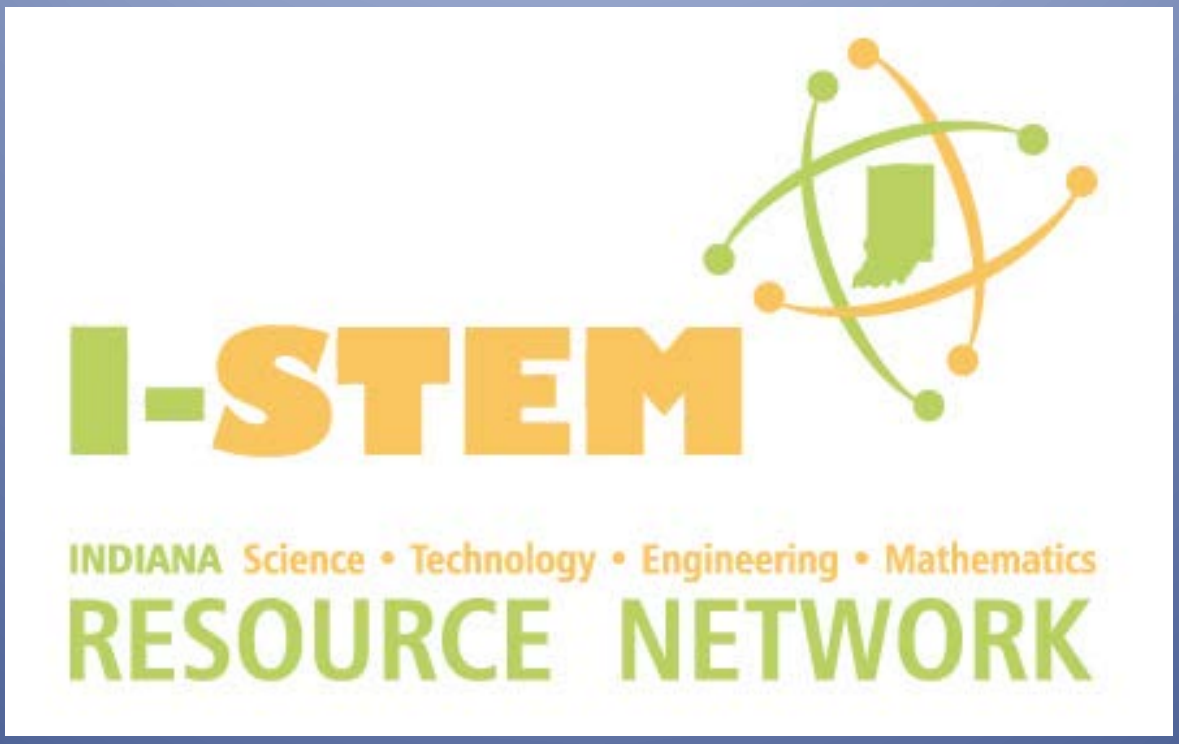
Report 2014-02



Indiana Science Initiative Update: Science Teacher Efficacy Study – TERC Evaluation Report

March 2014

The I-STEM vision is for Indiana to be a national leader in student achievement and to demonstratively improve college and career readiness in the STEM disciplines.

I-STEM Resource Network is supported by the Lilly Endowment, the Lilly Foundation, Biocrossroads, the Indiana Department of Education, and Purdue University.

STEM Education Evaluation Center at TERC:

Improving education through evaluation

Indiana Science Initiative

Science Teacher Efficacy Study

Evaluation Report

Heather Lavigne, Ed.M.
Karen Mutch-Jones, Ed.D.
Lindsay Demers, Ph.D.
TERC

STEM Education Evaluation Center
SEEC at TERC

TERC

2067 Massachusetts Avenue
Cambridge, Massachusetts 02140
617.873.9600 phone
617.873.9601 fax
http://evaluation.terc.edu

Indiana Science Initiative
Science Teacher Efficacy Study
March 2014

Heather Lavigne, Karen Mutch-Jones, Lindsay Demers
TERC

Introduction

With a goal to systemically reform K-8 science education in Indiana, teachers participating in the Indiana Science Initiative (ISI) have been learning to instruct with research-based curricular materials that supported guided inquiry. A science notebooking process was integral to the instruction, enhancing both science learning and literacy. While becoming an ISI teacher provided opportunities, there were also demands associated with adopting new curricula and developing new instructional practices. Teachers had to be open to innovation, persistent when new lessons did not go smoothly, and confident in their ability to make changes to instruction when students did not respond positively. Such characteristics are indicative of a strong sense of efficacy (Jerard, 2007, Protheroe, 2008). Therefore, a longitudinal study measuring ISI teacher efficacy and employing the Science Teaching Efficacy Beliefs Inventory (STEBI, Form A, Riggs and Enochs, 1990) was initiated by I-STEM project staff. In 2011, they collected baseline data from 813 ISI teachers and followed-up with data collection in 2012 and 2013 after these teachers had engaged in both professional development and classroom implementation of ISI materials. As the external evaluators from TERC, engaged in June 2012, we were asked to analyze the STEBI data collected over this three-year span.

According to Riggs and Enochs (1990), the STEBI-A is a valid instrument for the measurement of elementary teacher science efficacy beliefs and measures two distinct subscales: the *Personal Science Teaching Efficacy Belief* (PSTEB) and the *Science Teaching Outcome Expectancy* (STOE) scales. Participating teachers respond to 25 items within this survey instrument.

To check that the STEBI was, in fact, a valid and reliable measure of these two efficacy subscales for this ISI teacher sample, we conducted a confirmatory factor analysis of the baseline data. At the same time, we analyzed these data to identify the efficacy "starting point" of the teachers. Once all data were collected, we analyzed pre to post STEBI scores to measure whether there was change in teachers' sense of efficacy over time.

Confirmatory Factor Analysis

All eight hundred thirteen (813) teachers that were sampled in 2011 were included in this analysis. As shown in Figure 1 in Appendix A, each item on the STEBI pointed to the construct the authors claimed it would measure—either Personal Science Teaching Efficacy Beliefs (PSTEB) or Science Teaching Outcome Efficacy (STOE)—as all factor-loading t-values were significant ($\alpha = .05$). A factor-factor inter-correlation was calculated to determine the strength of the relationship between the PSTEB and STOE factors. Results suggest a small but significant factor correlation, $r = .108$, $t(812) = 2.62$, $p < .05$ (see Table 1 in Appendix A).

Reliability estimates were calculated for each item to determine the degree to which the STEBI produced consistent results. Reliability estimates for items within the PSTEB subscale ranged

from .195 to .530. Reliability estimates for items within the STOE subscale ranged from .109 to .501 (see full results in Tables 2 and 3 in Appendix A). Results suggest that the most reliable indicator for the PSTEB scale was item 22 ("When a student has difficulty understanding a science concept, I am usually at a loss as to how to help the student understand it better."). The most reliable indicator for the STOE subscale was item 15 ("Students' achievement in science is directly related to their teacher's effectiveness in science teaching.").

Composite reliability was also calculated for each latent factor, providing a sense of how well each collection of items does in predicting their latent construct. PSTE had a composite reliability.834. STOEB had a composite reliability calculation of .703. (Formula: $\rho c = (\Sigma\lambda)2 /[\Sigma\lambda 2 + \Sigma(\theta))$.

Validity was also calculated for each item to determine whether the STEBI actually measures the construct of teacher efficacy for these teachers. Estimates for items on the PSTEB ranged from .442 - .729. Estimates for items on the STOE ranged from .329 - .708. Consistent with the reliability results presented above, items that were found to be most reliable for each subscale were also those that were most valid (see Table 4 in Appendix A for full results).

Model fit statistics were also examined for the two-factor model described above. The Minimum Fit Function Chi-Square suggests a poor overall fit to the data ($df = 274$, $\chi 2 = 985.108$, $p < .001$). However, it has been suggested that this fit statistic can be negatively affected by large samples. To account for this issue, we calculated the Normed Chi-Square that divides by the model's degrees of freedom to adjust for sample size (NC = 3.595). Bollen (1990) suggests that Normed Chi-Squares between 2 and 5 suggest reasonable fit. The Root Mean Square Root of Approximation (RMSEA = .0391, 90% CI = 0.569, .0644) and the CFI (.9505) also lend evidence to suggest reasonable model fit.

