NURTURES Evaluation Report:

5 Year Summary
2011 –2016

Prepared by:

Acumen Research and Evaluation, LLC
1811 N. Reynolds Road, Suite 204
Toledo, OH 43615
419-265-1811

"Accuracy of observation is the equivalent of accuracy of thinking."--Wallace Stevens

On the behalf of:

THE UNIVERSITY OF TOLEDO 1872

National Science Foundation

Nurtures

Table of Contents

This report summarizes the activities and finding of the evaluators of the NSF MSP project at The University of Toledo entitled NURTURES from March 2011 through February 2017.

I. Executive Summary

The NURTURES project, consisting of three phases, was implemented per its original timeline. Hiring, planning, piloting, and implementation were all completed in a timely manner with regards to the original timeline. Elements of the project were modified for improvement based upon reflection and formative evaluation findings.

During the Planning Phase (Year 1), all staff members were hired and plans for recruiting, the Summer Institute, academic year follow-up, community events, research and data collection, and family packs were completed. The Planning Phase concluded having met its objectives. The groundwork for a successful Math Science Partnership (MSP) was laid for the Year 2 Training Phase. Hallmarks of the success of this Phase include:

- Adherence to timeline
- Frequent meetings that built teamwork and group identification
- In-depth planning for the Pilot Phase
- Adherence to staff and teacher qualifications when hiring as mapped out in the proposal

The Pilot Phase (Year 2) allowed the NURTURES team to test essential elements of the project prior to scaling up. During Year 2 the NURTURES partners met for a retreat that brought together the key players and familiarized them with the complementary learning paradigm as well as specific elements of the project. Piloting the Summer Institute provided an opportunity for the project leaders to try out various methodologies, schedules, and activities that were considered for inclusion in the actual Summer Institute that was offered the following year. Eight Master Teachers attended the Year 2 Summer Institute that included ideas for the actual Institute. Feedback from the Master Teachers was incorporated into the Summer Institute model. SciFUN Community Events were also piloted in Year 2. Families were enthusiastic about both the events and the Family Packs.

Process evaluation was an integral part of the Implementation Phase (Years 3–5). Findings verified fidelity of implementation of the various components of this comprehensive project and provided internal validity evidence. Corrections or modifications to implementation were made due to early detection of variance from implementation plans. For example, facilitators from SciFUN event hosts did not engage with families as designed during Year 3. Project leaders used process evaluation findings to better train facilitators with the result of close adherence to the event intentions.

The effect of participating in NURTURES on teachers was cataloged using several established instruments. Measures of teacher attitudes about teaching science showed that teachers' comfort with teaching science increased after participation as did their tendency to equate effective science instruction with cognitive learning theory. On the other hand, it was found that teaching inquiry-based science was more challenging after participating in NURTURES. This is most likely due to their realization that in order to teach inquiry-based science is more work, at least initially, than business as usual.

Classroom observations of teachers teaching science were conducted pre and post Summer Institute participation. Observations were scored using the Electronic Quality of Inquiry Protocol created through the project "Inquiry in Motion" at Clemson University. This tool was based upon the Next Generation Science Standards (NGSS). A chi-square test of independence compared pre- and post-frequencies and found a statistically significant change ($p < .001$)

suggesting that teachers were incorporating significantly more scientific inquiry into their practice than they did prior to participation in NURTURES. This finding supports teacher FOI of inquiry-based instructional strategies and verifies the obtainment of the improved science teaching objective.

Family science engagement was evaluated through observations of families in the SciFUN community events. Summative data show that the majority of family interactions were either balanced between the child and parent, or the parent slightly dominated the parent-child interaction. This is to be expected when working with young children as the children may need more guidance for some activities. Discourse between parent and child included question/short answer, why questions, and open-ended questions. Parent made frequent use of the SciFUN Event Guides to get ideas of questions to ask their children.

Improved student learning was measured using a standardized test of early literacy, mathematics, and reading (STAR K-12[1]). A hierarchical linear design (HLM) was employed to examine the differences in achievement scores between students of NURTURES teachers and students of control teachers in multiple participating schools (Toledo Public (OH), NW Ohio Parochial, Monroe County Schools (MI) and various preschools in NW Ohio. Participants consisted of 2899 students for the early literacy study, 2002 students for the mathematics study, and 1810 students in the reading study. All schools were high-needs (See Exhibit A in Appendix). Control students consisted of 2515 students for the early literacy study, 3028 students for the mathematics study, and 2448 students for the reading study, who had never had a NURTURES teacher within the same time frame. Results showed statistically significantly higher scores for NURTURES students on all three measures.

