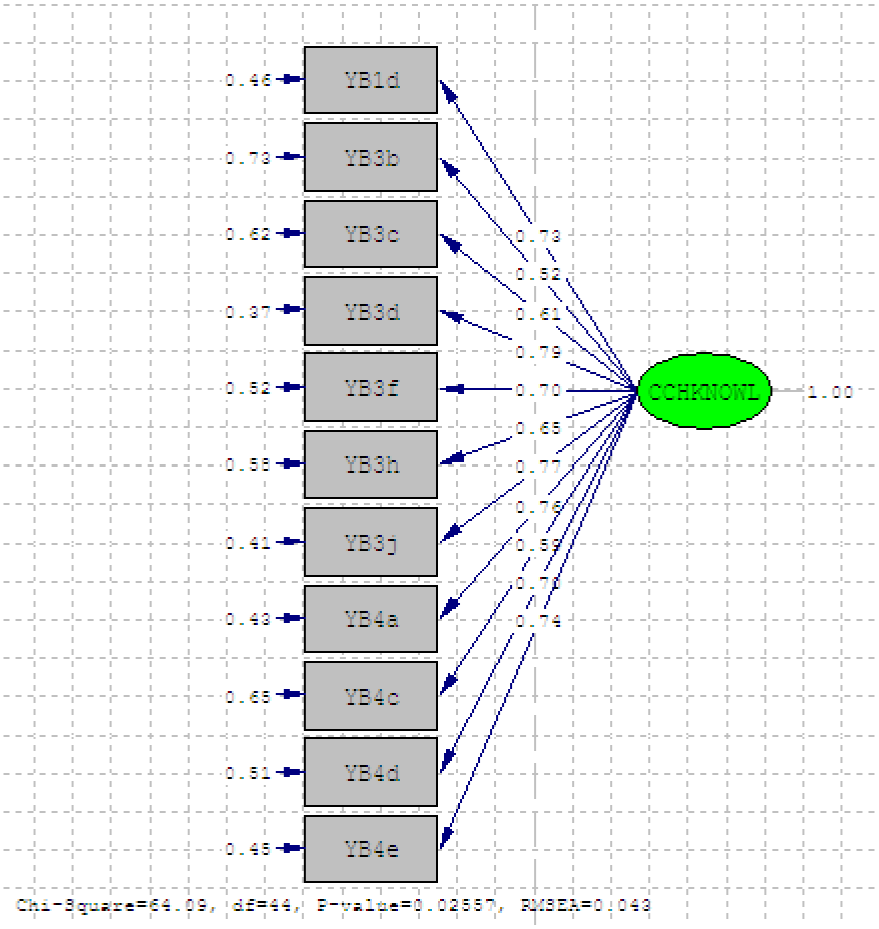
### Analysis of Pilot Data Collected Before the Project

As a first step in validation, a sample of 252 experienced coaches and other educators completed an early version of the CKS before the EMC project began. These data were recoded to reflect the conformity to theory so that each item was scored as a “0” or a “1”. These results were subjected to CFA to further understand the fit of the model with the data collected. Exhibit 3 displays the initial model for conducting CFA. Additional coefficients and descriptive statistics are included in Appendix X. Results demonstrated that the full model with 39 items would not converge using the polychoric matrix, so cluster analysis and reliability coefficient calculations were used to eliminate some items as the next step. Exhibit 4 displays the next model, which utilizes a subset of 18 items. This model was not adequate, as revealed by low loadings, negative loadings, and fit statistics. Exhibit 4, which displays the model constructed by eliminating items based on factor loadings, is a better fit for the pilot data, as evidenced by fit statistics summarized in Exhibit 5.

### EXHIBIT 3. INITIAL MODEL OF COACHING KNOWLEDGE SURVEY TOTAL SCALE (*N* = 252).



### EXHIBIT 4. FINAL MODEL OF COACHING KNOWLEDGE SURVEY TOTAL SCALE (*N* = 252).



### Exhibit 5. Fit Test Statistics for Models Using Pilot Sample Data (N=252).

| Fit Index | Heuristic | Initial Pilot Sample Data Model (18 items) | Final Pilot Selected Data Model (11 items) |
|---|---|---|---|
| Chi-square | Statistical Significance | 1833.89* | 64.09* |
| Root Mean Square Error of Approximation (RMSEA) | Close to .06 or below | 0.197 | 0.043 |
| Comparative Fit Index (CFI) | Close to .95 or greater | 0.922 | 0.927 |
| Standardized Root Mean Square Residual (SRMR) | Close to .08 or below | 0.370 | 0.153 |
| Nonnormed Fit Index (NNFI) | Close to .95 or greater | 0.913 | 0.909 |

*Note.* * p < .05. Heuristics are from Brown, 2006.

## Multidimensional Scaling (MDS)

Multidimensional scaling is a tool that is useful for identifying gaps in survey constructs, because it can provide visual representations of relationships between items. Items that are close together are similar, and items that are far apart are not similar. In a two-dimensional space, a “circle” of items in this case would be expected if the concept was truly captured. MDS was conducted on the pilot data set with the 11 items identified previously by using the ALSCAL routine (a MDS routine) in SPSS version 21. Distances were created from the data via the binary Euclidian option, and the level of measurement was specified as ordinal. The scree plot in Exhibit 6 displays stress values from 1 to 5 dimensions, and suggests that a 2-dimensional solution is best.

Exhibit 6. Scree Plot of Stress by Dimensions from 11 item MDS Solutions (N=252)

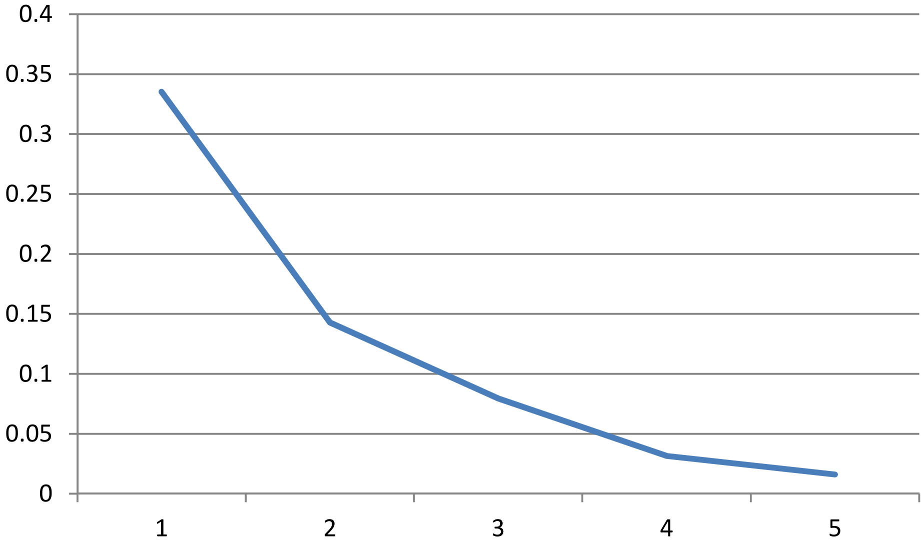
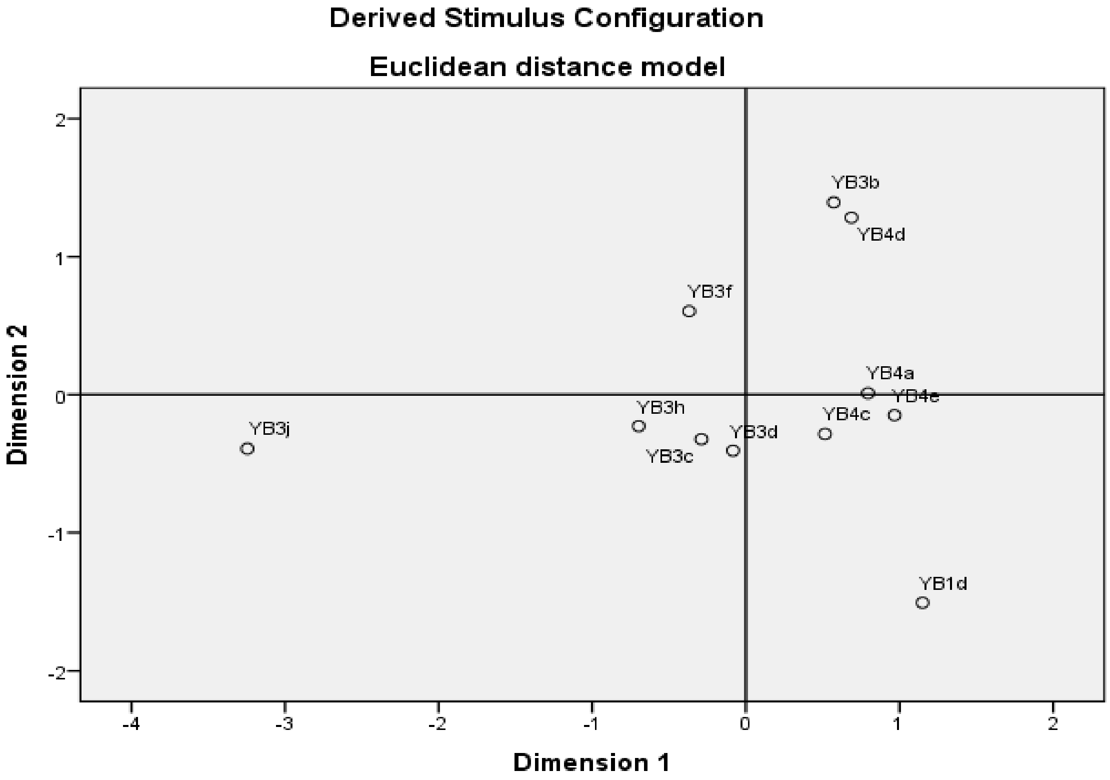


Exhibit 7 displays preliminary MDS results from the pilot test with the 11 items that have been selected through the preliminary CFA and alpha coefficient analyses. Results are displayed in two-dimensional space, and suggest that some gaps may exist if we think of the items as constituting a “circle”. For example, it may be interesting to explore the development of new items in the top left quadrant, which contains only question 3f, “I have difficult conversations with teachers, when necessary, about mathematics misconceptions they hold”.

Item 3j is by itself in the lower left quadrant, suggesting that it is measuring something that is not connected to the other items. Item 1j, “I provide feedback to teachers about whether or not the school is meeting its vision for mathematics instruction”, is the only one on the list that speaks to feedback to teachers about vision. Item 1d, “sometimes an effective mathematics coach has to oppose school or teacher actions that are not good for students’ mathematics learning may measure a strong position taken by coaches that may not be well-understood”, is by itself in the lower right quadrant.

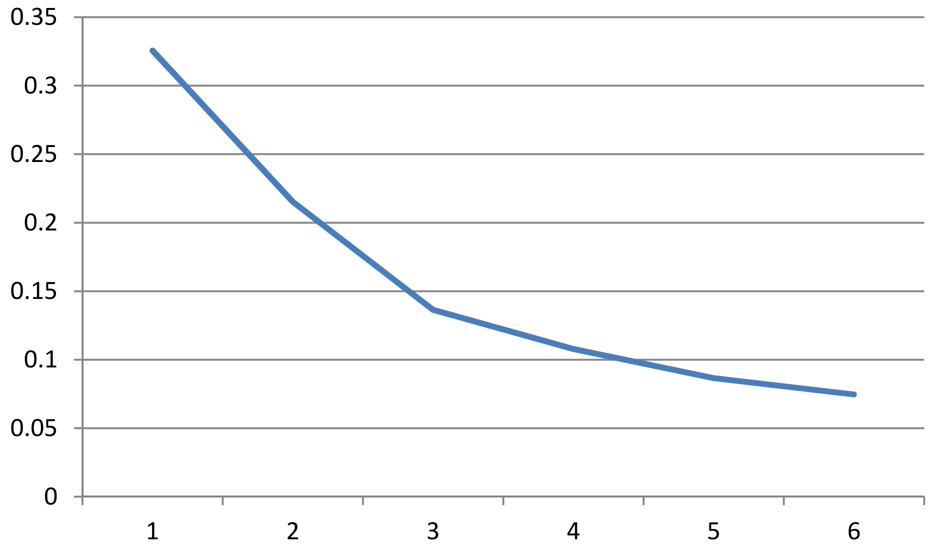
### Exhibit 7. MDS Solution for the 11 Item CKS Displayed in 2 Dimensions (N=252)



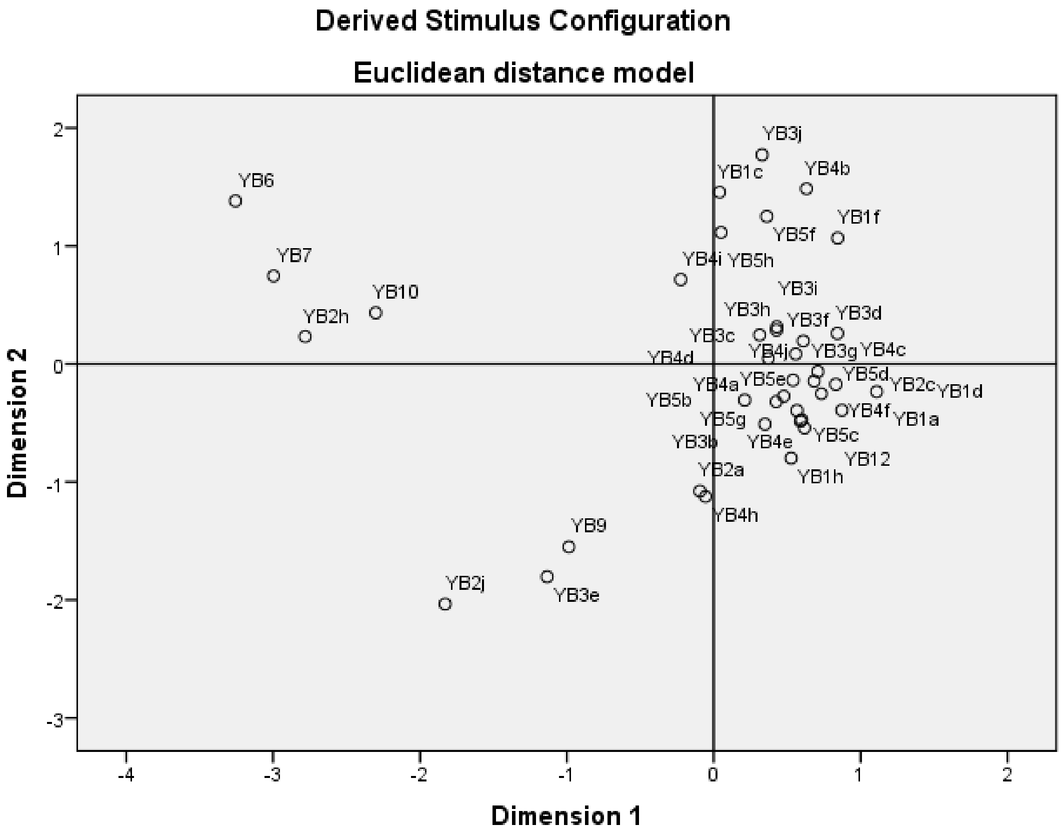
For the sake of comparison, a scree plot for the 39 item scale and a MDS mapping of results is displayed in Exhibits 8 and 9. Stress values across 5 dimensions were better for the 11 item solution (ranging from .34 to .02) than for the 39 item solution (.33 to .09)1. The scree plot in Exhibit 8 suggests that a 3 dimensional solution is best.

[1] In general, the smaller the stress is, the better the fit. Stress greater than .20 is poor, .10 fair, .05 good. .025 excellent, and .00 perfect (Wickelmaier, 2003).

### Exhibit 8. Scree Plot of Stress by Dimensions from 39 item MDS Solutions (N=252)



### Exhibit 9. MDS Solution for the 39 Item CKS Displayed in 2 Dimensions (N=252)



Finally, the polychoric correlation matrix technique was used to calculate an ordinal alpha or reliability coefficient for the 11 items used in this analysis. The ordinal reliability coefficient for these 11 items is .888, which is in the acceptable range for a scale used in research.

## Analysis of Project Data for 5 Years

Three different strategies were used to provide additional validity evidence through the use of project data collected longitudinally. What was learned from the analyses of the CKS pilot data served to guide analyses of project data collected over a five year time span. As a first step, CKS item Spearman correlations were calculated between items and selected teacher outcomes to provide evidence for predictive and convergent and divergent validity. Then, all CKS items were tested to determine whether they could discriminate between coaches trained in Cohort 1 and coaches trained in Cohort 2 to produce concurrent validity evidence. Items that predicted PD Group membership were also identified.

Additionally, a subset of supercoaches was identified statistically. Six coaches (three from Cohort 1, three from Cohort 2) who had at least two teachers improve on a combined measure of TS responses and classroom observation data from the first pretest to the final posttest. These growth scores, aggregated at the coach level, were rank-ordered and averaged to create a rank ordering of coaches. The following steps were taken to create the “supercoach” subset:

1. We calculated the growth between Time A and Time D on the TS for each teacher.
2. We calculated the growth between ITC-COP Capstone 7-point ratings at Time A and Time D for each teacher.
3. We calculated the mean or average growth on the TS and the ITC-COP measures for each coach.
4. We eliminated any coach or teacher who was not consistently paired from the first pretest to the last posttest.
5. We eliminated any coach who did not have data for more than one teacher. All coaches remaining in the sample had pretest and final posttest data for 2 or 3 teachers.
6. We averaged the ranks of TS growth and the ITC-COP growth.
7. We sorted the file by these ranks. There was a natural break in the data for the first six coaches. They were coded as a “1“, other coaches were coded as a “0”.
8. We identified the items that three or more of the supercoaches got “wrong”(i.e. nonconforming to the literature on the CKS).

Displayed in Exhibit 10 is a listing of the items that clarify the distinction between supercoaches and other coaches on the CKS, using data aggregated to the coach level on the last administration of the CKS (Time E).