Factor Analysis Conclusions: The results are fairly consistent with those found by Riggs and Enoch (1990). Overall, it is clear that the STOE subscale is less reliable in measurement as compared to the PSTEB subscale. The authors suggest that items included on the STOE scale may be perceived differently by different educators. Also, the original authors believed that the "internal nature" of the PSTEB items might contribute to higher reliability, such that it is easier for teachers to rate their own behaviors than it is for them to evaluate student outcomes, since the latter may depend on factors external to the one's teaching abilities. In our own STEBI data, we see more consistent and robust changes in the ISI teachers PSTEB scores and more variation in their STOE scores. While the latter may be influenced by several factors, the variation may be partially attributable to differing teacher interpretation of the questions.

This factor analysis also supported the notion that the two subscales are distinct constructs, as indicated by the factor correlation.

Descriptive Statistics for STEBI at Baseline

To get a sense of the starting level in 2011 for the ISI teachers and to ascertain whether we needed to be concerned by floor or ceiling effects, we analyzed all baseline data collected from ISI teachers in 2011. In particular, we had some concern that the STEBI items would sound like "typical" beliefs or behaviors that are expected of all teachers, and thus, the teachers might feel

compelled to rate themselves highly at the start, leaving little room on the scale for higher ratings later if they perceived changes in their sense of efficacy.

Therefore, we generated descriptive statistics for each of the twenty-five STEBI items. These are presented in Table 5 in Appendix A and are sorted by their subscale association. In addition, PSTEB and STOE composite scores were calculated by taking the sum of all appropriate item values for each subscale. This was done after assuring that proper reverse coding was completed on negative worded items. As a 5 was the highest possible score on each item, the highest possible composite scores were 65 for the PSTEB and 60 for the STOE.

Baseline conclusions: At baseline, the full sample of ISI teacher participants averaged 50.15 on the PSTEB (SD = 6.637) and 42.83 on the STOE (SD = 4.548). It was determined from the item-based descriptive statistics, as well as the composite subscale scores, that ceiling/floor effects were not present at baseline, allowing us to identify changes that may occur in teacher perceptions.

Change in ISI Teachers' Sense of Efficacy

To assess pre-post ISI training effects on teachers' ratings on the STEBI, repeated measures ANOVA was conducted with several subsets of the sample. Although the goal was to collect STEBI data from teachers annually, that did not always happen. Some teachers took the STEBI each year (2011, 2012, and 2013) as planned. Others only took it during the 2011 baseline year and then once again, in either 2012 or 2013. In addition, some teachers joined the ISI project after 2011, thus we received baseline data from them more recently. We will follow this latter group and report on their change in efficacy in the future. Below, we report on two sets of teachers for whom we have complete data at this juncture.

Our analyses focused on teacher change in personal efficacy and teaching outcomes expectancy via the PSTE and STOE subscales. A between-subjects factor was also included to assess whether being part of a district that received more intensive ISI support had a unique effect on the change between pre and post scores. These "districts of interest" included, Richmond, Avon, Logansport, and Evansville. We also looked at the relationship between years of experience and teachers' change scores.

Findings for teachers participating in ISI for one year: One-hundred and two (102) teachers provided a baseline STEBI assessment in 2011 and then took part in a post-training STEBI in 2012. Among teachers for which demographics were available (n=82), approximately 55% of teachers had a Master's Degree, 39% had a Bachelor's Degree, and 8% had another type of second-level degree. In regards to ethnicity, 73 percent (n = 71) of these participants were white. Seventy-one percent of the teachers (n = 69) were female. The average number of years of experience (in 2011) was 14.34 (SD = 12.00) with a median of 11.5.

Eighty (80) teachers within this sample were included in a test of pre to post change.[1]

[1] Teachers who did not have a district ID were excluded from the pre-post change analysis due to interest in whether or not the effect of time was contingent on teachers' affiliation with certain districts.

Personal Science Teacher Efficacy: There was a significant effect of time on pre to post PSTE scores, $F(1,78) = 13.61$, $p < .001$), such that, overall, the mean difference averaged 2.91 points higher on post as compared to pre (49.48 to 51.68). The between subjects factor of district was not statistically significant.