The summative examination of measures of outcomes compared to intentions showed that NURTURES achieved all of its outcomes and its overall goal of improving student learning. While the original intention was to improve student science learning specifically, there were no standardized, rigorous measures of science achievement available for the target young students. As a result, mathematics, reading, and early childhood literacy were substituted under the assumption that reading and mathematics achievement in particular are highly correlated with science achievement. Findings from the five-year evaluation indicate that NURTURES is a successful intervention for improving science teaching in the early childhood classroom as well as for increasing family science participation and the quality of that participation.

[1]STAR Assessments by Renaissance Learning are short adaptive tests that were administered to students in 41 schools in grades K-3. Subject areas include early literacy, mathematics, reading.

II. Planning Phase (Year 1)

Evaluation of the Planning Phase consisted of an examination of inputs listed on the project logic model. These inputs included hiring employees, recruiting graduate assistants, and identifying education, science, and engineering faculty members from The University of Toledo to participate in the five-year project. Additional measures used to examine fidelity of implementation and to identify potential roadblocks to success included observing and reviewing the Planning Team's group process procedures and outcomes. Data were collected through direct observation of the retreat, examination of the Summer Institute Master Teacher pilot, personal interviews with members of Planning Team, and examination of recruiting materials and efforts. Project documentation showed attainment of all Planning Phase outcomes within a timely manner. All employees were hired and trained as scheduled.

Evaluators made direct observations of several of the planning team's weekly meetings. Tangible outcomes, group progress, and group dynamics were examined. The team achieved the tangible outcomes expected during this phase. Interactions between members continually improved over the course of the year as group members became more familiar with one another. As far as group dynamics is concerned, the group worked as a cohesive entity towards mutual goals. The frequency of the meetings contributed to the collegiality among group members as they quickly came to identify with the group.

The evaluators conducted personal interviews with members of the Planning Team to further examine the group process—particularly from each member's perspective. Again, strong evidence of mutual respect and an atmosphere conducive to positive interaction and progress was provided. All appreciated that the meetings were well-organized and little time was wasted.

Master Teachers were recruited and hired. The group represented teachers from grades preschool through grade 3 (PreK-3) with experience teaching science and coaching or providing other teachers with professional development. Summer Institute surveys from the Implementation Phase verified that the selection of Master Teachers was appropriate for the success of the program.

Overall, Phase I was concluded having met its objectives. The groundwork for a successful MSP was laid for the Year 2 Training Phase. Hallmarks of the success of this Phase include:

- Adherence to timeline
- Frequent meetings that built teamwork and group identification
- In-depth planning for the Pilot Phase
- Adherence to staff and teacher qualifications when hiring as mapped out in the proposal

III. Training/Pilot Phase (Year 2)

During this phase, the project team piloted components of the program that were to be implemented during the Implementation Phase the following year. The Training Phase included the 2012 Summer Institute planning retreat and implementation as well as the planning retreat for the 2013 Summer Institute. NURTURES staff also piloted several community science events.

The 2012 Summer Institute (SI) was an opportunity for the project leaders to try out various methodologies, schedules, and activities that were considered for inclusion in the actual

Summer Institute that was to be offered in the Implementation Phase. Two measures were used to provide formative evaluation feedback to the project team: an observation of the 2012 planning retreat and a focus group interview with the participating teachers (n = 8). The retreat brought together the key players who helped implement the complementary learning paradigm reflecting the NURTURES mission and goals.

Partners present at the retreat included NURTURES leadership (n = 4), scientists and engineers (n = 7), prek-3 teachers (n = 8), and NURTURES staff (n = 6) including graduate assistants (n = 8). The retreat consisted of several segments: an icebreaker, science immersion (where attendees were introduced to PreK-3 science), piloting of several family backpack activities (participants in small groups tried and then commented on the activities including ease of instructions, age appropriateness, and application to classroom science), piloting of a potential community family event (sponsored by Challenger Space Center), an introduction to the Summer Institute and the opportunity to work on a team to draft a plan for one week of the Institute.

The formative evaluation of the retreat was based upon theory of the cyclical nature of interpersonal collaboration (Gajda & Koliba, 2007). Components of the cycle include dialogue, decision making, action (moving beyond planning), and evaluation (review or reflection). Ranking of the collaboration ranged from network (little interaction, no group identity, no common goal) to professional learning community (all members interact, group identity is strong, evidence of a common goal).

The evaluation examined the extent to which the retreat facilitated a professional learning community as evidenced by the amount and type of interaction experienced in the group as a whole as well as in the smaller break-out groups. There was evidence within each of the retreat's events of interpersonal collaboration that reflected high levels of teamwork. Members of the groups felt free to present ideas and all members participated in group activities. For events that required a consensus, the groups engaged in conversation and made decisions as a group. All members participated in the group decision making. The 2012 planning retreat was effective in its goal of preparing the partnership for the Summer Institute through group activities. The retreat offered the opportunity for preexisting groups to connect and for partnerships beyond pre-existing groups to be forged.