### Exhibit 10. 2013 CKS Items That Distinguish Supercoaches from Typical Coaches (*N*=50)

| Item | Percent Conforming | Other Coach Average (N=44) | Supercoach Average (N=6) |
|---|---|---|---|
| 1c. When a teacher says that she or he doesn't want any coaching, an effective mathematics coach respectfully does not try to persuade the teacher to accept coaching. | 42 | 43 | 33 |
| 1f. An effective mathematics coach gets input from a school's principal on which teachers need to improve their mathematics instruction. | 36 | 39 | 17 |
| 2h. An effective coach sticks to the coaching objectives established with a teacher at the beginning of the school year. | 50 | 52 | 33 |
| 3d. I coach teachers on needs that I observe in the teacher, even when the teacher is unaware of these needs. | 44 | 45 | 33 |
| 3h. I meet with the principal to discuss the school's vision for mathematics instruction. | 46 | 48 | 33 |
| 3j. I provide feedback to teachers about whether or not the school is meeting its vision for mathematics instruction. | 36 | 36 | 33 |
| 4b. I ask the principal what he or she believes the mathematics teachers' needs are. | 60 | 64 | 33 |
| 5f. I provide feedback to the principal about whether or not the school is meeting its vision for mathematics instruction. | 68 | 70 | 50 |
| 5h. When a teacher complains about the school's vision for mathematics, I ask the teacher about her or his vision for mathematics. | 40 | 39 | 50 |
| 6r. Which is the most powerful response to help the teacher take ownership of developing a personal knowledge base? (multiple choice) | 74 | 77 | 50 |
| 9r. "I think the teacher before me didn't teach subtraction very well." What should the coach do next? (multiple choice) | 52 | 52 | 50 |

A series of Spearman correlations were calculated with each of the 39 items on the CKS and selected teacher level outcomes, namely the TS total score and ITC-COP capsule ratings from classroom visits made by trained observers using a 7-point rating scale. These data points have been collected longitudinally across the project, and all correlations were calculated at each point in time. Exhibits A1 through A12 in the Appendix display these correlations, and reveal that some items are consistently correlated with teacher survey and observation outcomes. Specifically, CKS items 2h, 4h, 5c, 6, and 10 are correlated with more than one teacher outcome. Exhibit 11 summarizes these detailed results displayed in the Appendices in tabular form.

### Exhibit 11. CKS Items with Multiple Statistically Significant Correlations with Teacher Survey and ITC-COP Ratings Across Time

| Item | Number of Statistically Significant Spearman Correlations |
|---|---|
| 2h. An effective coach sticks to the coaching objectives established with a teacher at the beginning of the year. | 4 |
| 4h. I do not alter the coaching plan developed with the teacher at the beginning of the school year. | 4 |
| 5c. I take precautions to ensure that my demonstration lessons do not inadvertently send a message that I am the expert and the teacher is not. | 3 |
| 6. Base 10 Coach Scenario (multiple choice) | 3 |
| 10. Teaching Strategy Discussion Scenario (multiple choice) | 6 |

Similarly, Spearman correlations were calculated between CKS items and Coach MKT IRT scores and Coach CSI total scale scores. Exhibit 12 lists the items that were correlated with 10 or more coach measures (see Appendix).

### Exhibit 12. CKS Items with Multiple Statistically Significant Correlationswith Mathematical Knowledge for Teaching Coaching Skills Inventory Survey Scores Across Time

| Item | Number of Statistically Significant Spearman Correlations |
|---|---|
| 3c. When decisions about mathematics instruction are being made, I ensure that the decision-makers interpret research literature accurately. | 14 |
| 3f. I have difficult conversations with teachers, when necessary, about mathematics misconceptions they hold. | 15 |
| 3i. I encourage teachers to include, in each lesson they teach, summaries of what students learned. | 16 |
| 4i. I help teachers identify consistencies and inconsistencies between their won practices and the practices recommended by the National Council of Teachers of Mathematics. | 18 |
| 5b. I help teachers reflect on discrepancies between espoused beliefs and actual practices. | 17 |
| 5h. When a teacher complains about the school's vision for mathematics, I ask the teacher about her or his vision for mathematics. | 16 |

## DISCRIMINANT ANALYSIS OF CKS ITEMS TO PREDICT PD GROUP

Only a small number of CKS items distinguished PD Group 1 participants from PD Group 2 participants. Results are displayed in Appendix C graphically, with significant differences identified by the z test for two proportions.

In order to make a more rigorous determination about which CKS items significantly discriminated between training groups (PD Group 1 versus PD Group 2), five separate discriminant analyses using all CKS items was conducted to predict group membership. A conservative approach was taken to the analyses by using data aggregated to the coach level. While jackknifed predictions ranged from 48.9% to 59.6% (little better than chance), some items emerged as important predictors. Results are summarized in Exhibit 13 for brevity.

## EXHIBIT 13. CKS ITEMS THAT DISCRIMINATE BETWEEN PROFESSIONAL DEVELOPMENT COHORTS

| | Survey Administration | | | | |
|---|---|---|---|---|---|
| | Time A | Time B | Time C | Time D | Time E |
| | Spring | Fall | Fall | Fall | Fall |
| Item | 2010 | 2010 | 2011 | 2012 | 2013 |
| 1a. An effective mathematics coach coaches only on teacher-stated needs. | | | x | | x |
| 1b. Beginning teachers need more coaching than 25-year veterans. | | | x | | |
| 1c. When a teacher says that she or he doesn't want any coaching, an effective mathematics coach respectfully does not try to persuade the teacher to accept coaching. | x | x | | | |
| 1d. Sometimes an effective mathematics coach has to oppose school or teacher actions that are not good for students' mathematics learning. | x | | | | |
| 1h. A coach should put no pressure on teachers to improve their practices. | | x | | | |
| 3d. I coach teachers on needs that I observe in the teacher, even when the teacher is unaware of these needs. | x | | x | | x |
| 4f. I try to help teachers understand my role as a mathematics coach. | | | x | | |
| 4i. I help teachers identify consistencies and inconsistencies between their own practices and the practices recommended by the National Council of Teachers of Mathematics. | | | x | | |

Note. Data aggregated to the coach level. Time A = Baseline, Time B = PD1 content, Time C = PD2 coaching, Time D = PD1 Coaching, and Time E = PD2 Content.

## IDENTIFICATION OF A CANDIDATE SUBSET OF POTENTIAL ITEMS TO VALIDATE IN COGNITIVE INTERVIEWS

Based upon previous analyses, a set of items that constitute candidates for further scale development has been identified. Exhibit 14 displays these items which were retained for further study and validation.

### Exhibit 14. Candidate CKS Items for Validation Identified with Multiple Techniques

| Item |
|---|
| 1a. An effective mathematics coach coaches only on teacher-stated needs. |
| 1b. Beginning teachers need more coaching than 25-year veterans. |
| 1c. When a teacher says that she or he doesn't want any coaching, an effective mathematics coach respectfully does not try to persuade the teacher to accept coaching. |
| 1d. Sometimes an effective mathematics coach has to oppose school or teacher actions that are not good for students' mathematics learning. |
| 1h. A coach should put no pressure on teachers to improve their practices. |
| 2h. An effective coach sticks to the coaching objectives established with a teacher at the beginning of the year. |
| 3b. I collect students' mathematics work from a teacher's classroom to guide our coaching conversations. |
| 3c. When decisions about mathematics instruction are being made, I ensure that the decision-makers interpret research literature accurately. |
| 3d. I coach teachers on needs that I observe in the teacher, even when the teacher is unaware of these needs. |
| 3f. I have difficult conversations with teachers, when necessary, about mathematics misconceptions they hold. |
| 3h. I meet with the principal to discuss the school's vision for mathematics instruction. |
| 3i. I encourage teachers to include, in each lesson they teach, summaries of what students learned. |
| 3j. I provide feedback to teachers about whether or not the school is meeting its vision for mathematics instruction. |
| 4a. I try to provide the teachers I coach with an understanding of how the mathematics they teach supports learning beyond the grade level they teach. |
| 4c. I encourage the teachers I coach to reflect on similarities and differences among mathematics topics in the curriculum. |
| 4d. I help teachers plan their lessons. |
| 4e. I ask the teachers I coach what aspects of mathematics teaching they need help with |
| 4f. I try to help teachers understand my role as a mathematics coach. |
| 4h. I do not alter the coaching plan developed with the teacher at the beginning of the school year. |
| 4i. I help teachers identify consistencies and inconsistencies between their own practices and the practices recommended by the National Council of Teachers of Mathematics. |
| 5b. I help teachers reflect on discrepancies between espoused beliefs and actual practices. |
| 5c. I take precautions to ensure that my demonstration lessons do not inadvertently send a message that I am the expert and the teacher is not |
| 6. Base 10 Coach Scenario (multiple choice) |
| 10. Teaching Strategy Discussion Scenario (multiple choice) |

## Next Steps

A convenience sample of coaches will be interviewed to learn more about how they answered the items the way they did, when they did. Data from these interviews will be coded using methodologies explicated by Miles, Huberman, and Saldana (2014), and used to make determinations about whether respondents understood the directions, and answered the questions as EMC researchers intended. This information can then be used to understand results obtained, and make recommendations for future survey modifications.

The following items have been identified as items to use with a subset of coaches to gain insights into the thinking behind why they responded the way they did. Using EFA and IRT, 20 items were identified as the items of interest. Of the 20 items identified, three were dropped (5e, 5d, and 4a) from the interview protocol because nearly every respondent answered them the same way, leaving 17 items.

Of the remaining 17 items, those items were then compared to the items identified through the analyses of supercoaches. Six items (3d, 3h, 3j, 4b, 5f, and 5h) were identified by the EFA and IRT analyses as well as the supercoach analyses. All six of these items will be included in the interview protocol. Of the remaining 11 items identified by the EFA and IRT, three items (4i, 3b, and 4d) were identified as the “most difficult” and will be included in the interview protocol.

The eight remaining items identified by the EFA and IRT analyses, will also be included in the interview protocol. In the event that time does permit all 17 items to be addressed, the six items identified by the EFA and IRT analyses as well as the supercoach analyses will be prioritized, followed by the three items identified as the “most difficult” by the IRT, the three items (5g, 4c, and 4j) identified as the “easiest” by the IRT, and finally the remaining items (in ranking order of difficulty, as identified by the IRT. Three items deemed the “easiest”, followed by the remaining items (3i, 3c, 3g, 5b, and 3f).

The following items will be prioritized during the cognitive interviews and asked first:

1. 3d
2. 3h
3. 3j
4. 4b
5. 5f
6.5h
7. 4i
8. 3b
9. 4d

If time permits, the following items will be asked, in the order they appear in this list:

10. 5g
11. 4c
12. 4j
13. 3i
14. 3c
15. 3g
16. 5b
17. 3f

In order to complete the survey revision cycle, it is suggested that EMC provide coaches a copy of the survey and ask them to answer the following questions *after* they have completed the last study (after Cherasaro, 2012):

- What problems, if any, did you have completing the survey?
- Are the directions clear?  If not, why not?
- Are there any words or language in the survey that coaches might not understand? Please explain.
- Did you find any of the questions redundant or unnecessary?  If so, which ones?  Why?
- Were any of the questions difficult to answer?  If so, why?
- (item by item, or selected/balanced items, 39 items) What did you think this question was asking? How would you phrase it in your own words?
- Do the answer choices allow you to answer as you intended?  Please explain.
- Is there anything you would change about the instrument?  Please explain.

## REFERENCES

Alteras, V.H., Kostarelis, A., Tsitouridou, M., Niakas, D., & Nicolau, A. (2010). Development and preliminary validation of a questionnaire to measure satisfaction with home care in Greece: An exploratory factor analysis of polychoric correlations. BMC Health Services Research, 10: 189. Retrieved from: http://www.biomedcentral.com/1472-6963/10/189

Basto, M., & Pereia, J. M. (2012). An SPSS R-menu for ordinal factor analysis. *Journal of Statistical Software, 46*(4). Retrieved from: http://www.jstatsoft.org/

Bonanomi, A., Ruscone, M. N., & Osmetti, S. A. (n.d.). *The polychoric ordinal alpha, measuring the reliability of a set of polytomous ordinal items.*

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

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

Byrne, B. M. (1998). *Structural Equation Modeling with LISREL, PRELIS and SIMPLIS: Basic concepts, applications, and programming.* Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Cherasaro, T. (2012). *Examining evaluator feedback in teacher evaluation systems.* Unpublished manuscript, limited availability. Englewood, CO: Marzano Research Laboratory.

Cordray, D. S. & Pion, G. M. (2006). Treatment strength and integrity: Models and methods. In R. R. Bootzin & P. E., McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation (pp 103-124).* Washington, DC: American Psychological Association.

Costa, A. L., & Garmston, R. J. (2002). *Cognitive coaching: A foundation for Renaissance schools*. Norwood, MA: Christopher-Gordon Publishers.

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical and practice guide. *Practical Assessment, Research & Evaluation, 17*(3). Retrieved from: http://pareonline.net/getvn.asp?v=17&n=3

Greenwood, M., (2013).

Hill, H.C., Schilling, S.G., & Ball, D.L. (2004) Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal*, 105, 11-30.

Hulleman, C. S., & Cordray, D. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Intervention Effectiveness*, *2*(1), 88-110.

Jesse, D., et al (2013)

Knight, Knight, J. (2007b). *Instructional coaching: A partnership approach to improving instruction.* Thousand Oaks, CA: Corwin Press.

Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling*. (3rd ed.). New York: The Guilford Press.

MacCallum, R.C., Browne, M. W., & Sugaware, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130-149.

Miles, M.B., Huberman, A. M., and Saldana, J. (2014). Qualitative Data Analysis: A methods sourcebook. 3rd edition. Thousand Oaks, CA: Sage

Raudenbush, S. W., et al. (2011). *Optimal Design Software for Multi-level and Longitudinal Research* (Version 3.01) [Software]. Retrieved from: www.wtgrantfoundation.org.

Schreiber, et al Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A. and King, J. Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review**.** *Journal of Educational Research, 99*( 6), 323-337.

West, L., & Staub, F. C. (2003). *Content-focused coaching*. Pittsburgh,

Wickelmaier, F. (2003). An introduction to MDS. Aalforg, East, Denmark: Aalborg University. Retrieved from: http://homepages.uni-tuebingen.de/florian.wickelmaier/pubs/Wickelmaier2003SQRU.pdf

Yopp, D. (2008). Yopp, D.A. (2008, August). *Strengthening strands: Improving mathematics content and pedagogy in middle school teachers for Minidoka and Cassia Counties evaluation report.* (Unpublished Study)

Yopp, D, et al (2010). Yopp, D. A., Rose, J., and Meade, C. (2008). Validity and reliability estimates of a set of mathematics classroom coaching reflection instruments. *Education Evaluation and Policy*, submitted.

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. Journal of Modern Applied Statistical Methods, 6, 21-29. Retrieved from: http://educ.ubc.ca/faculty/zumbo/papers/ordinal_alpha_reprint.pdf

## Appendices

### A. CKS Spearman Correlations with Teacher Survey and ITC COP Observation Results

Exhibit A1: Year 1 CKS Items and Spearman Correlations with Teacher Survey and ITC COP Results across Time

Exhibit A2: Year 2 CKS Items and Spearman Correlations with Teacher Survey and ITC COP Results across Time

Exhibit A3: Year 3 CKS Items and Spearman Correlations with Teacher Survey and ITC COP Results across Time

Exhibit A4: Year 4 CKS Items and Spearman Correlations with Teacher Survey and ITC COP Results across Time

Exhibit A5: Year 5 CKS Items and Spearman Correlations with Teacher Survey and ITC COP Results across Time

### B. CKS Spearman Correlations with Coach MKT and CSI Results

Exhibit B1: Year 1 CKS Items and Spearman Correlations with Coach MKT and CSI Results across Time)

Exhibit B2: Year 2 CKS Items and Spearman Correlations with Coach MKT and CSI Results across Time

Exhibit B3: Year 3 CKS Items and Spearman Correlations with Coach MKT and CSI Results across Time

Exhibit B4: Year 4 CKS Items and Spearman Correlations with Coach MKT and CSI Results across Time

Exhibit B5: Year 5 CKS Items and Spearman Correlations with Coach MKT and CSI Results across Time

C. Descriptive Statistics for Items by PD Group

D. Original CKS survey

E. CKS Scoring using Item Response Theory methods

F. Descriptives on the 20 CKS Items Retained

## Exhibit A1: CKS Time 1 Item Correlations with Teacher Outcomes across Time (N = 41)

| | TSTotalA1_mean | TSTotalA2_mean | TSTotalB_mean | TSTotalC_mean | TSTotalD_mean | ITCCap7ptA_mean | ITCCap7ptB_mean | ITCCap7ptC_mean | ITCCap7ptD_mean |
|---|---|---|---|---|---|---|---|---|---|
| An effective mathematics coach coaches only on teacher-stated needs. | -.177 | -.213 | -.390* | -.241 | -.193 | -.086 | -.033 | -.209 | -.234 |
| | .268 | .181 | .012 | .129 | .226 | .591 | .839 | .190 | .142 |
| When a teacher says that she or he doesn't want any coaching, an effective mathematics coach respectfully does not try to persuade the teacher to accept coaching. | -.176 | -.163 | -.054 | -.123 | -.271 | -.103 | .122 | .218 | .094 |
| | .271 | .309 | .737 | .444 | .086 | .522 | .446 | .171 | .558 |
| Sometimes an effective mathematics coach has to oppose school or teacher actions that are not good for students' mathematics learning. | .013 | .009 | -.004 | -.013 | -.072 | .060 | .256 | .048 | .298 |
| | .937 | .955 | .980 | .937 | .657 | .712 | .106 | .765 | .059 |
| An effective mathematics coach gets input from a school's principal on which teachers need to improve their mathematics instruction. | -.134 | -.095 | -.039 | .006 | -.075 | -.024 | .061 | .124 | .064 |
| | .405 | .556 | .809 | .972 | .642 | .881 | .706 | .441 | .693 |
| A coach should put no pressure on teachers to improve their practices. | -.039 | -.067 | -.079 | .002 | -.133 | -.074 | -.010 | -.079 | .000 |
| | .808 | .675 | .623 | .991 | .405 | .645 | .950 | .622 | .999 |
| Once a teacher knows about a research-based strategy for improving student learning, the teacher will begin using the strategy. | -.032 | -.132 | .064 | .055 | -.093 | -.013 | -.099 | -.110 | -.067 |
| | .841 | .409 | .690 | .730 | .562 | .934 | .537 | .495 | .675 |