Science Teacher Outcomes Expectancy: There was a significant effect of time on pre to post STOE scores $F(1,78) = 10.31$ $p < .001$ such that overall, the mean difference averaged 2.29 points lower on post as compared to pre (40.59 compared to 42.88). The between subjects factor of district was statistically significant, $F(1,78) = 4.42$, $p < .05$, such that both pre and post scores were higher, on average, for the districts of interest (Richmond, Avon, Logansport, and Evansville) at baseline and at post assessment (42.37 versus 44.29 at pre, 39.69 versus 40.59 at post). However, the interaction was non-significant suggesting that the pre-to-post decrease occurs for both types of districts.

We examined whether or not being an experienced teacher influenced efficacy growth as well, but we found no significant correlations years between years of experience and teacher change scores.

Findings for teachers participating in ISI for two years: One hundred and fifty-seven (157) teachers, separate from the analyses presented above, provided a baseline STEBI assessment in 2011 and then took part in a post-training STEBI survey in 2013. Within this group, approximately 52% of teachers have a Master's Degree, 43% have a Bachelor's Degree, and 3% have another second-level degree. In regards to ethnicity, 90 percent (n = 141) of these participants were white. Eighty-six percent of participants (n = 134) were female. The average number of years of experience (in 2011) in this group was 13.74 (SD = 9.90) with a median of 11.5.

*Personal Science Teacher Efficacy (157 teachers)***:** There was a significant effect of time on pre to post PSTE scores $F(1,155) = 20.66$, $p < .001$) such that, overall, the mean difference averaged 2.27 points higher on post as compared to pre (49.04 to 51.25). The between subjects factor of district was not statistically significant.

Science Teacher Expectancy Outcomes (156 teachers): There was a marginally significant positive effect of time on pre to post STOE scores (41.90 to 42.44, $F(1,154) = 3.33$, $p = .07$); however a time by district interaction was present, $F(1,154) = 5.49$, $p = .02$, such that positive change occurred for the districts of interest teachers (n = 56, 41.36 pre versus 44.38 post) versus other district teachers (n = 70, 41.60 pre versus 42.45 post). The main effect of district was also significant, $F(1,154) = 5.77$, $p < .05$.

Once again, the relationship between years of experience and teachers' change scores were examined. As was true with teachers in the previous analysis, no significant correlations were present.

Summary of Efficacy Study Findings

As a result of our factor analysis, we feel reasonably confident that the Science Teaching Efficacy Beliefs Instrument (STEBI) is reliable and valid measure, appropriate for gauging the

level of teaching efficacy for the participating Indiana Science Initiative teachers. However, we continue to note that the STOE subscale items may have been interpreted less consistently by the ISI teachers.

Our findings suggest that participating teachers sense of ***personal teaching efficacy*** (their confidence in their own teaching abilities) grew while participating in ISI professional development and employing ISI-selected curricula and teaching practices. This significant increase in PSTE scores occurred whether teachers were involved in ISI for one or two years. While there may have been other things that occurred during this time period (e.g., school initiatives) and contributed to this growth, we found that neither one's years of teaching experience nor being part of a district of interest (Richmond, Avon, Logansport, and Evansville) influenced this positive change.

The results for ***teaching outcomes expectancy*** (teachers' beliefs about whether student learning can be influenced by effective teaching) were more complicated. For teachers who participated in one year of ISI, we found a significant decrease in their STOE scores over this one-year period. This negative change was not apparent, however, for those who participated for two years. Instead, there was a positive trend in the pre to post data that was approaching significance. Furthermore, for teachers who were part of our districts of interest, we found significant positive change in teaching outcomes expectancy. There continued to be no relationship between level of teaching experience and change in efficacy scores.

While the decrease in STOE results for one year ISI teachers was initially surprising, conversations with I-STEM staff and ISI teachers during the first year of the project point to a possible explanation: As teachers increased their science pedagogical content knowledge and raised their expectations of student sense-making, they became more cognizant of the skill and time it takes to truly support student learning and growth, and thus, they moderated their responses on the post-survey, rating themselves lower on the scale. This explanation is strengthened by an additional analysis conducted on data from those 2-year ISI teachers who had three data points (baseline STEBI in 2011, another in 2012, and then a final post score in 2013). For this group, we saw a similar pattern of a higher score at time one, a lower score—potentially "adjusted" to match their heightened awareness—at time 2, and then a higher score at time 3. The change from time points 1 to 3 was trending positive and approaching significance, even thought there was a drop at mid-point.