The four week 2012 SI provided the eight Master Teachers with a glimpse of what the actual SI would be like. Both the mornings and afternoons were broken into two or three sessions. Each session had a focus: pedagogy, science immersion, metacognition, and collaboration component. Teachers appreciated that each block of time (morning/afternoon) included an activity. The teachers admitted they became so engaged they lost track of time. Feedback from the teachers provided the NURTURES team with formative evaluative data that included:

- Provide Institute participants with optional iPad training prior to participation.
- To facilitate and clarify the education sessions, combine pedagogy and metacognition into one lesson or show clear linkages between the two.
- Science content sessions should increase time spent with scientists and engineers.

These suggestions were incorporated during the planning for the first SI offered in the Implementation Phase. Also piloted during this period were community science events and the Family Packs (take home science experiments). Formative data collected from families as they

experienced both activities provided valuable information regarding refining both the Packs and the community events.

The Training/Pilot Phase met its goal of furthering the planning of the SI and academic year follow up as well as continuing to build a strong MSP. Families were enthusiastic about the community events and the Family Packs. The partnership between NURTURES and its informal science organizations was balanced and open. Hallmarks of the success of the Training/Pilot Phase included:

- Adherence to timeline
- Continuance of regular team meetings
- Open communication and clear expectations provided to informal science partners
- Collection of formative feedback from relevant stakeholders and the implementation of suggestions for improvement

IV. Implementation Phase (Years 3 – 5)

The final three years of the project comprised the Implementation Phase where activities piloted during the Training Phase were implemented, reviewed, and revised as needed. The first year of Implementation (2013-14) recruited 40 participants and the subsequent two years recruited 146 (137 teachers and 9 administrators) and 141 (134 teachers and 7 administrators) respectively. There were two major outcomes identified in the logic model for the Implementation Phase as illustrated in Figure 1 (Outcome column):

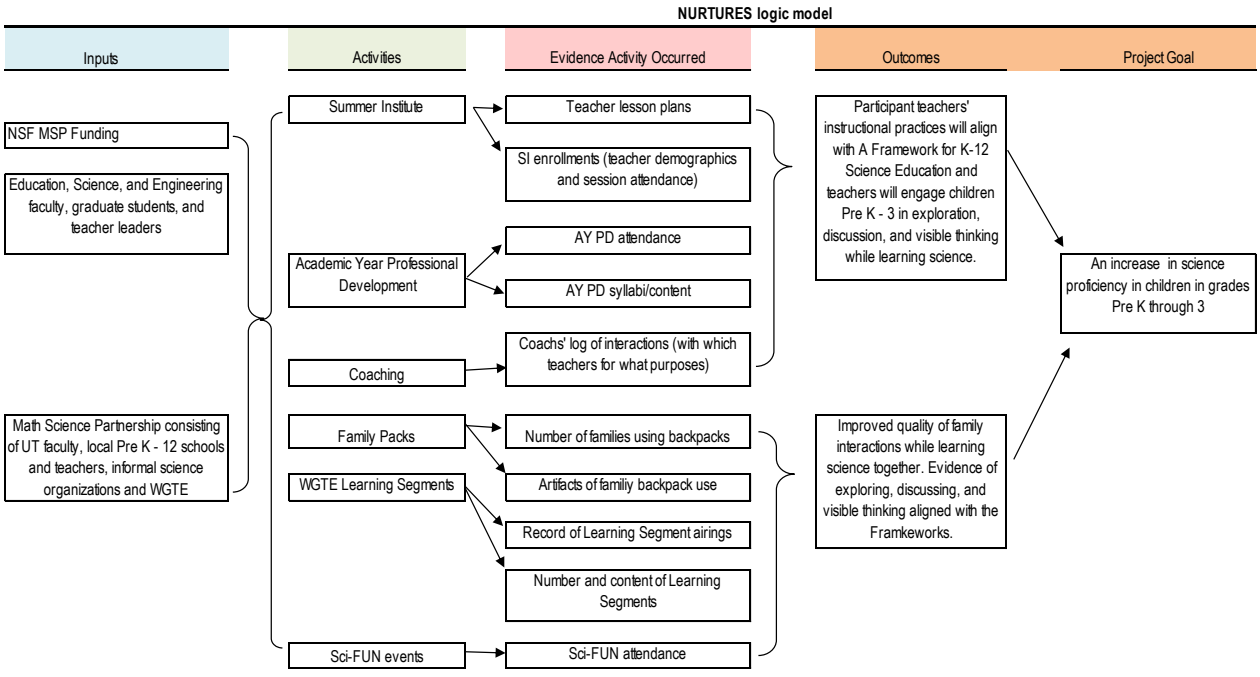


Figure 1. NURTURES logic model.