|  | TSTotalA1 _mean | TSTotalA2 _mean | TSTotalB_ mean | TSTotalC_ mean | TSTotalD_ mean | ITCCap7pt A_mean | ITCCap7pt B_mean | ITCCap7pt C_mean | ITCCap7pt D_mean |
|---|---|---|---|---|---|---|---|---|---|
| An effective mathematics coach provides teachers with an understanding of how the mathematics they teach supports learning beyond the grade level they teach. |  |  |  |  |  |  |  |  |  |
| An effective mathematics coach uses state mathematics assessment data when developing a coaching plan with teachers. | .033 | .059 | .026 | .002 | -.085 | .197 | .084 | -.067 | .058 |
|  | .839 | .715 | .871 | .988 | .598 | .218 | .602 | .679 | .721 |
| An effective coach sticks to the coaching objectives established with a teacher at the beginning of the school year. | .374* | .296 | .365* | .327* | .181 | .291 | .077 | -.021 | -.026 |
|  | .016 | .061 | .019 | .037 | .256 | .064 | .634 | .898 | .874 |
| An effective mathematics coach gives feedback to the principal about teachers who are struggling in the classroom. | .229 | .240 | -.072 | -.133 | -.036 | .016 | -.120 | -.053 | .049 |
|  | .149 | .130 | .652 | .406 | .825 | .920 | .455 | .743 | .759 |
| I collect students' mathematics work from a teacher's classroom to guide our coaching conversations. | .031 | -.011 | -.062 | -.182 | -.030 | .322* | -.035 | -.204 | -.059 |
|  | .848 | .947 | .698 | .255 | .854 | .040 | .827 | .200 | .713 |
| When decisions about mathematics instruction are being made, I ensure that the decision-makers interpret research literature accurately. | .202 | .041 | .151 | .146 | .075 | .300 | .049 | -.149 | .010 |
|  | .206 | .799 | .346 | .362 | .640 | .057 | .760 | .352 | .949 |

| | TSTotalA1<br>_mean | TSTotalA2<br>_mean | TSTotalB_<br>mean | TSTotalC_<br>mean | TSTotalD_<br>mean | ITCCap7pt<br>A_mean | ITCCap7pt<br>B_mean | ITCCap7pt<br>C_mean | ITCCap7pt<br>D_mean |
|---|---|---|---|---|---|---|---|---|---|
| I coach teachers on needs that I observe in the teacher, even when the teacher is unaware of these needs. | .045 | .086 | -.082 | -.118 | -.186 | .070 | .028 | .091 | -.025 |
| | .781 | .593 | .611 | .462 | .243 | .664 | .864 | .570 | .877 |
| As a mathematics coach, I support mathematics teachers by tutoring their struggling students. | -.083 | -.044 | -.224 | -.110 | -.176 | .022 | .035 | .071 | .187 |
| | .606 | .783 | .159 | .492 | .271 | .892 | .827 | .660 | .243 |
| I have difficult conversations with teachers, when necessary, about mathematics misconceptions they hold. | .218 | .100 | .101 | .111 | .260 | .255 | .013 | -.214 | -.003 |
| | .170 | .535 | .531 | .491 | .101 | .107 | .935 | .178 | .986 |
| I always make sure that coaching conversations with mathematics teachers are grounded in the mathematics content. | -.027 | -.074 | -.060 | .093 | -.063 | -.040 | -.165 | .179 | .034 |
| | .867 | .646 | .709 | .562 | .697 | .802 | .304 | .263 | .835 |
| I meet with the principal to discuss the school's vision for mathematics instruction. | -.189 | -.192 | -.122 | .027 | -.139 | -.143 | -.233 | -.330* | -.337* |
| | .237 | .230 | .447 | .867 | .387 | .374 | .143 | .035 | .031 |
| I encourage teachers to include, in each lesson they teach, summaries of what students learned. | .022 | .078 | .026 | .153 | .013 | .267 | .164 | .118 | -.149 |
| | .891 | .626 | .870 | .341 | .933 | .091 | .305 | .461 | .351 |
| I provide feedback to teachers about whether or not the school is meeting its vision for mathematics instruction. | -.165 | -.296 | -.043 | -.059 | -.117 | -.210 | -.127 | -.192 | -.407** |
| | .303 | .061 | .788 | .713 | .468 | .188 | .427 | .230 | .008 |

| | TSTotalA1_mean | TSTotalA2_mean | TSTotalB_mean | TSTotalC_mean | TSTotalD_mean | ITCCap7ptA_mean | ITCCap7ptB_mean | ITCCap7ptC_mean | ITCCap7ptD_mean |
|---|---|---|---|---|---|---|---|---|---|
| I try to provide the teachers I coach with an understanding of how the mathematics they teach supports learning beyond the grade level they teach. | -.044 | -.110 | -.025 | -.028 | -.025 | .220 | .034 | -.064 | -.035 |
| | .783 | .492 | .878 | .860 | .878 | .167 | .831 | .691 | .830 |
| I ask the principal what he or she believes the mathematics teachers' needs are. | .011 | .041 | -.085 | .033 | -.039 | .090 | .159 | .045 | -.262 |
| | .943 | .801 | .595 | .835 | .811 | .575 | .320 | .779 | .098 |
| I encourage the teachers I coach to reflect on similarities and differences among mathematics topics in the curriculum. | -.043 | -.148 | -.032 | -.047 | .024 | .100 | .012 | .078 | .002 |
| | .788 | .357 | .842 | .769 | .884 | .534 | .942 | .627 | .988 |
| I help teachers plan their lessons. | .015 | .015 | -.046 | .038 | -.014 | .238 | .059 | -.032 | .105 |
| | .928 | .928 | .777 | .815 | .929 | .134 | .713 | .841 | .512 |
| I ask the teachers I coach what aspects of mathematics teaching they need help with. | .254 | .241 | .241 | .254 | .267 | .095 | -.208 | -.020 | .067 |
| | .109 | .130 | .130 | .109 | .091 | .556 | .192 | .901 | .676 |
| I try to help teachers understand my role as mathematics coach. | .205 | .082 | .178 | .190 | .214 | -.132 | -.130 | -.288 | -.121 |
| | .198 | .611 | .266 | .235 | .179 | .410 | .419 | .067 | .451 |
| I do not alter the coaching plan developed with the teacher at the beginning of the school year. | .216 | .226 | .195 | .084 | .024 | .009 | .033 | .094 | .061 |
| | .175 | .155 | .222 | .599 | .884 | .953 | .838 | .560 | .705 |
| I help teachers identify consistencies and inconsistencies between their own practices and the practices recommended by the National Council of Teachers of Mathematics. | .255 | .151 | .122 | .004 | .139 | .382* | .119 | -.249 | .135 |
| | .108 | .346 | .446 | .980 | .387 | .014 | .457 | .116 | .399 |

|  | TSTotalA1_mean | TSTotalA2_mean | TSTotalB_mean | TSTotalC_mean | TSTotalD_mean | ITCCap7ptA_mean | ITCCap7ptB_mean | ITCCap7ptC_mean | ITCCap7ptD_mean |
|---|---|---|---|---|---|---|---|---|---|
| I work with principals or other administrators to form a clear message to teachers about effective mathematics instruction. | -.009 | -.129 | -.261 | -.375* | -.285 | -.019 | -.005 | -.163 | -.096 |
|  | .956 | .423 | .099 | .016 | .071 | .905 | .973 | .308 | .552 |
| When a teacher says something I find confusing, I say, "That confused me," and ask the teacher to rethink it. | -.082 | -.141 | -.193 | -.170 | -.143 | -.086 | -.083 | -.173 | -.172 |
|  | .610 | .378 | .227 | .288 | .373 | .591 | .604 | .279 | .283 |
| I take precautions to ensure that my demonstration lessons do not inadvertently send a message that I am the expert and the teacher is not. | .095 | .044 | .178 | .263 | .253 | .064 | -.180 | .000 | -.125 |
|  | .554 | .783 | .266 | .097 | .110 | .689 | .260 | .998 | .437 |
| I reflect on state assessment data to identify curriculum areas that need to be strengthened. | -.141 | -.062 | -.114 | -.073 | .061 | -.087 | .127 | -.168 | -.072 |
|  | .378 | .698 | .477 | .650 | .707 | .587 | .429 | .294 | .655 |
| I use student work when coaching mathematics teachers. | .056 | -.052 | .036 | -.053 | -.038 | .052 | -.021 | .021 | -.035 |
|  | .730 | .747 | .824 | .741 | .812 | .746 | .896 | .897 | .827 |
| I provide feedback to the principal about whether or not the school is meeting its vision for mathematics instruction. | -.163 | -.165 | -.018 | -.018 | -.152 | .003 | .022 | -.206 | -.142 |
|  | .308 | .302 | .910 | .913 | .342 | .984 | .893 | .197 | .376 |
| I encourage teachers to set personal improvement goals for mathematics instruction. | -.068 | -.223 | -.098 | -.098 | -.008 | .041 | -.113 | -.218 | -.138 |
|  | .670 | .162 | .541 | .543 | .962 | .798 | .482 | .170 | .390 |

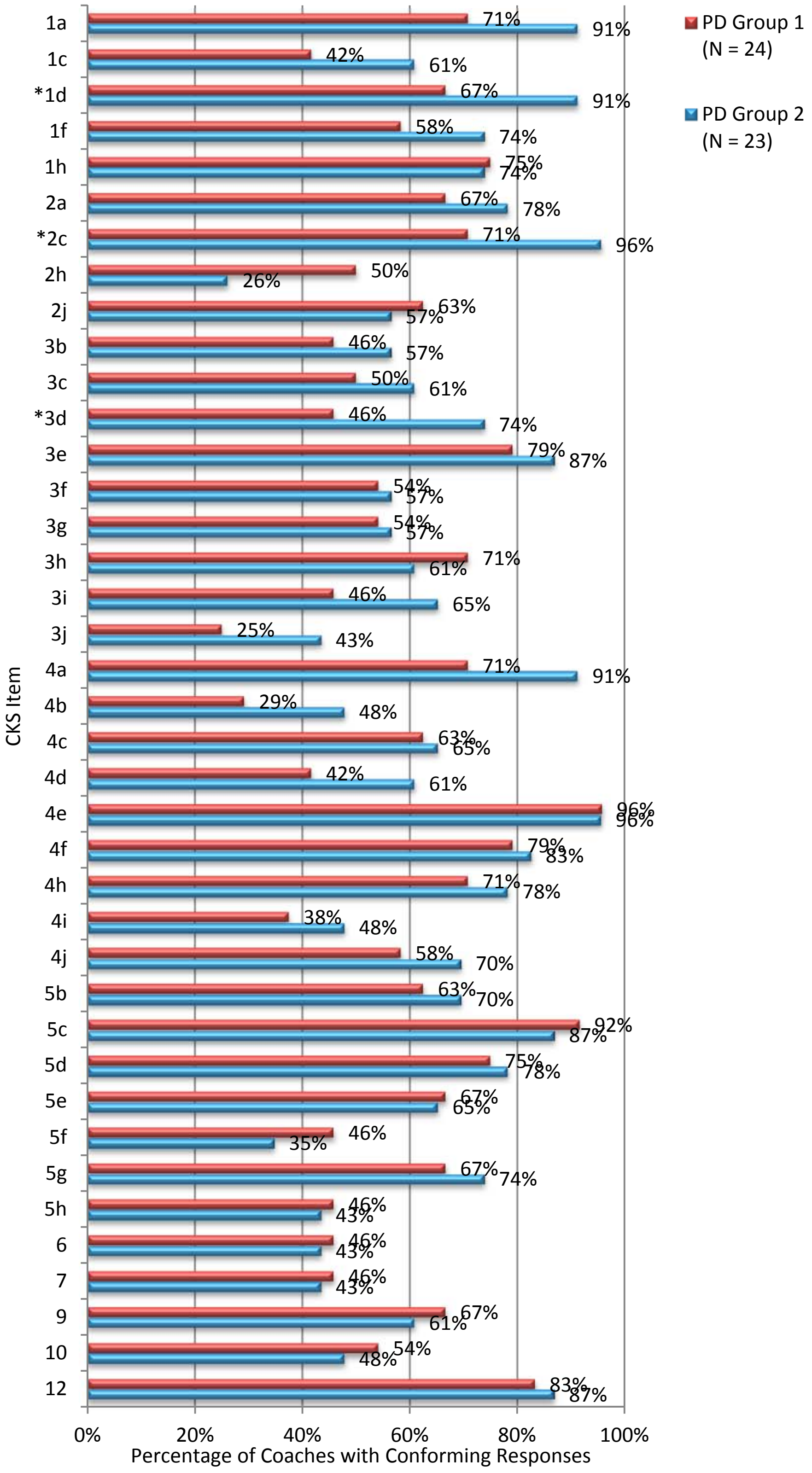
|  | TSTotalA1_mean | TSTotalA2_mean | TSTotalB_mean | TSTotalC_mean | TSTotalD_mean | ITCCap7ptA_mean | ITCCap7ptB_mean | ITCCap7ptC_mean | ITCCap7ptD_mean |
|---|---|---|---|---|---|---|---|---|---|
| When a teacher complains about the school's vision for mathematics, I ask the teacher about her or his vision for mathematics | .012 | -.023 | -.073 | .004 | .054 | .024 | .099 | .058 | -.009 |
|  | .942 | .885 | .650 | .980 | .736 | .880 | .536 | .718 | .958 |
| Base 10 Coach Scenario | .003 | .010 | .092 | .155 | -.034 | -.080 | .040 | -.007 | .100 |
|  | .986 | .949 | .568 | .335 | .832 | .617 | .805 | .966 | .536 |
| Ordering fractions Scenario. | .122 | .036 | .003 | -.072 | -.073 | .164 | .099 | .105 | .130 |
|  | .448 | .823 | .984 | .654 | .651 | .307 | .536 | .514 | .417 |
| Subtraction lesson observation scenario | .140 | .120 | .197 | .106 | .070 | .243 | .025 | -.145 | .266 |
|  | .381 | .453 | .216 | .511 | .664 | .126 | .876 | .367 | .093 |
| Teaching strategy discussion scenarios. | .318* | .370* | .339* | .355* | .199 | .164 | -.052 | .074 | -.006 |
|  | .042 | .017 | .030 | .023 | .213 | .305 | .746 | .647 | .970 |
| Which of the following is true about teachers and professional development without a coaching component? | -.088 | -.213 | -.093 | -.053 | .085 | .028 | -.130 | -.127 | -.242 |
|  | .585 | .182 | .563 | .742 | .599 | .860 | .418 | .430 | .128 |

*Note.* *Correlation is significant at the 0.05 level (2-tailed). **Correlation is significant at the 0.01 level (2-tailed)

## EXHIBIT B1: DISCRIMINANT ANALYSIS FOR QUESTION 1

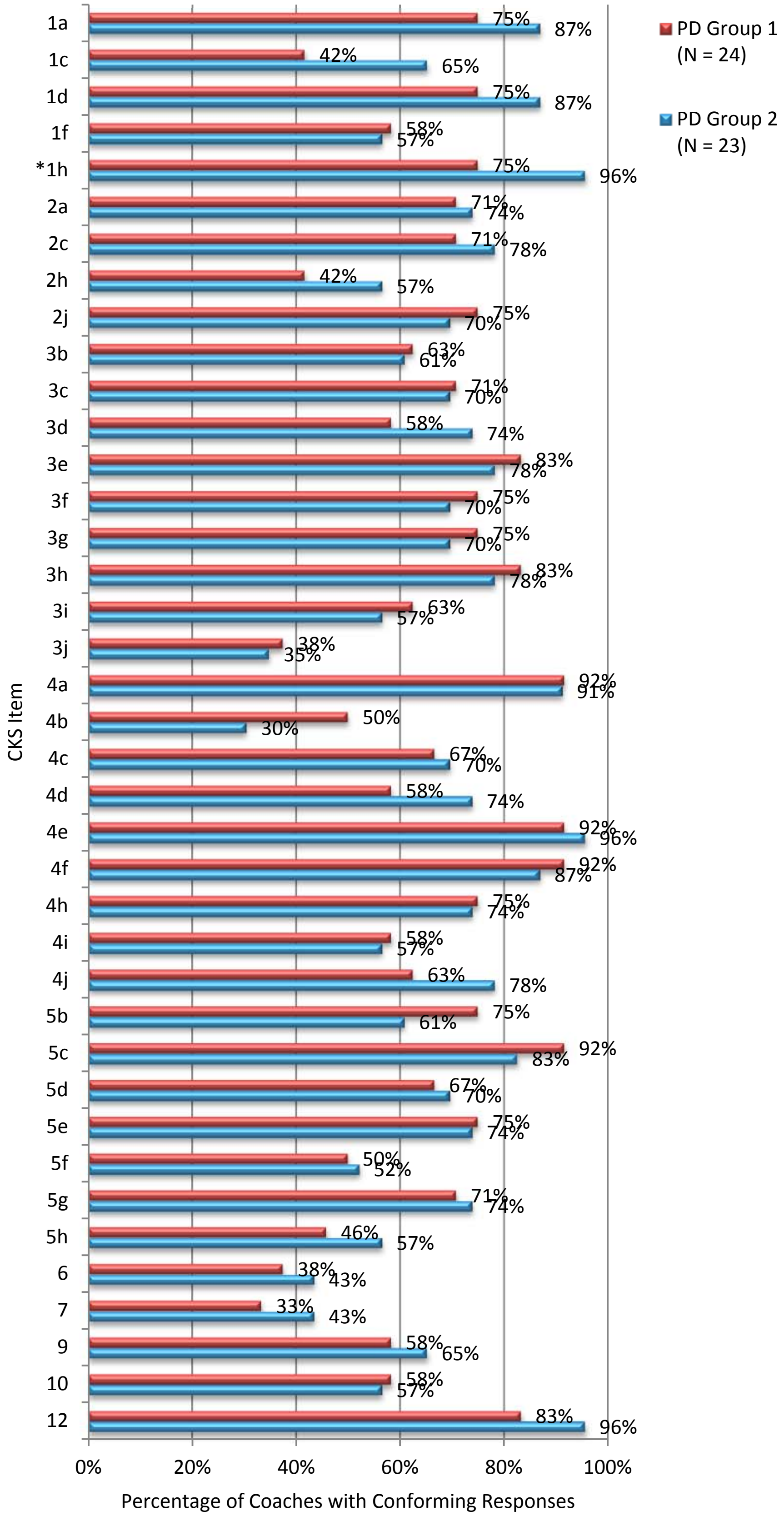
| | Time A | Time B | Time C | Time D | Time E |
|---|---|---|---|---|---|
| | Spring 2010 | Fall 2010 | Fall 2011 | Fall 2011 | Fall 2013 |
| a. An effective mathematics coach coaches only on teacher-stated needs. | | | X | | X |
| b. Beginning teachers need more coaching than 25-year veterans. | | | X | | |
| c. When a teacher says that she or he doesn't want any coaching, an effective mathematics coach respectfully does not try to persuade the teacher to accept coaching. | X | X | | | |
| d. Sometimes an effective mathematics coach has to oppose school or teacher actions that are not good for students' mathematics learning. | X | | | | |
| e. Teachers will adapt to whatever method of coaching is used. | | | | | |
| f. An effective mathematics coach gets input from a school's principal on which teachers need to improve their mathematics instruction. | | | | | |
| g. Number sense is a prerequisite for algebraic thinking. | | | | | |
| h. A coach should put no pressure on teachers to improve their practices. | | X | | | |
| i. In general, teachers need coaches to model a lesson with a particular strategy before they will incorporate it with fidelity. | | | | | |
| j. A teacher can learn new mathematics, but the teacher's basic mathematical intelligence cannot be changed. | | | | | |