These data also suggest that when these teachers participated in ISI for a longer period of time and when the PD support was intensified (as with the districts of interest), their sense that they could impact student learning was strengthened.

Further study of the trajectory of teacher change over a longer period of time, and the factors that support and inhibit change at various points would provide valuable context for these STEBI findings.

References

Bollen, K. (1990). Overall fit in covariance models: Two types of sample size effects. *Psychological Bulletin,* 107(2), 256.

Jerald, C. D. (2007). *Believing and achieving (Issue Brief).* Washington, DC: Center for Comprehensive School Reform and Improvement.

Protheroe, N. (2008). Teacher efficacy: What is it and does it matter? *Principal,* Retrieved on February 5, 2014 at https://www.naesp.org/resources/1/Principal/2008/M-Jp42.pdf.

Riggs, I.M. & Enoch, L.G. (1990). Toward Development of an Elementary Teachers Science Teaching Efficacy Belief Instrument. *Science Education*, 74(6), 625-637.

Appendix A
Statistics for Confirmatory Factor Analysis

Figure 1. STEBI-A Factor Analysis: Two Factor Model

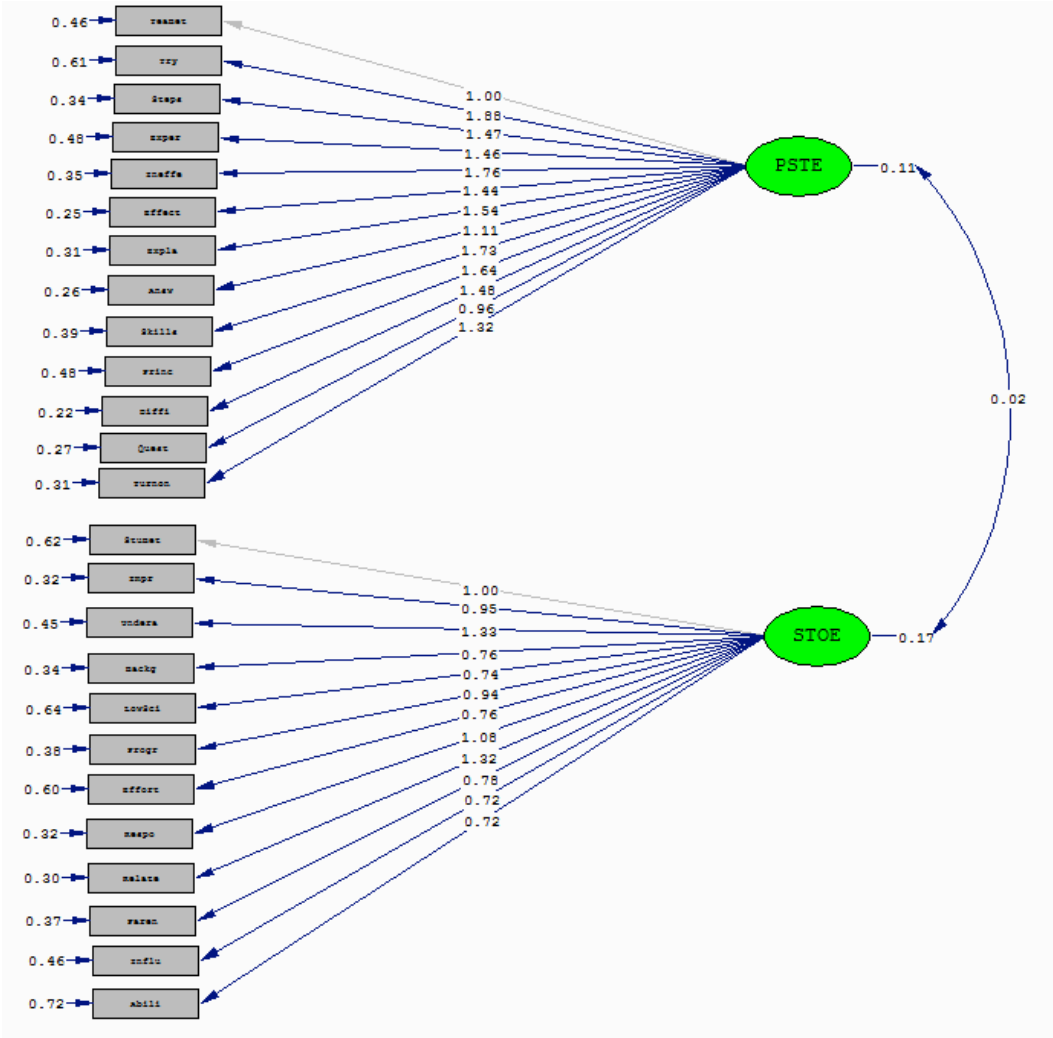


Table 1. Factor-factor Inter-correlations

Factor #1		Factor #2			
Factor Name	Estimated Variance	Factor Name	Estimated Variance	Estimated Factor-factor Covariance	Estimated Factor-factor Correlation