Several variables were measured within those two outcomes to verify outcome achievement. Table 1 provides these variables and their corresponding measures. Details about the instruments can be found in NURTURES annual reports.

Table 1

Implementation Phase Outcomes, Variables, and Measures

Teacher Outcome		
Method	**Variable Measured**	**Purpose**
Observations of participants teaching science coded with the EQUIP	Implementation of inquiry-based science teaching	To gain a better understanding of how PreK-3 teachers teach science, to compare teaching with Summer Institute goals, to examine change over time
Preschool Teachers Attitudes & Beliefs about Science (P-TABS)	Teachers' attitudes and beliefs toward science teaching	To determine whether participation in NURTURES alters teachers' attitudes about teaching science
Teacher Beliefs about Effective Science Teaching (TBEST)	Effects of professional development on teacher behavioral beliefs about science teaching	To determine whether participation in NURTURES alters teachers' beliefs about science teaching
Family Science Outcome		
Community science event activity observations using revised project developed observation rubric	In general, how families do science together; what families discuss when doing science together	To understand the quality of family interactions while learning science together; to determine the extent to which Community event instructions and materials support family science
Project Goal (Student Outcome)		
STAR achievement tests	Student academic growth/proficiency	To determine effects of NURTURES on children's academic growth

A. Process Evaluation

The implementation phase logic model served as a tollgate to making causal conclusions regarding the relationship between the implementation of the project activities or events and outcome attainment. A process evaluation, based upon the logic model, provided evidence of fidelity of implementation (FOI) and as such verifies summative evidence of the effectiveness of the program. Process evaluation provides evidence of internal validity—a necessary step prior to drawing conclusions about causal relationships. Elements of the process evaluation included observations of the SI sessions, observations of the SciFUN community events, observations of

teachers implementing inquiry-based science teaching in the classroom, and a comparison between-project inputs and the logic model and timeline.

Observations of both the SI and SciFUN events focused on instructor/staff implementation of the intervention with regards to project defined criteria. Evaluation data collected during the Training Phase provided project leaders with assessments of the implementation of the interventions including areas of inconsistency. Additional information to inform leaders about FOI was collected during the Implementation Phase and adjustments were made by project leaders when necessary.

To ensure FOI of the SI, evaluators were embedded in each of the courses. Observations were scored using the *Local Systemic Change through Teacher Enhancement Professional Development Observation Protocol* developed by Horizon Research, Inc. This tool includes five major elements of math/science teacher professional development: design, implementation, content, pedagogy, and culture. With regards to providing quality professional development and adherence to fidelity of implementation of professional development best practices, the NURTURES SI met and exceeded minimum standards as outlined by Banilower, Boyd, Pasley, & Weiss (2006). Details of the FOI for the SI can be found in the 2016 NURTURES Annual Evaluation Report.

To assess FOI of the SciFUN events, in depth case notes were taken and analyzed using a phenomenological approach to uncover common practices among those who facilitated the events (Patton, 2015). Three phases of coding were employed (Corbin & Strauss, 2015). Open coding examined practices for themes both anticipated (relevant to project delineated practices) and unanticipated. Open coding was reviewed to determine dominant themes and then axial coding mapped relationships between codes. Finally, selective coding was applied to describe the common experiences and relate these to project guidelines. During the first year of Implementation (Year 3) several discrepancies were observed particularly among volunteer facilitators. These facilitators were associated with the host organization but did not receive prior training or information regarding their role in the SciFUN activities. Frequently the volunteers were observed "taking over" the activity by working directly with the child or by completing the activity for the child thereby eliminating the opportunity for the family science experience. Project leaders reflected upon the information gathered in Year 3 and made significant changes by providing more NURTURES staff to assist at events and informing host organizations of the expectations for their facilitators. They were then encouraged to "hand pick" volunteers and provide them with the expectations prior to their event. As a result, Years 4 and 5 saw close adherence to SciFUN FOI.

Teacher implementation of inquiry-based science instructions was measured using several instruments outlined in the next section (B Teacher Outcome).

B. Teacher Outcome

Year 3 data collection was not as comprehensive as the two subsequent, full-scale years as evaluation measures were continued to be piloted during Year 3. Teachers completed the P-TABS and the TBEST all three years to measure teacher perceptions and self-efficacy regarding teaching science. These instruments are explained in detail in previous NURTURES annual evaluation reports. During Year 3 the instrument used for classroom observation was the *Ohio Continuum of Teacher Development* that is based on the Ohio Standards for the Teaching

Profession. A modified version of the instrument was used for the evaluation that included only those categories that could be assessed through observation of teaching. Elements of teaching included Lesson Delivery, Differentiation, Resources, Environment, and Assessment. During Years 4 and 5, the *Electronic Quality of Inquiry Protocol* created through the project "Inquiry in Motion" at Clemson University was used for classroom observations. This tool was based upon the Next Generation Science Standards and proved to be a better tool for the evaluation of NURTURES teaching goals. Again, the Ohio Continuum and EQUIP are explained in detail in previous reports.