## Exhibit C1. Percentage of Coaches with Conforming Responses in Year 1 by Item



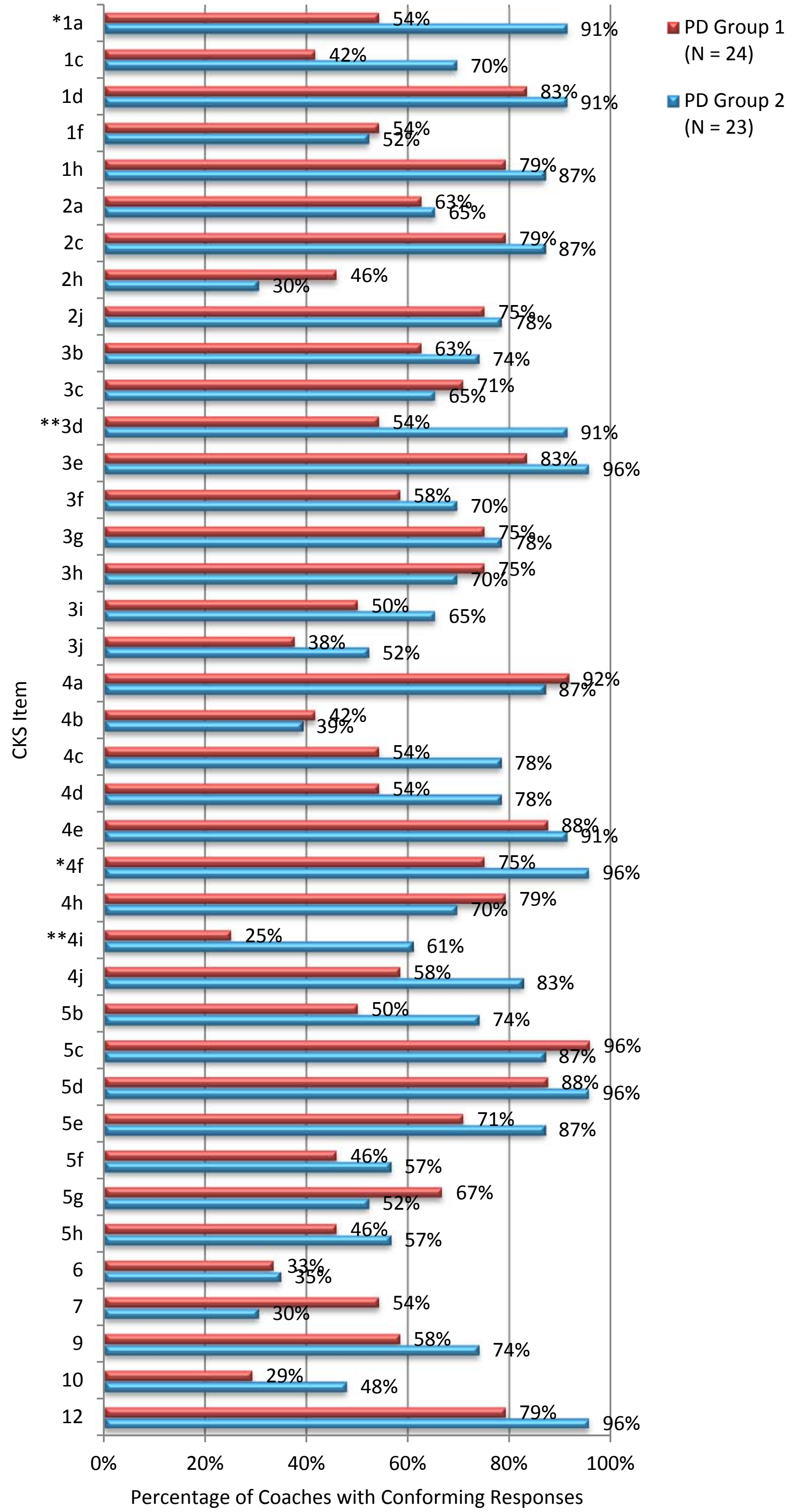
*$p < .05$; ** $p < .01$; ***$p < .001$.

## Exhibit C2. Percentage of Coaches with Conforming Responses in Year 2 by Item



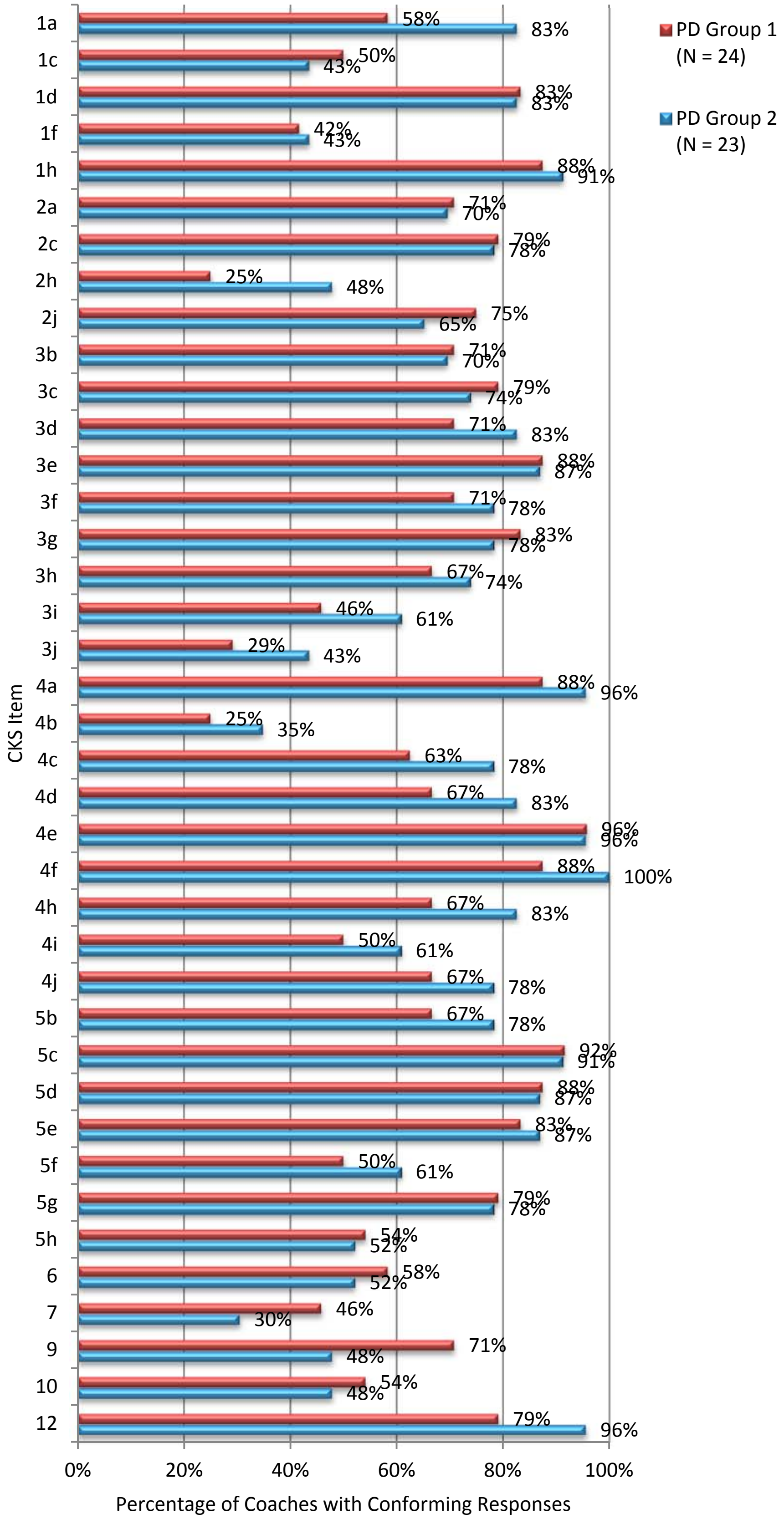
*$p$ < .05.

## Exhibit C3. Percentage of Coaches with Conforming Responses in Year 3 by Item

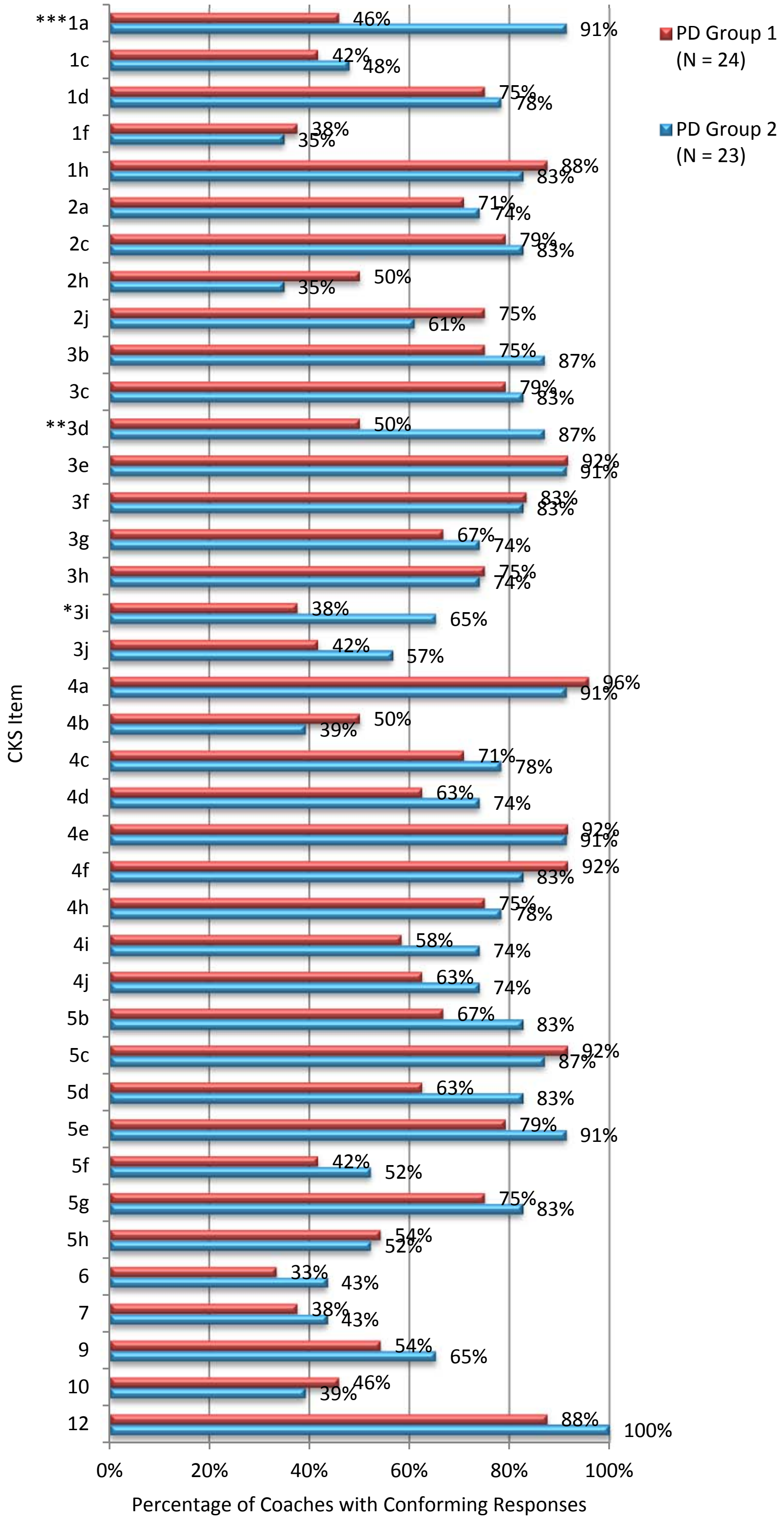


**p* < .05; ** *p* < .01.

## Exhibit C4. Percentage of Coaches with Conforming Responses in Year 4 by Item



## Exhibit C5. Percentage of Coaches with Conforming Responses in Year 5 by Item



** *p* < .01; ****p* < .001.

# CKS Scoring using Item Response Theory methods

Prepared as an internal report for EMC consideration by Mark Greenwood 3/14/2014

**Option 1: Generate a single latent trait with all items having a decent "fit" by starting with 1 factor exploratory factor analysis and then dropping any items that cause problems in the IRT.**

Benefits:

- Focuses on single most clearly identified underlying factor.
- All loadings were in the correct direction and the results pass all IRT diagnostics reasonably well.
- Single score for future analyses.
- Fewer number of items relative to sample size so closer to meeting rules of thumb for using IRT methods.
- Simpler to explain: screened items for single best underlying trait using EFA and fit an IRT model to top items, selecting between Rasch and more complicated models based on observed data.

Drawbacks:

- Only uses 20 of the 39 items.
- May ignore other "traits" of knowledge that the instrument is measuring.

**Option 2: Determine optimal number of traits for all items and then proceed with determining items that relate to certain traits.**

Benefits:

- Can detect multiple underlying subscales present in instrument.

Drawbacks:

- Number of factors is unclear – different methods suggest different numbers of latent traits.
- Multi-dimensional IRT is not standard. Scoring methods also may not be as clearcut.
    - o Splitting data set into groups of variables and running separate IRTs partially alleviate this issue but require many additional models to be reported/discussed.
- Arbitrariness of rotational method now part of methods to discuss. Varimax used but not the only option.
- Some items load negatively onto latent traits which are purported to be knowledge of aspects of literature but can't be if based on reversals of "scored" direction.
- Extra factors contain small number of items and still leave many items out of final models.

**Selecting the number of factors:**

For binary responses, the tetrachoric correlation is used to estimate the correlation. This method can create non-positive semi-definite correlation matrices and so a "smoothing" method is used to create an invertible matrix. The smoothing involves setting the negative eigenvalues to 0 and rescaling the remaining correlations appropriately. It is unclear how this impacts some of the methods for selecting the appropriate number of factors in the related EFA so the standard results may be taken with some additional caution.

The latent root criterion and parallel analysis suggest an extremely large number of factors as displayed in Figure 1. The Optimal Coordinates criterion is impacted by the "smoothing" procedure applied and without smoothing suggests 4 factors. The acceleration factor is much more conservative here and is also not impacted by the "smoothing", selecting the factor that corresponds to the most abrupt change in the eigenvalues. Its suggestion is to use only a single factor.

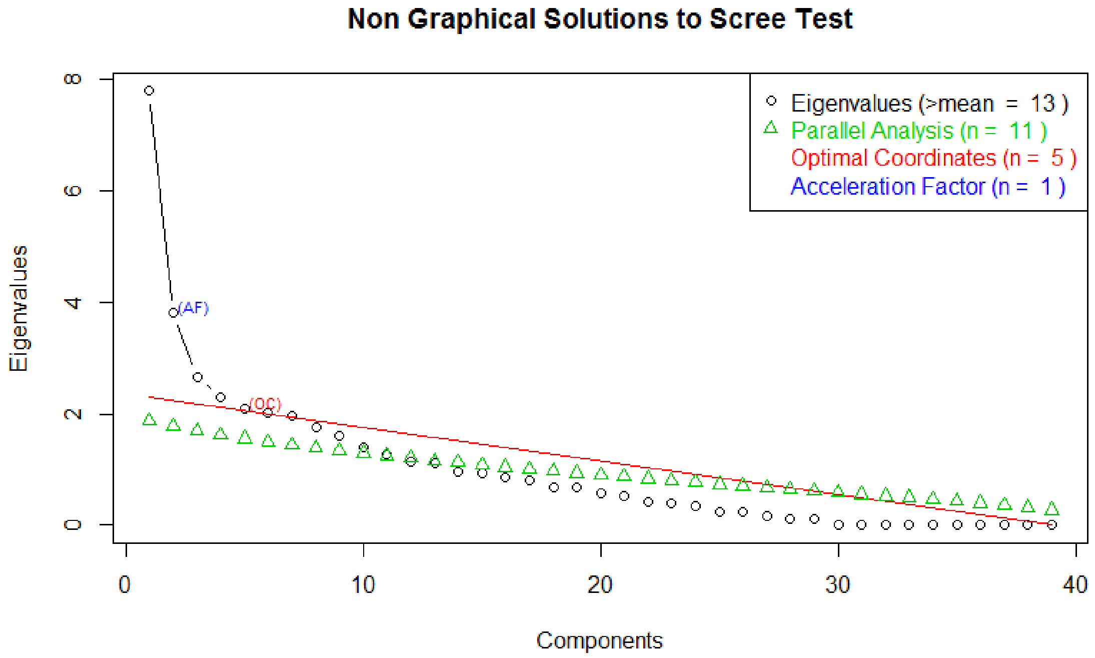


Figure 1. Plot of eigenvalues from smoothed tetrachoric correlation matrix with some selection criteria.

It seems then that either a 1 factor or 4 or 5 factor solutions are supported by the methods. It is hard to justify using 2 or 3 factors based on these results.