PSTEB	.112	STOE	.171	.015	.108

Correlation = r = cov/SD1*SD2 = .015/(√ .112*.171) = .015/(.335*.414) = .108

Table 2. PSTEB Indicator Reliabilities (Item Reliability Estimates)

		Estimated Indicator Parameters		
Indicator (Item#)	Observed Variance	Error Variance	True Variance	Reliability
TeaBet (2)	.574	.462	.112	.195
Try (3)	1.007	.610	.397	.394
Steps (5)	.579	.340	.239	.413
Exper (6)	.724	.484	.240	.331
Ineffe (8)	.699	.353	.346	.495
Effect (12)	.484	.252	.232	.479
Explan (17)	.580	.315	.265	.457
Answ (18)	.396	.259	.137	.346
Skills (19)	.721	.385	.336	.466
Princ (21)	.780	.478	.302	.387
Diffi (22)	.462	.217	.245	.530
Quest (23)	.376	.274	.102	.271
Turnon (24)	.510	.314	.196	.384

Reliability = True/Observed

Table 3. STOE Indicator Reliabilities (Item Reliability Estimates)

		Estimated Indicator Parameters		
Indicator (Item#)	Observed Variance	Error Variance	True Variance	Reliability
StuBet (1)	.790	.618	.172	.218
Impr (4)	.474	.320	.154	.325
Undera (7)	.757	.455	.302	.399
Backg (9)	.436	.336	.100	.230
LowSci (10)	.738	.643	.095	.129
Progr (11)	.530	.380	.150	.283
Effort (13)	.694	.596	.098	.141
Respo (14)	.524	.322	.202	.385
Relate (15)	.597	.298	.299	.501
Paren (16)	.480	.375	.105	.219
Influ (20)	.551	.464	.087	.158
Abili (25)	.807	.719	.088	.109

Reliability = True/Observed

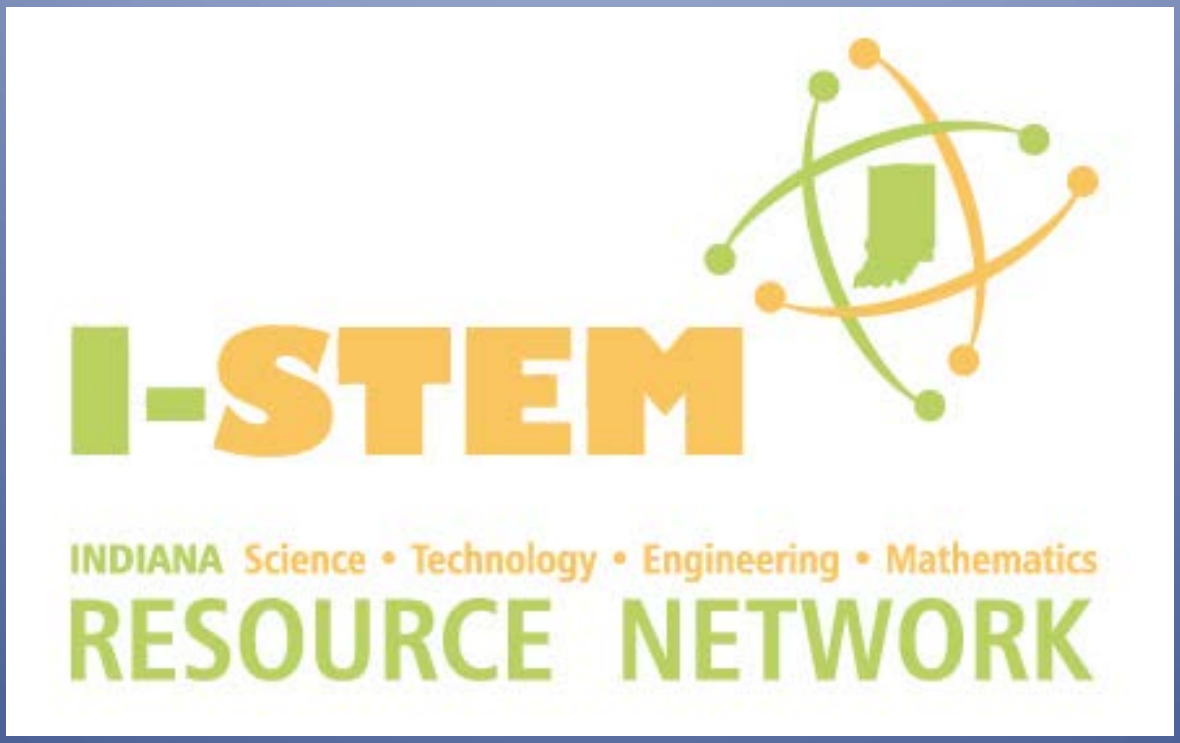
Table 4. Indicator-factor Correlation (Validity Estimates)