1. Teacher attitudes (P-TABS and TBEST): Table 2 illustrates the results by year of the P-TABS and TBEST. In Year 3 the tests were administered as pre/post participation in the SI. Years 4 and 5 included three administrations to extend the examination to long-term effects (pre, 3 months post, and 6 months post). In general, teacher attitudes about teaching science did not change as measured on the P-TABS. However, because scores were higher than the expected mean on the pretest, it is likely that there was a little room for improvement (ceiling effect). In Year 4, however, there was a statistically significant improvement in scores on both the comfort and challenges scales that occurred at the first posttest and remained consistent at follow up.

Table 2

Summary of P-TABS and TBEST Scores Years 3 - 5

Test	Variable	Year 3	Year 4	Year 5
P-TABS*	Comfort teaching science	$p > .05$	$p < .05$	$p > .05$
	Challenges teaching science	$p > .05$	$p < .05$	$p > .05$
	Benefits to child of learning science	$p > .05$	$p > .05$	$p > .05$
TBEST	Learning theory	$p < .05$	$p > .05$	$p < .05$
	Confirmatory science	$p > .05$	$p > .05$	$p > .05$
	Hands-on science	$p > .05$	$p < .05$	$p > .05$

*All administrations of the instrument realized scores that exceeded the expected mean score indicating that the teachers scored higher than expected on the pretest.

On the TBEST, the *learning theory scale* measured the degree to which the teachers agreed with effective science instruction aligned with cognitive learning theory. *Confirmatory science* measures preference for instruction that controls the student so that only the "correct" outcome occurs and *hands-on all the time* is a measure of preference for instruction that does hands-on activities for the sake of hands-on activities. It is the goal of NURTURES that the learning theory scores will increase while the other two scale scores decrease. As Table 2 shows, learning theory did indeed increase in Year 3 and Year 5. The increase in the hands-on science scale in Year 4 could have been due to the fact that while teachers saw the value in engaging students in their learning, they had not yet reached the level where they were selective in the types and frequency of hands-on activities. This information was shared with project leaders and more emphasis on the role of hands-on activities to make science relevant was included in the final year. As a result, there was no statistically significant change in that construct in Year 5.

2. Teaching practice: To determine the degree to which teachers adhered to NURTURES intended inquiry-based teaching strategies and Next Generation Science Standards, pre- and post-participation SI classroom observations of a random sample of teachers teaching science in their classrooms were conducted. As noted earlier, during Year 3 the Ohio Continuum for Teacher Development was used while Electronic Quality of Inquiry Protocol (EQUIP) was used in Years 4 and 5. To measure long-term effects, pre-participation observations of the Year 5 participants who were repeating from Year 4 were also examined.

In Year 3, 27 teachers were randomly selected for observation and scored using the Ohio Continuum. The scores are provided in Table 3 and are listed by category. It should be noted that not every category was scored for every teacher so row totals may not always equal 27. Year 3 observations provided the project leaders with feedback on what areas the teachers in general had mastered and where more professional development emphasis was needed. Assessment was clearly an area that needed improvement; however, a third of the teachers scored in the Developing category for lesson delivery and differentiation also. Using this information, project leaders adapted the Year 4 and 5 SIs to assist teachers in these areas.

Table 3

Summary of Ohio Continuum of Teaching Scores

Category	Ineffective	Developing	Proficient	Accomplished
Lesson Delivery	0	9	16	1
Differentiation	0	9	13	4
Resources	0	5	12	7
Environment	0	6	15	6
Assessment	2	9	13	3
Total	**2**	**38**	**69**	**21**

During Years 4 and 5 the EQUIP was used to score observations. The EQUIP rubric measures four factors associated with inquiry instruction and is based upon NGSS—instruction factors, discourse factors, assessment factors, and curriculum factors. Within these four factors are 19 indicators. Scores on pre- and post-participation observations were compared first to determine if there were patterns of proficiency among the time-based observations (process evaluation) and second to determine if there were areas of improvement between the observations. The rubric included four levels—pre-inquiry, developing, proficient, and exemplary. In addition to the EQUIP, observers scored NGSS practices using a checklist. Observed instances of all of the scientific practices increased from pre to post observations. A chi-square test of independence compared pre and post frequencies and found a statistically significant change ($p < .001$) suggesting that teachers were incorporating significantly more scientific inquiry into their practice than they did prior to participation in NURTURES. This supports teacher FOI of inquiry-based instructional strategies.