**Results for Option 1:**

With the support of the acceleration factor for selecting 1 factor, a 1-factor EFA follows. In this maximum likelihood EFA, there are 21 items with absolute values of factor loadings over 0.3 and 20 with loadings over 0.4. All are positively loading on the one factor. The remaining 18 items are discarded from the following analysis and the EFA is re-fit producing the following graphical display of the estimated factor analysis model:

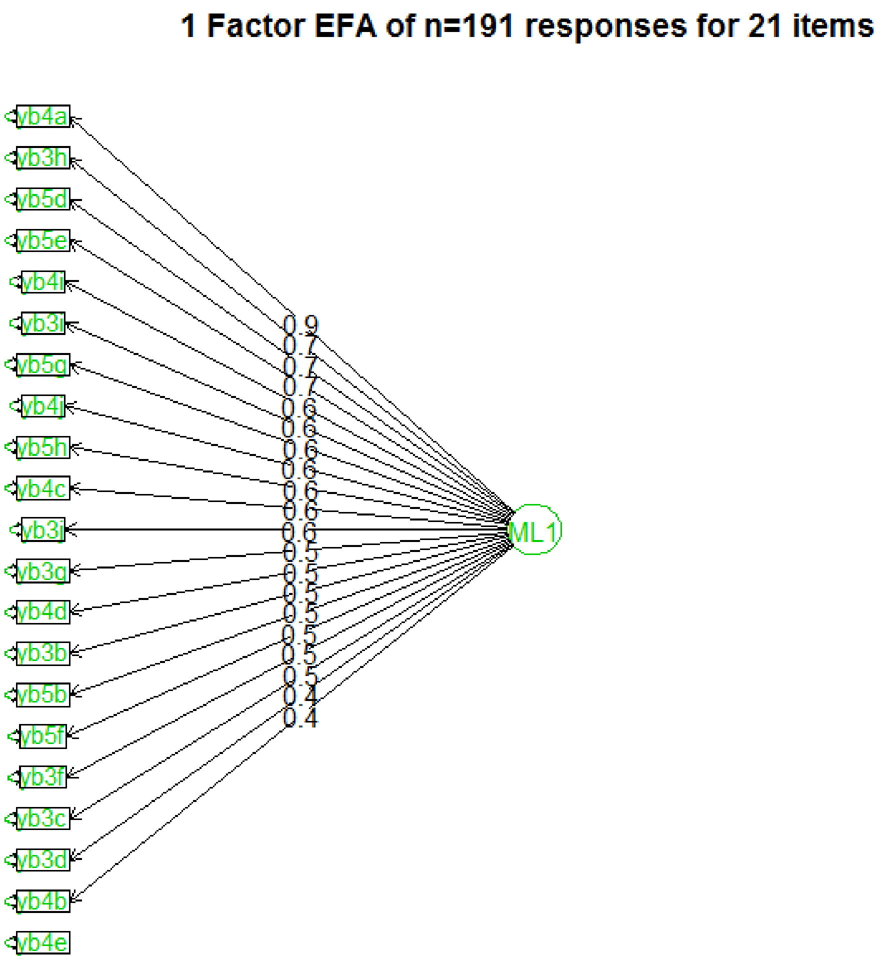


Figure 2. Path diagram of maximum likelihood EFA with one factor on 21 items suggested by initial EFA on 39 items.

The results are similar to those obtained with all 39 items except that item 4e now drops out with its loading going from 0.33 to 0.26. Since it was only marginally included in the first model, its exclusion is reasonable.

With this set of 20 items selected, an IRT model is fit to directly link the binary results to the same underlying trait identified previously and provide results for item difficulty and discrimination as well as methods for scoring new observations on this latent “coaching knowledge” trait.

In the IRT process, the difficulty of an item is related to how often respondents get the item correct (get a 1). The following table summarizes these results for the selected 20 items. It shows that all of the items are generally fairly “easy” with the hardest question getting positive responses 65% of the time. The easiest question had 96% positive responses.

Table 1. Proportions of n=191 responses for 20 selected items.

| | 0 | 1 |
|---|---|---|
| yb3j | 0.35 | 0.65 |
| yb5h | 0.33 | 0.67 |
| yb4b | 0.30 | 0.70 |
| yb4i | 0.29 | 0.71 |
| yb5f | 0.27 | 0.73 |
| yb3b | 0.21 | 0.79 |
| yb3i | 0.21 | 0.79 |
| yb4d | 0.21 | 0.79 |
| yb3c | 0.20 | 0.80 |
| yb3g | 0.19 | 0.81 |
| yb5b | 0.18 | 0.82 |
| yb3d | 0.18 | 0.82 |
| yb3f | 0.17 | 0.83 |
| yb4j | 0.17 | 0.83 |
| yb4c | 0.15 | 0.85 |
| yb5g | 0.15 | 0.85 |
| yb3h | 0.13 | 0.87 |
| yb5e | 0.08 | 0.92 |
| yb5d | 0.07 | 0.93 |
| yb4a | 0.04 | 0.96 |

With this subset of items, a Rasch model can be compared to an IRT (2-parameter) model. This comparison provides information about whether a model that provides different discrimination is a better description of the data set than the Rasch model that fixes the discrimination to be the same for all the items. A likelihood ratio test comparing the simpler Rasch model to the IRT model produces statistic of 29.39 which from a Chi-squared distribution with 19 degrees of freedom produces a p-value of 0.06. This is marginal but suggestive evidence to support the need to go to the more complicated model. It is also possible to do an overall goodness of fit test for the Rasch model and it provides a bootstrap p-value of 0.62, suggesting no major problem with the fit of the Rasch model. The Rasch model requires fewer parameters so follows closer to the rules of thumb for IRT/Rasch models of sample size versus number of estimated parameters.

For the 20 items selected, Cronbach's alpha is reasonably high, estimated to be 0.81 (95% CI from 0.732 to 0.862). This suggests good, but not excellent, internal reliability of these items.

For comparison, the estimated Item Information and Item Characteristic Curves from the Rasch and 2-parameter IRT models are provided in Figure 3. While the discrimination does vary between items in Figure 3b, the impacts of the item characteristic curves are minimal. The limited differences between panels c and d suggest that the Rasch model is a reasonable approximation of the structure in the items. Because the IRT model and the original EFA can "reverse" the direction of the relationship between the latent trait and the scored items and does not, this provides reassuring evidence that the model is estimating a valid latent trait.

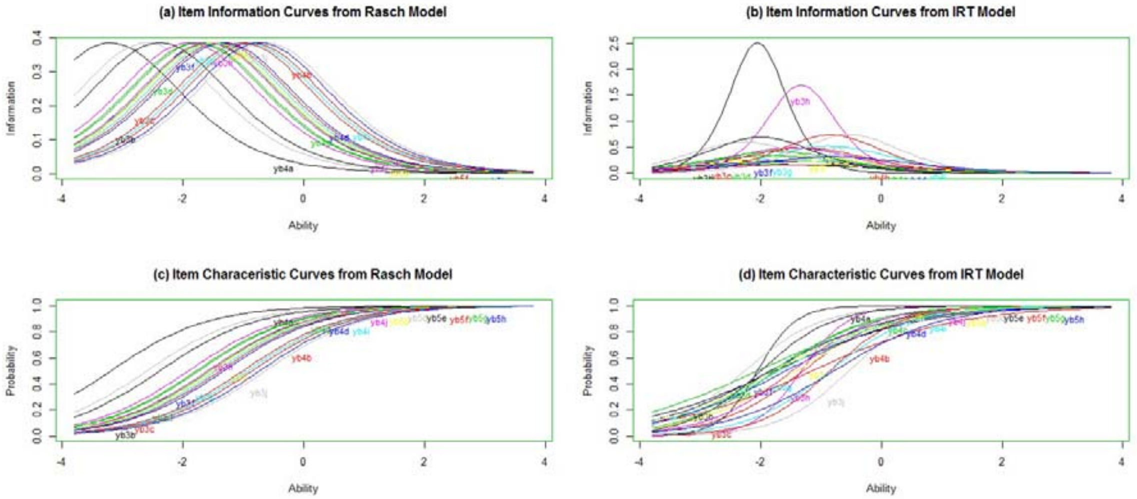


Figure 3. Item information and characteristic curves from Rasch and IRT models for 20 selected items.

While it is reassuring that all items have similar directions of relationships with the latent trait in the IRT, it is also possible to test whether each item "fits" with the overall model. Large test statistics and small p-values suggest that the item is performing differentially with respect to the IRT model (Reise, 1990). With 20 tests considered, moderately small p-values should be taken with a grain of salt as just by chance, one would expect one p-value lower than 0.05. To avoid inflated type I errors with these results, the p-values are Bonferroni corrected, with the results reported in Table 2. The smallest results are for question 3h, 4c, and 5f but none provide strong evidence of a lack of fit once the number of tests is considered in the correct p-values.

Table 2. Item fit test results for Rasch model.

| Item | Test Statistic | P-value | Corrected p-value |
|---|---|---|---|
| yb3b | 6.24 | 0.75 | 1.00 |
| yb3c | 12.41 | 0.16 | 1.00 |
| yb3d | 10.82 | 0.26 | 1.00 |
| yb3f | 7.84 | 0.54 | 1.00 |
| yb3g | 4.51 | 0.91 | 1.00 |
| yb3h | 18.50 | 0.02 | 0.44 |
| yb3i | 10.77 | 0.30 | 1.00 |
| yb3j | 12.87 | 0.30 | 1.00 |
| yb4a | 8.30 | 0.28 | 1.00 |
| yb4b | 17.93 | 0.06 | 1.00 |
| yb4c | 17.11 | 0.03 | 0.54 |
| yb4d | 6.06 | 0.79 | 1.00 |
| yb4i | 13.23 | 0.21 | 1.00 |
| yb4j | 8.65 | 0.46 | 1.00 |
| yb5b | 8.50 | 0.52 | 1.00 |

| | | | |
|---|---|---|---|
| yb5d | 6.97 | 0.53 | 1.00 |
| yb5e | 10.64 | 0.21 | 1.00 |
| yb5f | 18.82 | 0.03 | 0.64 |
| yb5g | 6.52 | 0.69 | 1.00 |
| yb5h | 12.76 | 0.26 | 1.00 |

It is also possible to test for lack of fit of the model to a subject using the person fit tests (Reise, 1990). These tests can be used to identify and possibly remove subjects with “deviant” response patterns. Across the 191 subjects, the smallest observed p-value was 0.017 which does not suggest strong evidence of a problem given that a test is performed for each subject.

The last concern is that the items have been combined across more than one latent trait. This is addressed directly with a test of unidimensionality with a null hypothesis of a single underlying trait versus an alternative of more than one trait, assessed using a test statistic based on the second eigenvalue of the observed data set. The results do not suggest any evidence of an additional trait in these items with a p-value of 0.227.

The estimated Rasch model has coefficients described in Table 3. The negative difficulty parameter estimates correspond to the earlier results that suggest that all the questions are relatively easy, centering each Item Information curve below 0 on the difficulty scale (Figure 3a). The items still provide some discrimination for individuals with above average CKS but most of the discrimination is focused on the below average results since most of the subjects in the pilot data set (generally) got these items correct. In agreement with the initial results, item 4a was the easiest and 3j was the hardest (but only centered at -0.63). The common discrimination parameter for all items is estimated to be 1.24 (95% CI from 1.05 to 1.43).

Table 3. Rasch model estimated parameters.

| Model Parameter | Estimate | SE | Z |
|---|---|---|---|
| Dffclt.yb3b | -1.3247 | 0.1934 | -6.8511 |
| Dffclt.yb3c | -1.3874 | 0.1974 | -7.0288 |
| Dffclt.yb3d | -1.5554 | 0.2091 | -7.4397 |
| Dffclt.yb3f | -1.591 | 0.2117 | -7.5154 |
| Dffclt.yb3g | -1.4525 | 0.2018 | -7.1989 |
| Dffclt.yb3h | -1.9095 | 0.2378 | -8.0299 |
| Dffclt.yb3i | -1.3246 | 0.1933 | -6.8509 |
| Dffclt.yb3j | -0.6348 | 0.1609 | -3.9462 |
| Dffclt.yb4a | -3.1965 | 0.398 | -8.0312 |
| Dffclt.yb4b | -0.854 | 0.1688 | -5.0604 |
| Dffclt.yb4c | -1.7416 | 0.2235 | -7.7933 |
| Dffclt.yb4d | -1.3248 | 0.1934 | -6.8514 |
| Dffclt.yb4i | -0.9307 | 0.1721 | -5.4091 |
| Dffclt.yb4j | -1.5902 | 0.2116 | -7.5137 |
| Dffclt.yb5b | -1.5201 | 0.2065 | -7.361 |
| Dffclt.yb5d | -2.597 | 0.3108 | -8.3547 |

| | | | |
|---|---|---|---|
| Dffclt.yb5e | -2.3861 | 0.2858 | -8.3492 |
| Dffclt.yb5f | -1.0096 | 0.1758 | -5.7445 |
| Dffclt.yb5g | -1.7821 | 0.2268 | -7.8573 |
| Dffclt.yb5h | -0.7305 | 0.164 | -4.4535 |
| **Discrimination** | **1.242** | **0.0963** | **12.897** |

Together, these results suggest that the Rasch model is a reasonable approximation of the results observed in the pilot data set and provides a single CKS trait. Further exploration of the items selected for the unidimensional CKS trait will provide insights into the meaning of the CKS scores produced. The selected model can be used to score responses over time from the coaches in the EMC project. These scores will be used both as explanatory variables for teacher level responses and as outcome variables to assess impacts of the PD on coaching knowledge in this domain.

**Option 2:**

The initial screening criteria were not clear about the number of factors to use, but if more than one trait is considered for the suite of 39 original items, it is possible to explore additional dimensions of information from the CKS responses. These additional dimensions should follow the same general pattern as before – whatever the underlying trait is that is considered, the items should strongly load in a positive direction with that trait and only with that trait (double loadings would mean that the same item would be used as part of two scores). As before, the EFA of the tetrachoric correlation matrix provides a useful starting point for doing IRT analyses of the items. The focus of the EFA is on identifying the variables associated with underlying latent traits. It is also useful to verify that the estimated direction of relationships between items and latent traits are in a direction that would provide latent traits that exclusively mean higher levels of conforming to the literature on coaching. Negative loadings in a two-parameter IRT would likely correspond to negative discrimination coefficients and an induced lack of clarity in the meaning of the underlying trait(s).

Maximum likelihood estimation was attempted for the two through four factor models considered below, but failed to converge for more than 2 factors. The minimum residual method available in the fa function from the psych package was used to estimate the models. Figures 4, 5, and 6 below provide the varimax rotated, minimum residual estimated EFAs for two, three, and four factors. Only loadings over 0.4 are displayed in the fa diagrams as a high number of double and negative loadings are encountered if a lower threshold is used.

For two factors, displayed in Figure 4, the first factor matches the previous single factor results except that item 4b is dropped. However, item 3h double loads with the second factor. For the second factor, it seems to include some new items excluded from the one factor solution except that one of the loadings is negative while five are positive including the double loading 3h. This creates a difficult interpretation from an IRT perspective for this new trait with higher “ability” on this trait related to higher chances of conforming to the literature on most items and disagreeing with literature on another. As a study on perceptions of coaching, this might be interesting to explore further but creates difficulty in creating an instrument that relates to knowledge of the coaching literature. The additional complication of double loadings creates uncertainties in how the scores should be created. This solution only uses information from four more items than the first solution since item 4b drops from factor 1 into factor 2.

For the three factor solution in Figure 5, the items loading on the first two factors basically match the two factor solution. The third factor contains only four new items with two negatively loading on the trait. Ten items are still not loading at a high level on any factor.

For the four factor solution in Figure 6, the first factor retains the same components as in previous solutions. However, there are now two items that double load with the second factor. The second factor has one negative loading. The third factor matches the third factor in the previous solution. The new fourth factor includes four new items in the results with one negative loading. Seven items would still be dropped from the analysis.

For all of the higher dimensional solutions, there are still items that would be dropped from the interpretation and further analyses/scoring and so would retain a similar issue as was encountered in Option 1. In each of the potential solutions in Option 2, there are negative and double loadings encountered. Finally, the extra factors are only associated with a few items. This does not create much information to create a score from.

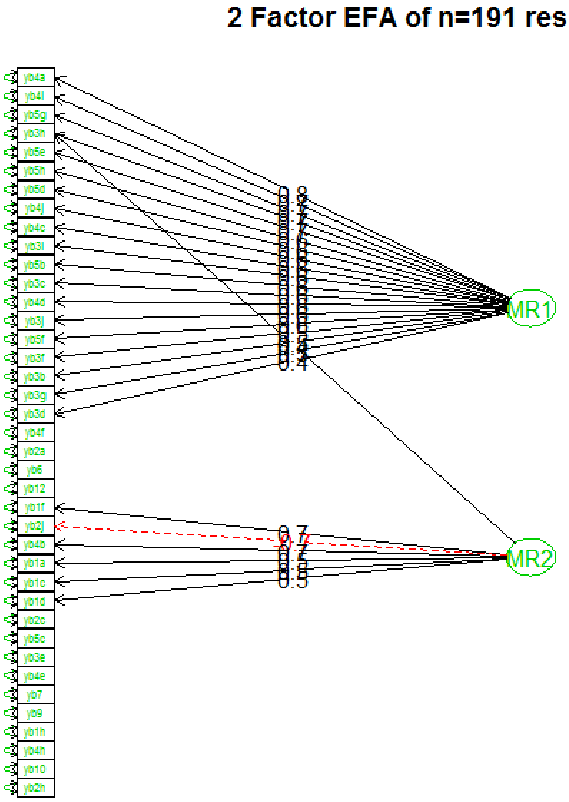


Figure 4. Minimum residuals, varimax rotated, two-factor EFA path diagram with loadings over 0.4 displayed. Red, dashed lines correspond to negative loadings.

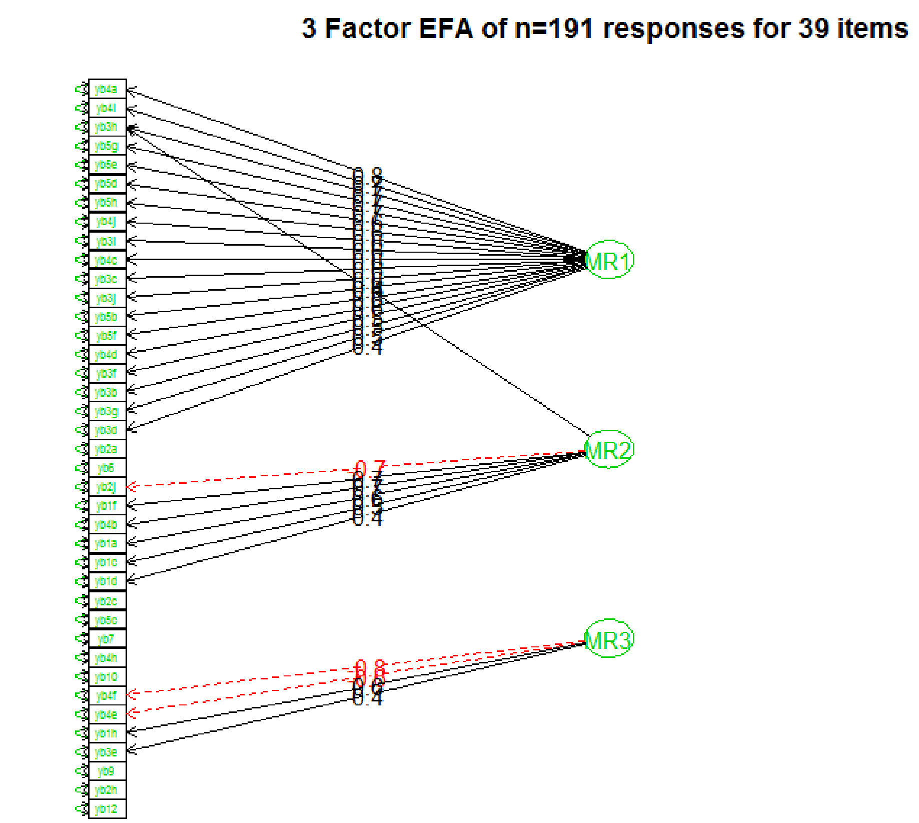


Figure 5. Minimum residuals, varimax rotated, three-factor EFA path diagram with loadings over 0.4 displayed. Red, dashed lines correspond to negative loadings.