			Estimates		
Indicator (Item#)	Observed Indicator Variance	Factor	Factor Variance	Factor Loading	Indicator-factor Correlation
TeaBet (2)	.574	PSTE	0.112	1.000	.442
Try (3)	1.007			1.884	.628
Steps (5)	.579			1.465	.644
Exper (6)	.724			1.465	.576
Ineffe (8)	.699			1.759	.704
Effect (12)	.484			1.442	.694
Explan (17)	.580			1.542	.678
Answ (18)	.396			1.110	.590
Skills (19)	.721			1.733	.683
Princ (21)	.780			1.644	.623
Diffi (22)	.462			1.480	.729
Quest (23)	.376			0.956	.522
Turnon (24)	.510			1.323	.620
StuBet (1)	.790	STOEB	.171	1.000	.465
Impr (4)	.474			0.949	.570
Undera (7)	.757			1.328	.631
Backg (9)	.436			0.764	.478
LowSci (10)	.738			0.743	.358
Progr (11)	.530			0.938	.532
Effort (13)	.694			0.757	.375
Respo (14)	.524			1.085	.620
Relate (15)	.597			1.322	.708
Paren (16)	.480			0.784	.468
Influ (20)	.551			0.716	.399
Abili (25)	.807			0.715	.329

Indicator– Factor correlation: Loading*(√(Factor variance/Indicator variance)

Table 5. Descriptive Statistics for Baseline 2011 STEBI-A Item Responses

PSTEB		STOE	
Item (# on STEBI)	Mean (SD)	Item (# on STEBI)	Mean (SD)
TeaBet (2)	4.06 (.785)	StuBet (1)	3.72 (.875)
Try (3)	3.31 (1.02)	Impr (4)	3.90 (.578)
Steps (5)	3.65 (.753)	Undera (7)	3.28 (.791)
Exper (6)	3.64 (.915)	Backg (9)	3.88 (.622)
Ineffe (8)	3.88 (.792)	LowSci (10)	2.80 (.882)
Effect (12)	4.00 (.712)	Progr (11)	3.66 (.691)
Explan (17)	3.80 (.674)	Effort (13)	3.67 (.803)
Answ (18)	3.92 (.622)	Respo (14)	3.57 (.677)
Skills (19)	3.78 (.889)	Relate (15)	3.52 (.759)
Princ (21)	3.81 (.964)	Paren (16)	3.60 (.607)
Diffi (22)	3.91 (.711)	Influ (20)	3.77 (.719)
Quest (23)	4.38 (.648)	Abili (25)	3.45 (.947)
Turnon (24)	4.01 (.702)		



For more information, please contact:
Paul J. Ainslie, Ph.D.
Managing Director, I-STEM Resource Network
Purdue University
Mann Hall B041
203 S. Martin Jischke Dr.
West Lafayette, IN 47907
Office: 765-494-0557 Mobile: 317-531-7301