The evaluators randomly selected 24 teachers in Year 4 and 59 teachers in Year 5 for observation and a dependent t-test was conducted to test the null hypothesis that there was no statistically significant change in teachers' EQUIP observation scores after participation in NURTURES. As noted in earlier reports, ordinal ranking scores were converted to an interval

scale using the Rasch Rating Scale measurement model. Table 4 shows that there were statistically significant improvements of scores on each of the four scales between the pre and post assessment indicating that participation in NURTURES improved teachers' inquiry-based instruction.

Table 4

*Pre/post Comparison of EQUIP Observation Scores (*n = *83)*

Scale	*t*	*p*
Instruction	4.623	<.0001
Discourse	3.701	<.0001
Assessment	6.067	<.0001
Curriculum	3.877	<.0001

A subsample of teachers attended the 2016 SI ($n = 40$) provided a third observation in spring 2016 allowing for an examination of the sustained impact of the NURTURES program. A repeated measures ANOVA results provided in Table 5 indicate that teachers improved between the pre- and the first post- assessment, but there was no statistically significant change in teaching practice between the fall and subsequent spring observations.

Table 5

*Repeated Measures Comparison of EQUIP scores (*n = *40)*

Scale	Comparisons		
Instruction	1 < 2 ($p = .002$)	1 < 3 ($p < .0001$)	2 = 3 ($p > .05$)
Discourse	1 < 2 ($p < .0013$)	1 < 3 ($p < .0001$)	2 = 3 ($p > .05$)
Assessment	1 < 2 ($p < .0001$)	1 < 3 ($p < .0001$)	2 = 3 ($p > .05$)
Curriculum	1 < 2 ($p = .004$)	1 < 3 ($p < .0001$)	2 = 3 ($p > .05$)

This suggests that the gains made in the Fall 2015 were maintained through Spring 2016 and that teachers continued to include inquiry-based instructional best practices in their science teaching.

C. Family Science Learning

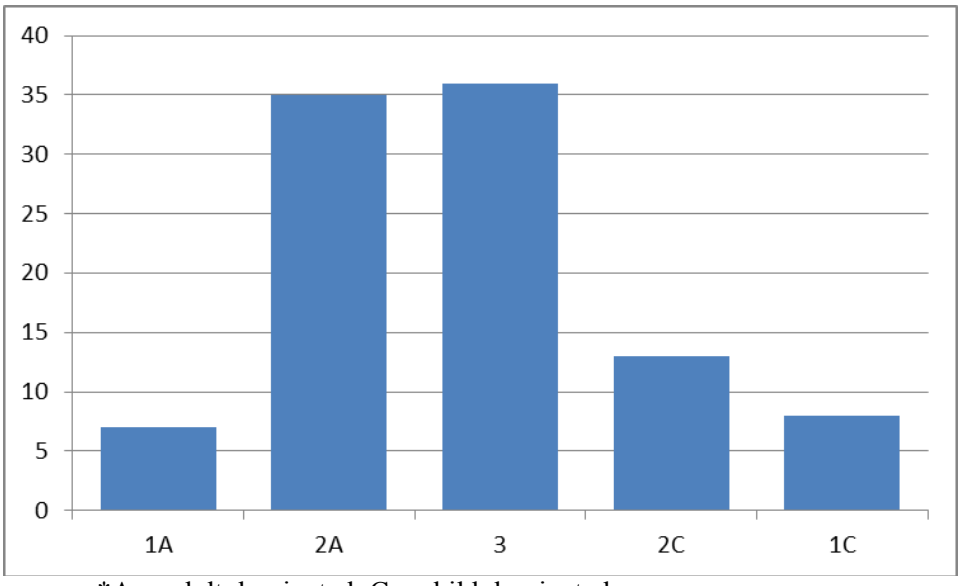
The NURTURES Family Packs and SciFUN events were aligned with NGSS and provided inquiry-based science and engineering family activities to improve quality of family interactions while learning science together. Evaluation looked for evidence of exploring, discussing, and visible thinking aligned with the *A Framework for K-12 Science Education* [*Framework]*.