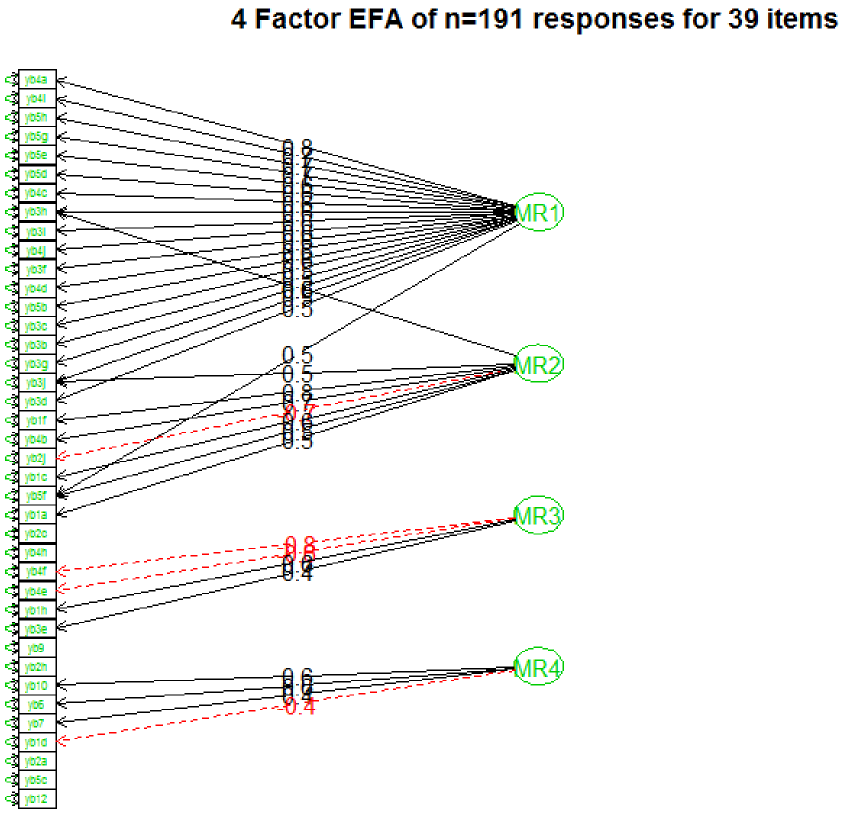


Figure 6. Minimum residuals, varimax rotated, four-factor EFA path diagram with loadings over 0.4 displayed. Red, dashed lines correspond to negative loadings.

**General conclusion:**

Selecting a subset of items that loaded heavily in a single factor EFA provided a reliable set of 20 items to use in a Rasch model. The Rasch model is supported by the n=191 observations from the pilot data set and can be readily applied to EMC CKS longitudinal responses to create a single score for each year of the study. Additional discussion of the items retained/lost may provide further insight into the latent trait actually being measured by these items.

Attempts to use additional factors encounter problems with maximum likelihood estimation and shows negative and double loading on the underlying traits. Additionally, extra factors that are only associated with a few items. The criteria to aid in selecting a particular number of factors were generally inconclusive in these data and the added complexity versus added information retained seem to suggest retaining only the first main factor measured in the CKS items.

One additional benefit of the support for a Rasch model in the pilot data set is that it provides an opportunity to directly compare the analysis of the scored items using this scoring model as they change over time to a GLMM approach that uses the longitudinal data only to both score and, simultaneously, analyze changes over time.

There is one potential complication that scoring of patterns in new observations not observed in the original pilot data set could create a small issue in generating a new. Additionally, all scores will need to be created within R so a protocol for working with our current database management methods will need to be developed. This issue will be encountered with any of the proposed factor solutions when converted into an IRT scoring model(s).

**References (more in finalization of this report):**

Reise, S. (1990) A Comparison of Item- and Person-fit Methods for Assessing Model-Data Fit in IRT. *Applied Psychological Measurement*, 14(2), 127-137.

## Descriptives on the 20 CKS Items Retained

Based on a set of analyses conducted by EMC research staff 20 items within the CKS were found to load onto one factor and were retained for future analyses. The following set of exhibits show the amount of agreement among coaches over time for these 20 items. This analysis utilized the coach only data set, rather than the teacher plus coach data set, and was restricted to include only coaches who completed all five administrations of the CKS.

The frequency of coaches with conforming responses is presented in the following exhibits, both in the aggregate and by PD group. Exhibit 1 contains findings for items within question 3 that were identified in the previous analysis. From Year 1 to Year 5, the number of coaches in the aggregate with conforming responses increased for seven of the eight items and decreased for one item.

### Exhibit 1. The Number of Coaches with Conforming Responses Over Time for Question 3

|  | Number of Coaches with Conforming Responses | | | | | |
|---|---|---|---|---|---|---|
| **CKS Items** | **Year 1** | **Year 2** | **Year 3** | **Year 4** | **Year 5** | **Trend** |
| **3b. I collect students' mathematics work from a teacher's classroom to guide our coaching conversations.** | | | | | | |
| Aggregate (*N* = 47) | 24 | 29 | 32 | 33 | 38 | + |
| PD Group 1 (*N* = 24) | 11 | 15 | 15 | 17 | 18 | + |
| PD Group 2 (*N* = 23) | 13 | 14 | 17 | 16 | 20 | + |
| **3c. When decisions about mathematics instruction are being made, I ensure that the decision-makers interpret research literature accurately.** | | | | | | |
| Aggregate (*N* = 47) | 26 | 33 | 32 | 36 | 38 | + |
| PD Group 1 (*N* = 24) | 12 | 17 | 17 | 19 | 19 | + |
| PD Group 2 (*N* = 23) | 14 | 16 | 15 | 17 | 19 | + |
| **3d. I coach teachers on needs that I observe in the teacher, even when the teacher is unaware of these needs.** | | | | | | |
| Aggregate (*N* = 47) | 28 | 31 | 34 | 36 | 32 | + |
| PD Group 1 (*N* = 24) | 11 | 14 | 13 | 17 | 12 | + |
| PD Group 2 (*N* = 23) | 17 | 17 | 21 | 19 | 20 | + |
| **3f. I have difficult conversations with teachers, when necessary, about mathematics misconceptions they hold.** | | | | | | |
| Aggregate (*N* = 47) | 26 | 34 | 30 | 35 | 39 | + |
| PD Group 1 (*N* = 24) | 13 | 18 | 14 | 17 | 20 | + |
| PD Group 2 (*N* = 23) | 13 | 16 | 16 | 18 | 19 | + |
| **3g. I always make sure that coaching conversations with mathematics teachers are grounded in the mathematics content.** | | | | | | |
| Aggregate (*N* = 47) | 26 | 34 | 36 | 38 | 33 | + |
| PD Group 1 (*N* = 24) | 13 | 18 | 18 | 20 | 16 | + |
| PD Group 2 (*N* = 23) | 13 | 16 | 18 | 18 | 17 | + |
| **3h. I meet with the principal to discuss the school's vision for mathematics instruction.** | | | | | | |
| Aggregate (*N* = 47) | 31 | 38 | 34 | 33 | 35 | + |
| PD Group 1 (*N* = 24) | 17 | 20 | 18 | 16 | 18 | + |
| PD Group 2 (*N* = 23) | 14 | 18 | 16 | 17 | 17 | + |

| | | | | | | |
|---|---|---|---|---|---|---|
| **3i. I encourage teachers to include, in each lesson they teach, summaries of what students learned or discovered.** | | | | | | |
| Aggregate (*N* = 47) | 26 | 28 | 27 | 25 | 24 | - |
| PD Group 1 (*N* = 24) | 11 | 15 | 12 | 11 | 9 | - |
| PD Group 2 (*N* = 23) | 15 | 13 | 15 | 14 | 15 | * |
| **3j. I provide feedback to teachers about whether or not the school is meeting its vision for mathematics instruction.** | | | | | | |
| Aggregate (*N* = 47) | 16 | 17 | 21 | 17 | 23 | + |
| PD Group 1 (*N* = 24) | 6 | 9 | 9 | 7 | 10 | + |
| PD Group 2 (*N* = 23) | 10 | 8 | 12 | 10 | 13 | + |

*Note*. PD Group 1 received coaching professional development in summer 2012 and PD Group 2 received coaching professional development in summer 2011. A + indicates an increase from year 1 to year 5, a * indicates no change, and a – indicates a decrease from year 1 to year 5.

Exhibit 2 contains findings for items within question 4 that were retained. In the aggregate, the frequency of coaches with conforming responses increased from Year 1 to Year 5 on all six of the items.

### Exhibit 2. The Number of Coaches with Conforming Responses Over Time for Question 4

| | **Number of Coaches with Conforming Responses** | | | | | |
|---|---|---|---|---|---|---|
| **CKS Items** | **Year 1** | **Year 2** | **Year 3** | **Year 4** | **Year 5** | **Trend** |
| **4a. I try to provide the teachers I coach with an understanding of how the mathematics they teach supports learning beyond the grade level they teach.** | | | | | | |
| Aggregate (*N* = 47) | 38 | 43 | 42 | 43 | 44 | + |
| PD Group 1 (*N* = 24) | 17 | 22 | 22 | 21 | 23 | + |
| PD Group 2 (*N* = 23) | 21 | 21 | 20 | 22 | 21 | * |
| **4b. I ask the principal what he or she believes the mathematics teachers' needs are.** | | | | | | |
| Aggregate (*N* = 47) | 18 | 19 | 19 | 14 | 21 | + |
| PD Group 1 (*N* = 24) | 7 | 12 | 10 | 6 | 12 | + |
| PD Group 2 (*N* = 23) | 11 | 7 | 9 | 8 | 9 | - |
| **4c. I encourage the teachers I coach to reflect on similarities and differences among mathematics topics in the curriculum.** | | | | | | |
| Aggregate (*N* = 47) | 30 | 32 | 31 | 33 | 35 | + |
| PD Group 1 (*N* = 24) | 15 | 16 | 13 | 15 | 17 | + |
| PD Group 2 (*N* = 23) | 15 | 16 | 18 | 18 | 18 | + |
| **4d. I help teachers plan their lessons.** | | | | | | |
| Aggregate (*N* = 47) | 24 | 31 | 31 | 35 | 32 | + |
| PD Group 1 (*N* = 24) | 10 | 14 | 13 | 16 | 15 | + |
| PD Group 2 (*N* = 23) | 14 | 17 | 18 | 19 | 17 | + |
| **4i. I help teachers identify consistencies and inconsistencies between their own practices and the practices recommended by the National Council of Teachers of Mathematics.** | | | | | | |
| Aggregate (*N* = 47) | 20 | 27 | 20 | 26 | 31 | + |
| PD Group 1 (*N* = 24) | 9 | 14 | 6 | 12 | 14 | + |
| PD Group 2 (*N* = 23) | 11 | 13 | 14 | 14 | 17 | + |

| | | | | | | |
|---|---|---|---|---|---|---|
| **4j. I work with principals or other administrators to form a clear message to teachers about effective mathematics instruction.** | | | | | | |
| Aggregate (*N* = 47) | 30 | 33 | 33 | 34 | 32 | + |
| PD Group 1 (*N* = 24) | 14 | 15 | 14 | 16 | 15 | + |
| PD Group 2 (*N* = 23) | 16 | 18 | 19 | 18 | 17 | + |

*Note*. PD Group 1 received coaching professional development in summer 2012 and PD Group 2 received coaching professional development in summer 2011. A + indicates an increase from year 1 to year 5, a * indicates no change, and a – indicates a decrease from year 1 to year 5.

Items within question 5 that were retained are shown in Exhibit 3. The frequency of coaches in the aggregate increased over time for five items and decreased on one item.

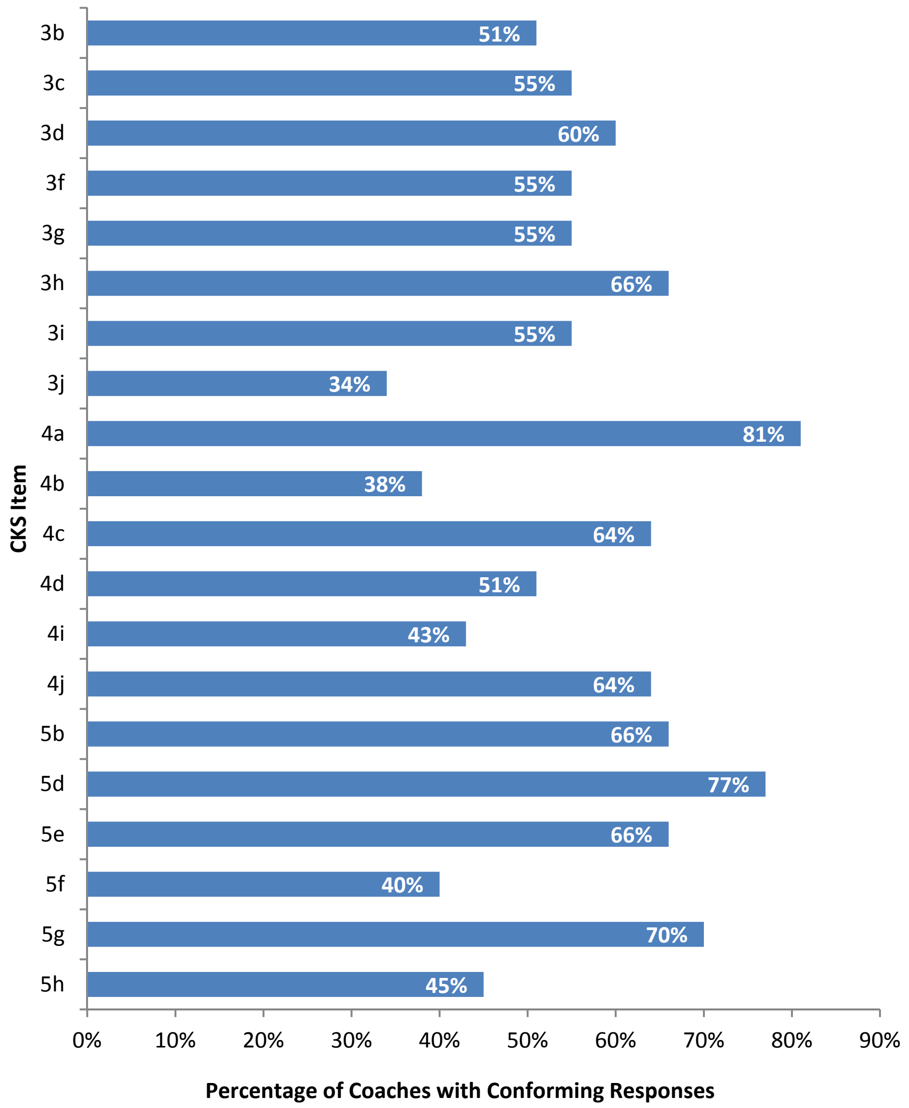
### Exhibit 3. The Number of Coaches with Conforming Responses Over Time for Question 5

| | **Number of Coaches with Conforming Responses** | | | | | |
|---|---|---|---|---|---|---|
| **CKS Items** | **Year 1** | **Year 2** | **Year 3** | **Year 4** | **Year 5** | **Trend** |
| **5b. I help teachers reflect on discrepancies between espoused beliefs and actual practices.** | | | | | | |
| Aggregate (*N* = 47) | 31 | 32 | 29 | 34 | 35 | + |
| PD Group 1 (*N* = 24) | 15 | 18 | 12 | 16 | 16 | + |
| PD Group 2 (*N* = 23) | 16 | 14 | 17 | 18 | 19 | + |
| **5d. I reflect on state assessment data to identify curriculum areas that need to be strengthened.** | | | | | | |
| Aggregate (*N* = 47) | 36 | 32 | 43 | 41 | 34 | - |
| PD Group 1 (*N* = 24) | 18 | 16 | 21 | 21 | 15 | - |
| PD Group 2 (*N* = 23) | 18 | 16 | 22 | 20 | 19 | + |
| **5e. I use student work when coaching mathematics teachers.** | | | | | | |
| Aggregate (*N* = 47) | 31 | 35 | 37 | 40 | 40 | + |
| PD Group 1 (*N* = 24) | 16 | 18 | 17 | 20 | 19 | + |
| PD Group 2 (*N* = 23) | 15 | 17 | 20 | 20 | 21 | + |
| **5f. I provide feedback to the principal about whether or not the school is meeting its vision for mathematics instruction.** | | | | | | |
| Aggregate (*N* = 47) | 19 | 24 | 24 | 26 | 22 | + |
| PD Group 1 (*N* = 24) | 11 | 12 | 11 | 12 | 10 | - |
| PD Group 2 (*N* = 23) | 8 | 12 | 13 | 14 | 12 | + |
| **5g. I encourage teachers to set personal improvement goals for mathematics instruction.** | | | | | | |
| Aggregate (*N* = 47) | 33 | 34 | 28 | 37 | 37 | + |
| PD Group 1 (*N* = 24) | 16 | 17 | 16 | 19 | 18 | + |
| PD Group 2 (*N* = 23) | 17 | 17 | 12 | 18 | 19 | + |
| **5h. When a teacher complains about the school's vision for mathematics, I ask the teacher about her or his vision for mathematics.** | | | | | | |
| Aggregate (*N* = 47) | 21 | 24 | 24 | 25 | 25 | + |
| PD Group 1 (*N* = 24) | 11 | 11 | 11 | 13 | 13 | + |
| PD Group 2 (*N* = 23) | 10 | 13 | 13 | 12 | 12 | + |

*Note*. PD Group 1 received coaching professional development in summer 2012 and PD Group 2 received coaching professional development in summer 2011. A + indicates an increase from year 1 to year 5, a * indicates no change, and a – indicates a decrease from year 1 to year 5.

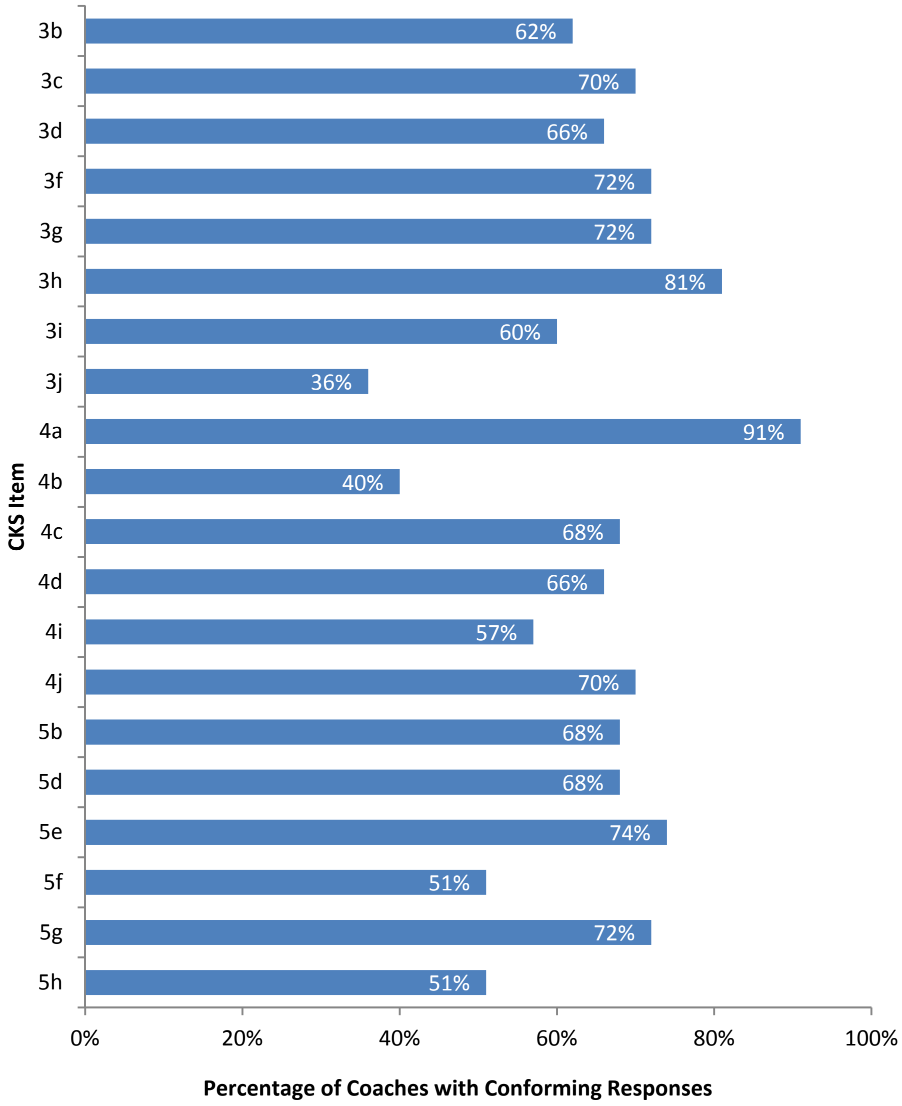
The percentage of coaches in the aggregate with conforming responses to the 20 items across each year is presented in the following five exhibits. Only Items 4a and 5d had 75% or more of coaches conforming to the statements. Less than half of the coaches conformed to Items 3j, 4b, 4i, 5f, and 5h.

Exhibit 4. Percentage of Coaches with Conforming Responses in Year 1 by Item (N = 47)



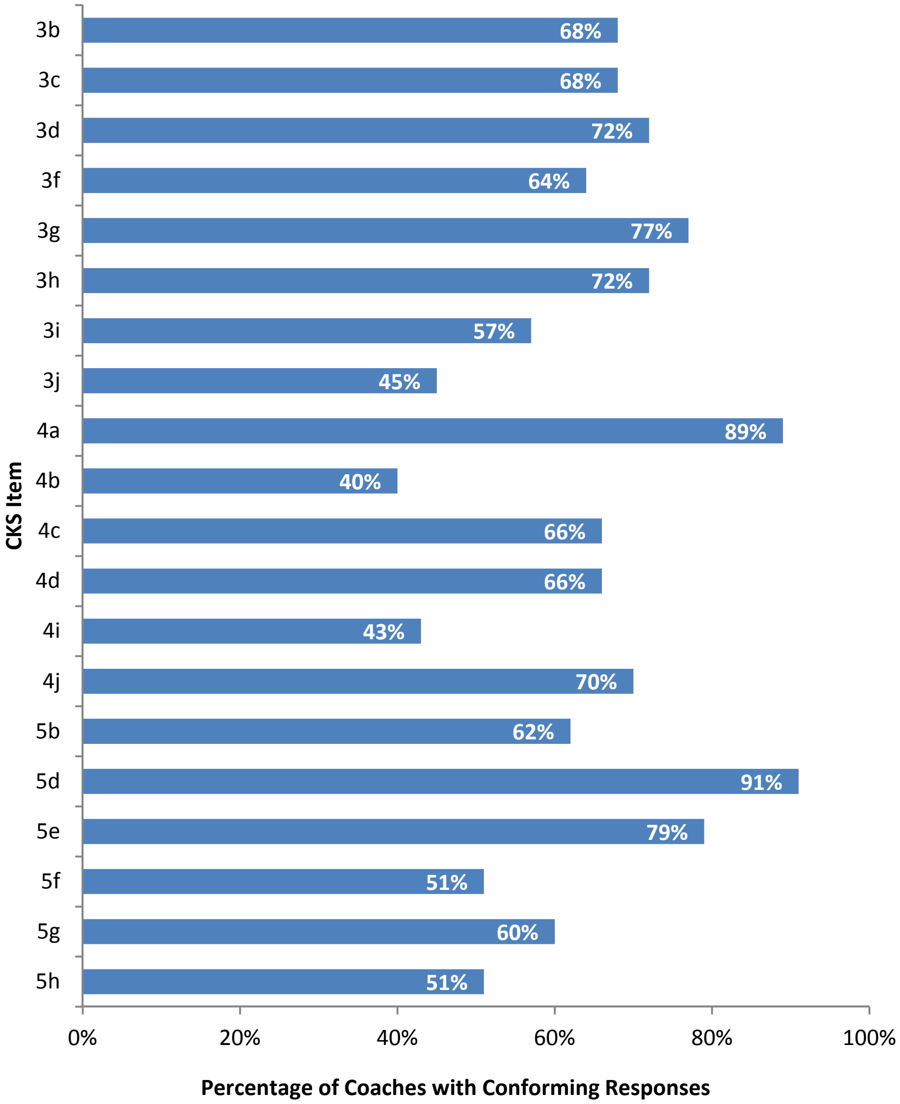
Items for which there were 75% or more of coaches with conforming responses in Year 2 included Item 3h and Item 4a again, with item 5e close behind.  Items 3j and 4b continued to have less than 50% of the coach scores conform to the statement.

### EXHIBIT 5. PERCENTAGE OF COACHES WITH CONFORMING RESPONSES IN YEAR 2 BY ITEM (N = 47)



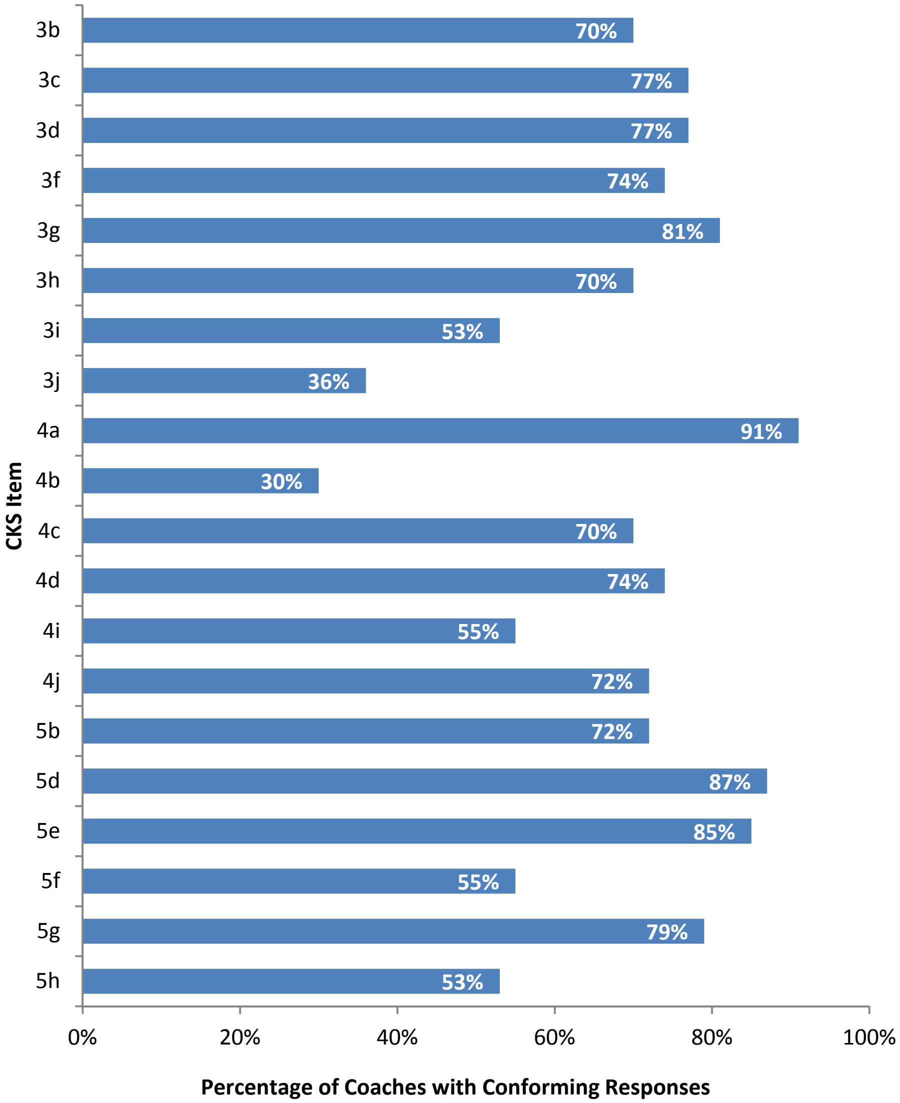
In Year 3 four items had 75% or more of coaches with conforming responses including Items 3g, 4a, 5d, and 5e, with Items 3d and 3h trailing close behind. Items which had the lowest frequency of coaches with conforming responses continued to include Items 3j and 4b, along with Item 4i.

EXHIBIT 6. PERCENTAGE OF COACHES WITH CONFORMING RESPONSES IN YEAR 3 BY ITEM (N = 47)



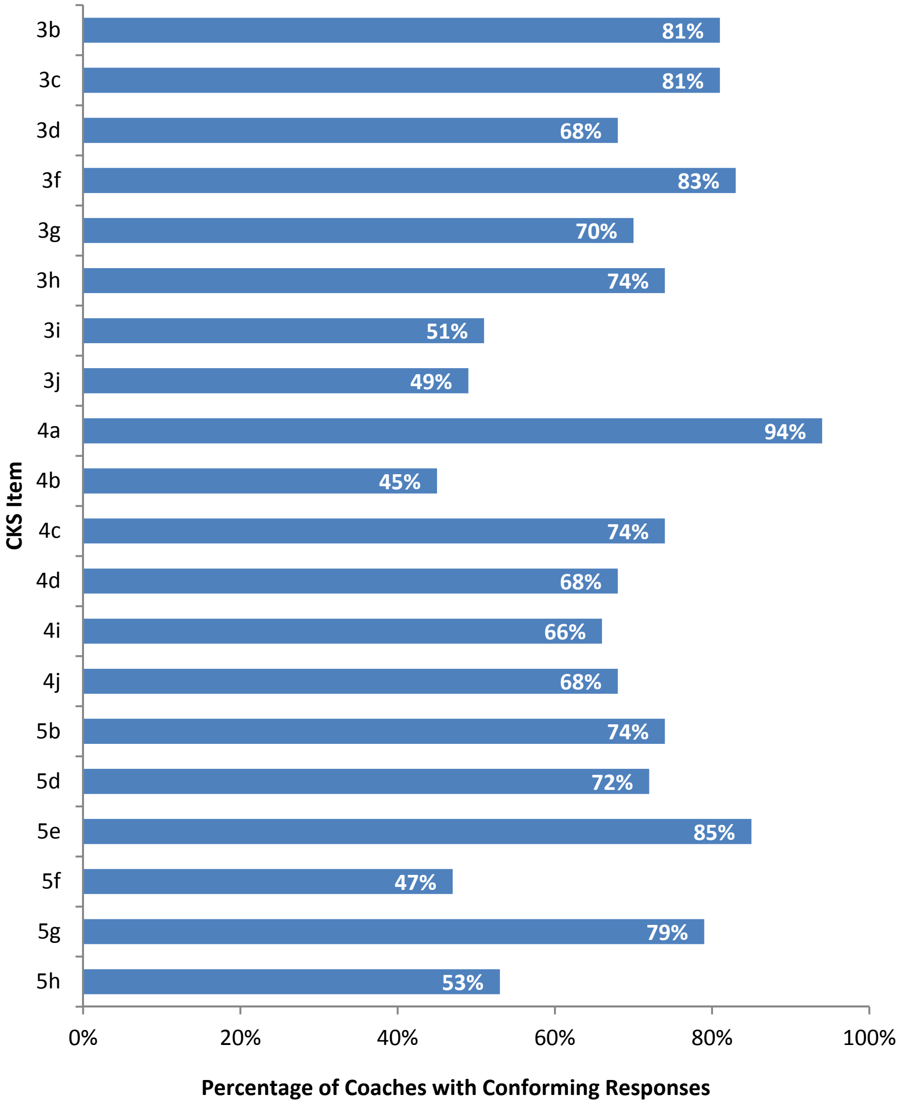
Items for which there were 75% or more of coaches with conforming responses in Year 4 included items 3c, 3d, 3g, 4a, 5d, 5e, and 5g. Items 3f and 4d also had high rates of conforming responses. Items which had the lowest frequency of coaches with conforming responses continued to include items 3j and 4b.

### EXHIBIT 7. PERCENTAGE OF COACHES WITH CONFORMING RESPONSES IN YEAR 4 BY ITEM (N = 47)



In Year 5, the following items had 75% or more of the coaches conforming to the statements: 3b, 3c, 3f, 4a, 5e, and 5g. Items 3h, 4c, 5b, and 5d were just under 75%. As with other years, Items 3j and 4b, along with Item 5f had the lowest percentage of coaches conforming.

EXHIBIT 8. PERCENTAGE OF COACHES WITH CONFORMING RESPONSES IN YEAR 5 BY ITEM (N = 47)



Exhibits 9 through 13 present similar information as above, but with the percentage of coaches with conforming responses disaggregated by PD Group, rather than in the aggregate. Exhibit 9 shows that coaches in either PD Group scored similarly on a handful of items such as 3f, 3g, and 5e. PD Group 1 had a higher percentage of coaches conforming to two items, 3h and 5f. PD Group 2 coaches were much more likely to have a higher percentage of conforming responses on several items, including Items 3d, 4a, 4b, 4d, and 4j.

### EXHIBIT 9. PERCENTAGE OF COACHES WITH CONFORMING RESPONSES IN YEAR 1 BY ITEM

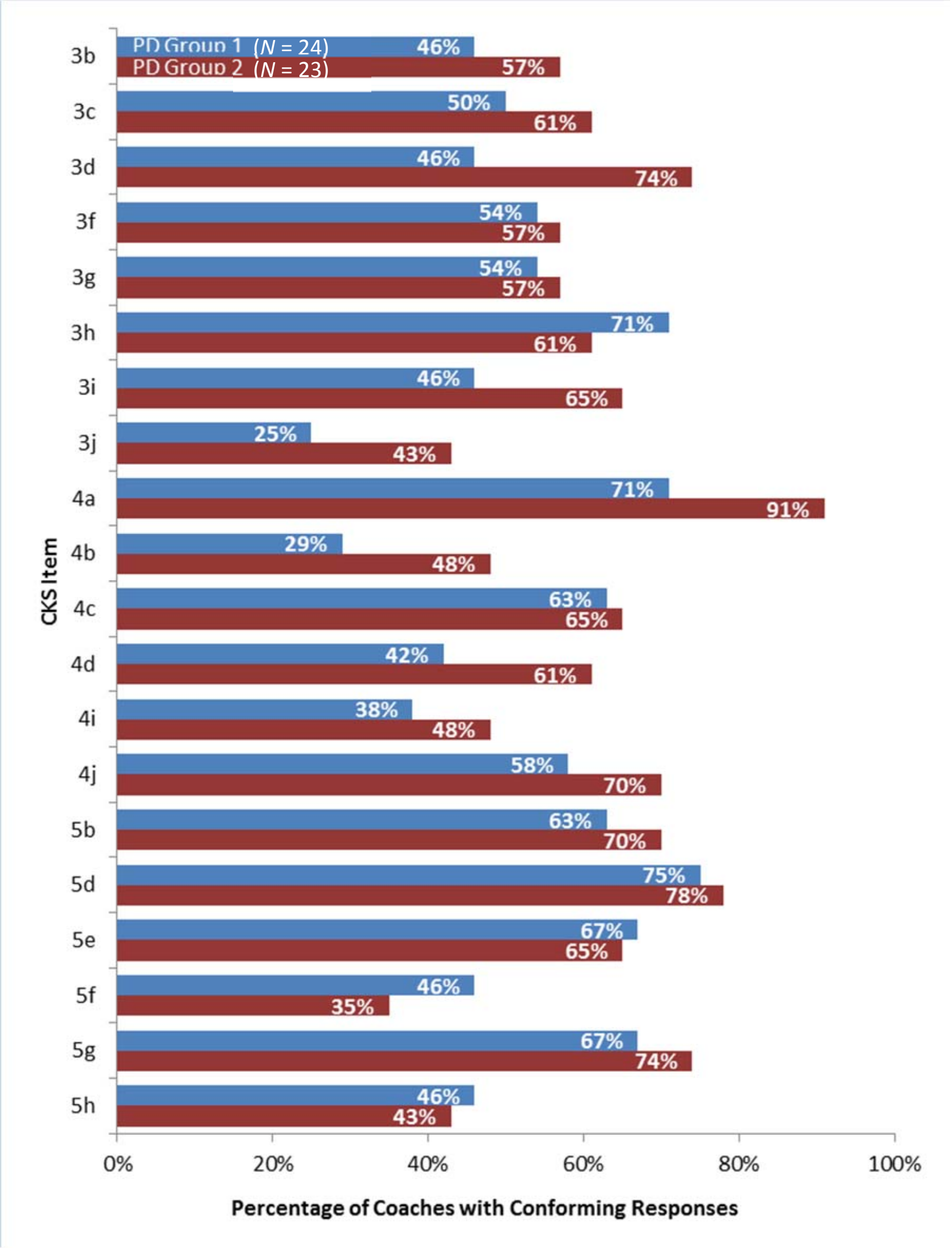


Exhibit 10 shows that the two PD groups had more similar rates of conforming responses in Year 2 than in comparison to Year 1. When large differences were found between the percentages of coaches with conforming responses, PD Group 2 usually had a higher percentage of coaches such as with Items 3d, 4d, 4j, and 5h.  However, Item 4b did have a much larger percentage of coaches in PD Group 1 than PD Group 2 with conforming responses.