1. Family Packs: Review of the activities included in the Family Packs verified adherence to NGSS and the *Framework*. Measuring impact on families, however, proved elusive. The intention was to observe families as they completed family packs and score using a project-designed rubric based upon intended discourse and interaction outcomes. In Year 3, 18

observations were scheduled—3 PreK, 8 K, 2 1st-grade, 3 2nd-grade, and 2 3rd-grade. Of the 18, only five families actually met with observers. Findings from the observations demonstrated that parents interacted to varying degrees; however, all adhered to the activity guides and followed the instructions. In general, parents did not require more than a one word response from their child as they explored the activity. In Year 4, to recruit families, flyers went home to one classroom per grade level (PreK-3) or five classes of about 22 students each. Of the over one hundred flyers sent home, six families were willing to participate. However, three of those families did not show up for their sessions resulting in an unrepresentative sample. No families were observed during Year 5. Although impact of the Family Packs was not verified through the evaluation, a dissertation was completed as part of NURTURES research showed the packs encouraged positive family interaction (Strickler-Eppard, 2016). The evaluation did determine that the Packs included age-appropriate inquiry-based activities and the instructions/guides that accompanied them promoted family discourse and learning.

2. SciFUN Community Events: Year 3 SciFUN observations collected the data necessary to develop an observation rubric used during Years 4 and 5 (Discourse and Inquiry in Family Science--DIFS). This tool, explained in detail in the 2015 and 2016 Evaluation Reports, measured interaction levels between adult and child, recorded the level of discourse between adult and child (no discourse, question/short answer, questions that require an explanation, and open-ended questions or discussion), and documented parent/adult use of recommended Talking Tips (Kaderavek, North, Rotshtein, Dao, Liber, Milewski, Molitor, & Czerniak, 2015; Michaels, Shouse, & Schweingruber, 2007). Combined with case notes, the tool provided a rich description of how families engaged in the SciFUN event activities.

Seventy-four observations were made in Year 4 at three events and 62 observations were made in Year 5 at three events. Specific findings per event/year and activity are chronicled in the respective annual evaluation reports. The summative ratings are provided in Figures 1 (interaction) and 2 (discourse).



*A = adult dominated; C = child dominated.

Figure 1. NURTURES Summative SciFUN Interaction Scores*

The majority of interactions observed were balanced or with some adult dominance but clear participation from both the adult and the child was evident indicating that the families worked together as they participated in the activities. Discourse scores indicated that conversations were occurring but that the question/short answer pattern was the most prevalent. Many parents were observed attempting to ask questions that elicited more than a short or one word response but they were often happy to get any answer from their child. Perhaps they did not have the skill, experience, or patience to move beyond the question/answer exchange to move to a discussion. While level 1 was the most frequently observed type of discourse, more complex discourse that encouraged open ended questioning and thoughtful responses was also clearly evident.

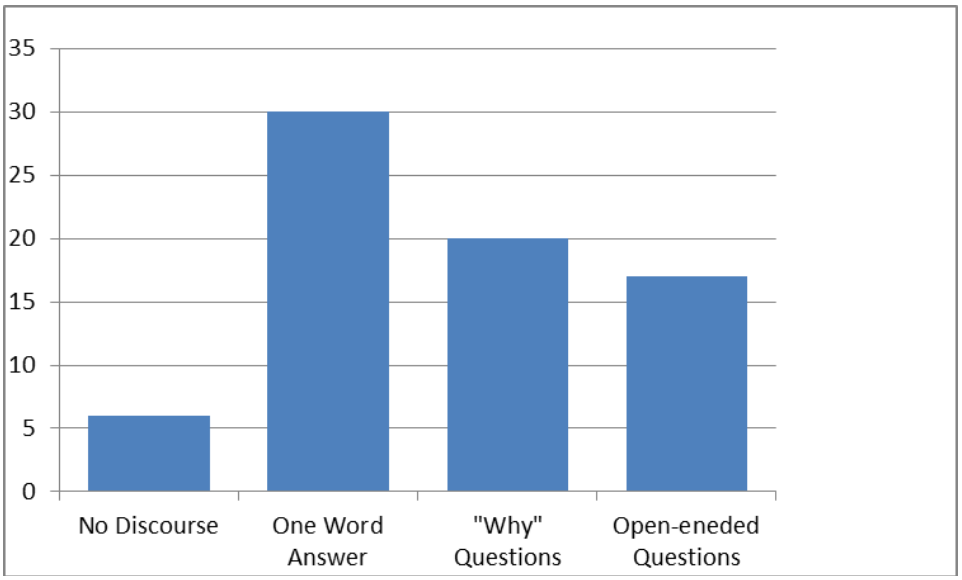


Figure 2. NURTURES Summative SciFUN Discourse Scores

The event guides included many of the following Talking Tips targeted to specific activities:

1. Re-voice what the young scientist said.
2. Ask young scientist to restate someone else's reasoning.
3. Ask young scientist to apply his/her own reasoning to someone else's reasoning.
4. Prompt young scientist for further participation.
5. Ask young scientist to explicate his/her own reasoning.
6. Use wait time.
7. Encourage young scientists to extend what they learn to other situations.
8. Encourage your young scientist to observe.
9. Encourage your young scientists to make predictions.
10. Encourage your young scientist to figure out how to solve a problem.
11. Encourage your young scientists to talk about what they observed.
12. Encourage your young scientist to provide a reason for their answer.