EXHIBIT 10. PERCENTAGE OF COACHES WITH CONFORMING RESPONSES IN YEAR 2 BY ITEM

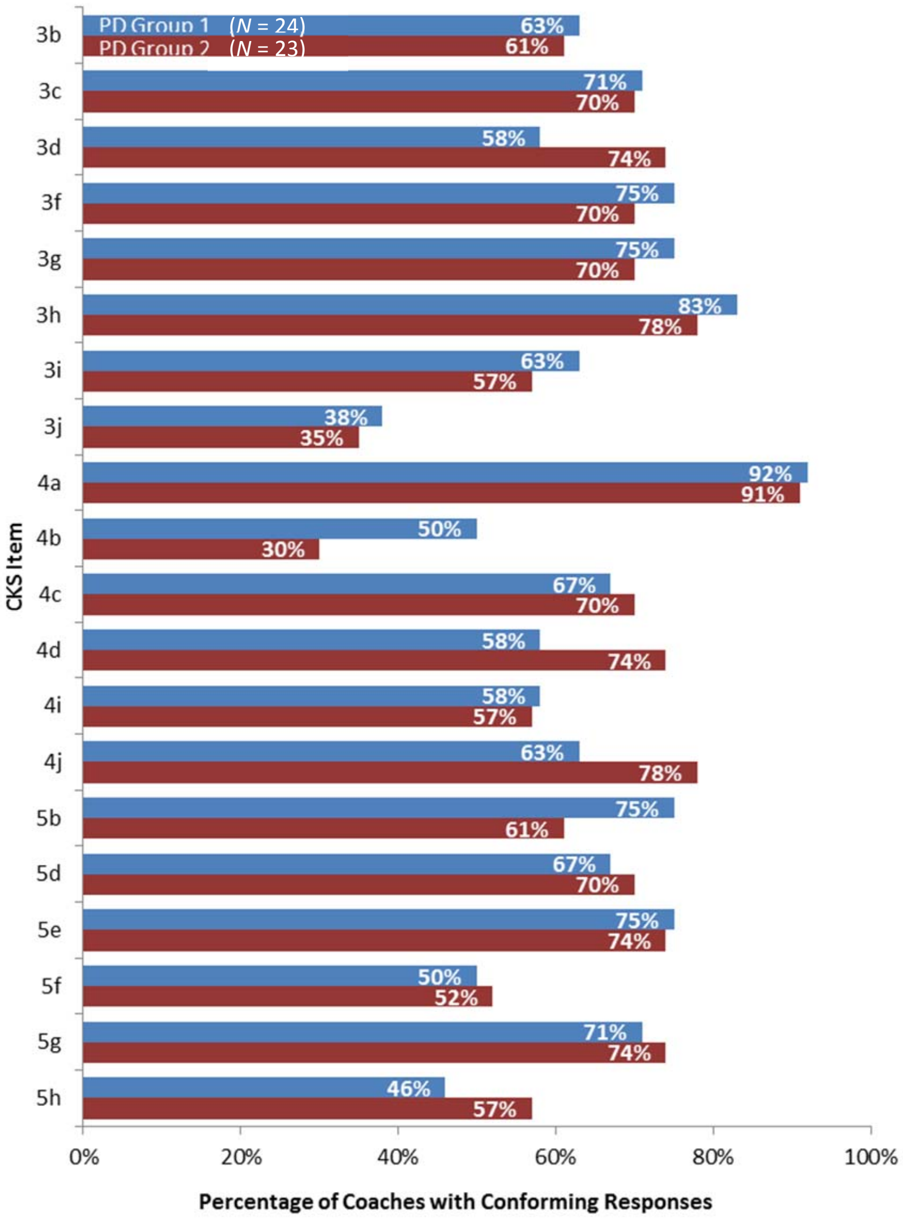
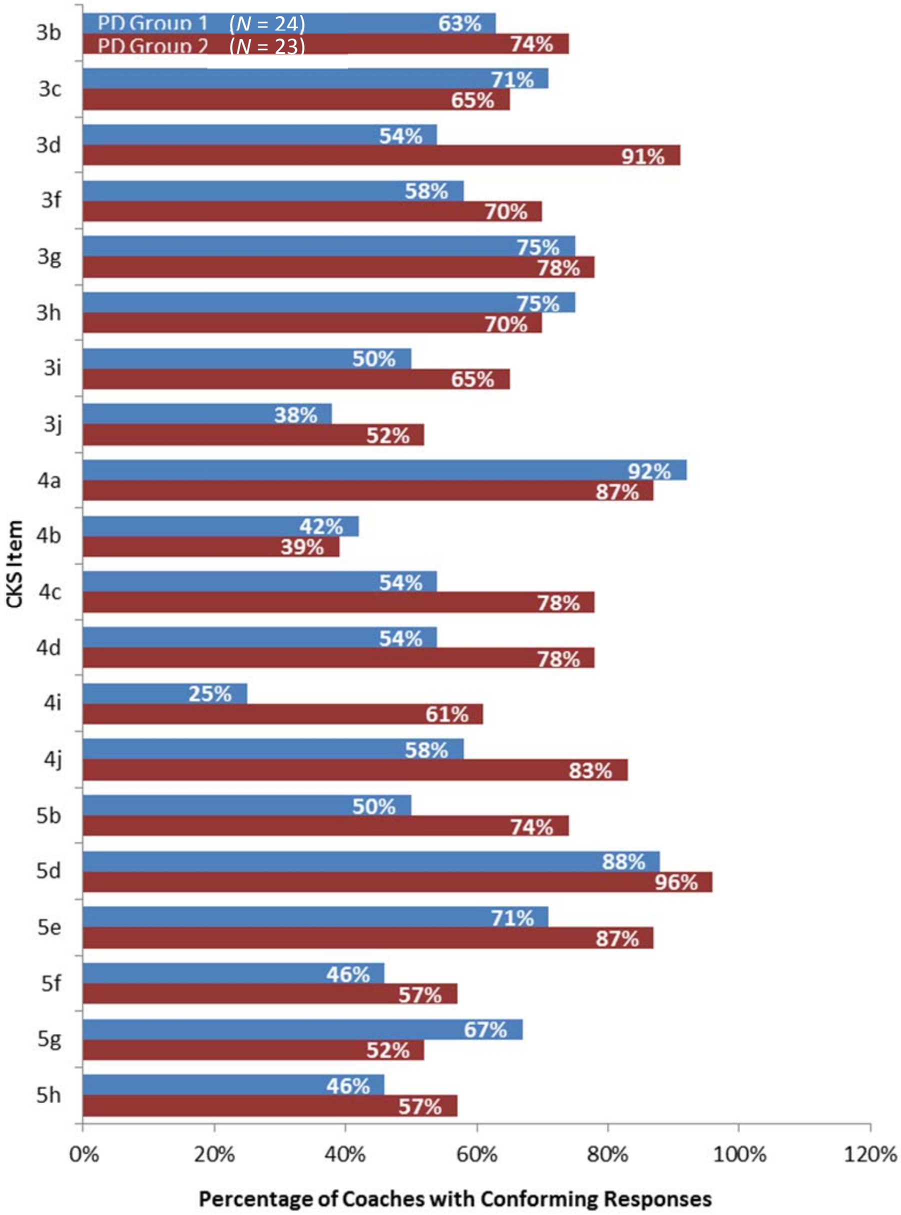


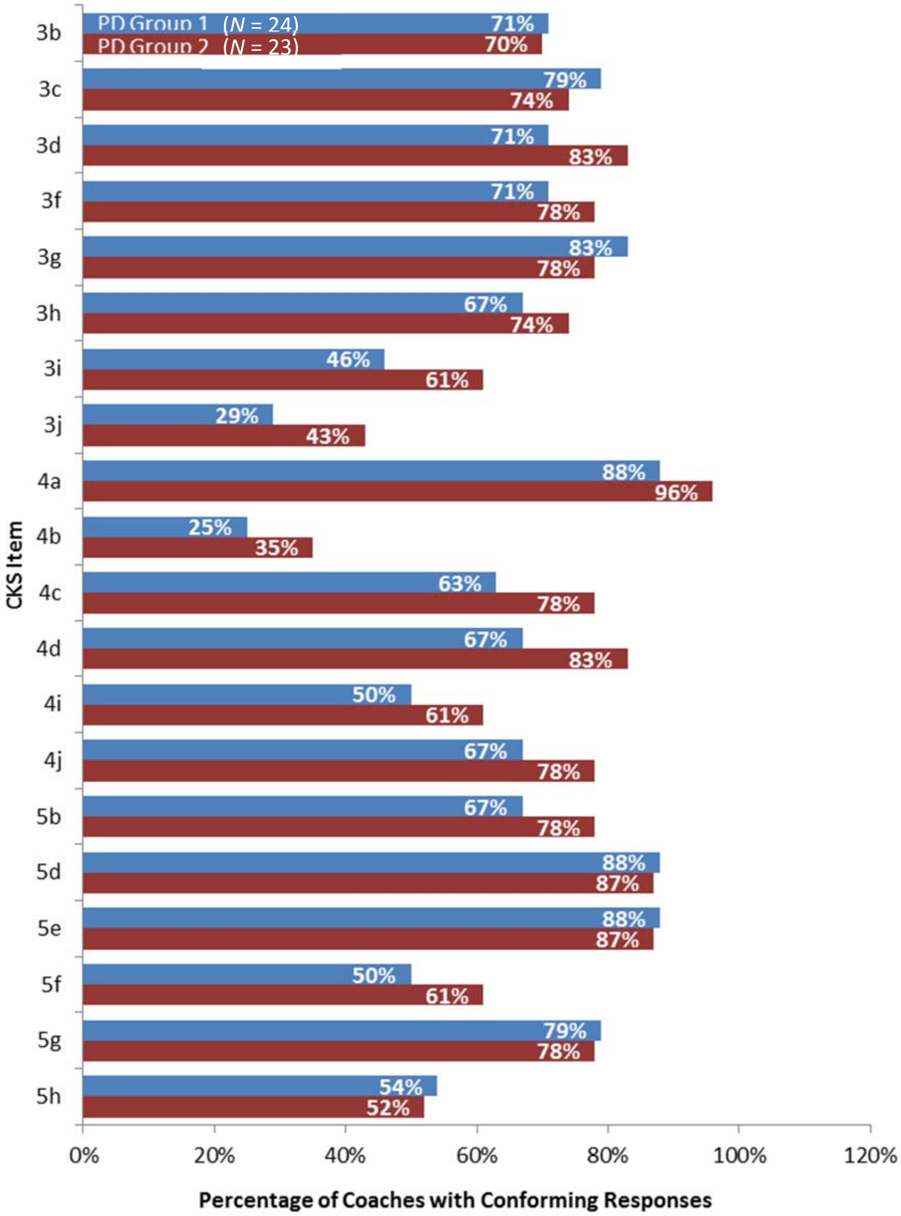
Exhibit 11 displays the percentage of coaches with conforming responses to the CKS in Year 3. The Exhibit illustrates that there were several items in which the PD groups had large differences, such as Item 3d, 4c, 4d, and 4j where PD Group 2 had a larger percentage of coaches conforming to the statement than PD Group 1. When comparing Year 3 percentages with Year 2, PD Group 2 showed increases in the percentage of coaches with conforming responses while PD Group 1 had a decrease to Items 3i, 4b, 4i, 4j, and 5b. Both groups showed an increase for Item 5d and showed decreases for Items 3h and 5g.

### EXHIBIT 11. PERCENTAGE OF COACHES WITH CONFORMING RESPONSES IN YEAR 3 BY ITEM



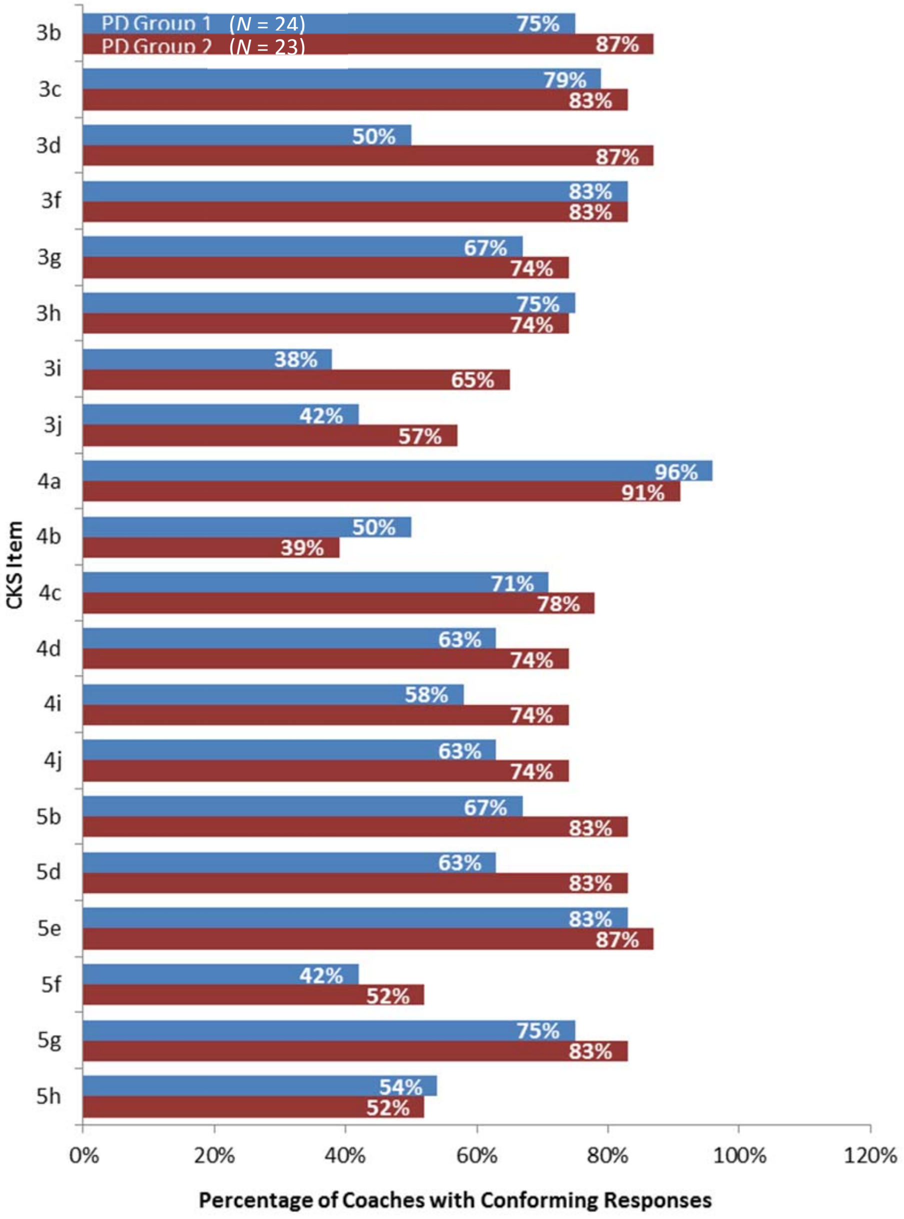
The percentages of coaches in either PD Group with conforming responses to the CKS in Year 4 are displayed in Exhibit 12. PD Group 2 continued to have a larger percentage of coaches with conforming responses than PD Group 1, but there were no longer large differences between PD groups which were present in Year 3. In comparison to the year before, the percentage of PD Group 1 coaches increasing and PD Group 2 coaches decreasing were found for Items 3b, 3d, 4a, 4j, and 5h. Both groups had an increase for Items 3c, 3f, 3g, 4d, 5b, 5f, and 5g and a decrease for Items 3i and 4b.

### EXHIBIT 12. PERCENTAGE OF COACHES WITH CONFORMING RESPONSES IN YEAR 4 BY ITEM



The percentages of Year 5 conforming responses by PD Group are shown in Exhibit 13. At this administration there were slightly larger differences in PD Groups than compared to Year 4. PD Group 2 continued to have a higher percentage of coaches with conforming responses than PD Group 1, a trend that was apparent across all administrations.  Both groups increased from Year 4 to Year 5 on Items 3b, 3f, 3j, 4b, and 4i, and the two groups both decreased on Items 3g, 4d, 4j, 5d, and 5f.  Lastly, PD Group 2 increased while PD Group 1 decreased on Items 3d, 3i, 4a, and 5g.

### EXHIBIT 13. PERCENTAGE OF COACHES WITH CONFORMING RESPONSES IN YEAR 5 BY ITEM


PD Group 1 (N = 24)
PD Group 2 (N = 23)
CKS Item
3b 75% 87%
3c 79% 83%
3d 50% 87%
3f 83% 83%
3g 67% 74%
3h 75% 74%
3i 38% 65%
3j 42% 57%
4a 96% 91%
4b 50% 39%
4c 71% 78%
4d 63% 74%
4i 58% 74%
4j 63% 74%
5b 67% 83%
5d 63% 83%
5e 83% 87%
5f 42% 52%
5g 75% 83%
5h 54% 52%
Percentage of Coaches with Conforming Responses