Figure 3 illustrates the dispersion of Talking Tips observed during observation. Talking Tip 4 was the most frequently noted and an examination of the Event Guide suggestions

revealed that parents were often encouraged to nudge their children to investigate further on most activities. A close second, Talking Tip 5, was also frequently highlighted in the Event Guide. This suggests that parents are willing to interact with their children as they do science together but in general need some guidance. That parents frequently implement the suggestions from the Event Guide indicates that when given assistance, they take it thereby reinforcing the conclusion that the Guides are useful. Talking Tip 7 was not observed; however, upon closer examination of the SciFUN activities, none really offered the opportunity for transfer of learning to other situations.

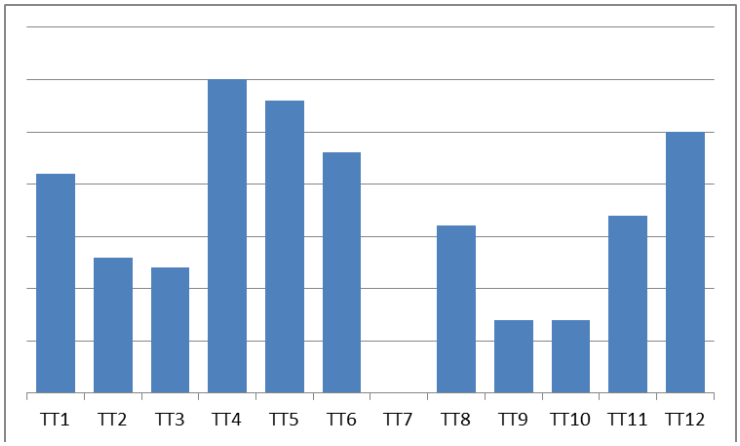


Figure 3. Talking Tips dispersion during SciFUN events

D. Student Academic Achievement

During the course of the project, the STAR K-12 assessments of students were used to measure effects of NURTURES on student achievement. This measure was explained in detail in the Year 4 Evaluation Report; however, to summarize the STAR tests consist of operational items that align to a set of skills derived from exemplary state standards as well as the Common Core State Standards and current research. The STAR assessments have a high degree of internal consistency; overall, it's .85 for Early Literacy and .97 for both Math and Reading. It also has strong predictive validity evidence (Brown & Coughlin, 2007).

This aspect of the evaluation/assessment used a quasi-experimental design with control and treatment students drawn from students at from 41 participating elementary schools. Treatment participants consisted of students who had at least one NURTURES teacher during Years 3, 4 and/or 5 of the project (teachers' participation in NURTURES could have occurred in any of those academic years). Participants consisted of 2899 students for the early literacy study, 2002 students for the mathematics study, and 1810 students for the reading study. Control students consisted of 2515 students for the early literacy study, 3028 students for the mathematics study, and 2448 students for the reading study, who had never had a NURTURES teacher within the same time frame. *Comparisons were made using a binary approach* (students either had the intervention at some time during the study (treatment) or they did not (control)).

1. Method: A two-level hierarchical model was used to assess the equivalency between the treatment and control cohorts; three separate analyses were performed for the three respective

years. The first level equation that predicted individual student's mean achievement included the intercept value and the participant's gender, ethnicity, intervention type variables and a random error component. To capture the effects of individual schools, a second-level, unconditional equation was added. The intercept and intervention type coefficients were considered random, and the effects of gender and ethnicity were considered fixed at the school level. The results for the treatment type coefficients for all three years indicated no statistically significant difference between the groups: $t(40) = -6.66$, $p = .242$ for Fall 2013; $t(40) = 3.32$, $p = .359$ for Fall 2014; and $t(40) = 0.87$, $p = .777$ for Fall 2015 data. The weighted average absolute value effect size for intervention (Hedges' g) was .047, which is considered to be a negligible effect size, so no statistical correction for baseline was used during subsequent data analyses.

Although a multivariate approach to the dependent variables is possible, the present study focused on the analysis of one outcome variable at a time. Therefore, the first-level of data consists of repeated observations of the assessment data in one domain (a level-one unit) nested within a specific student (a level-two unit). Students in turn are nested within schools (a level-three unit).

At the first level equation, the individual student mean achievement was predicted from one time-variant variable: grand-mean centered testing occasion (levels: 0 = Fall 2015, 1 = Winter 2015, and 2 = Spring 2016). The first-level equation included student's intercept (mean value of student achievement) and his/her slope or individual growth trajectory over the measurement occasions, plus a random error interpreted as a residual temporal variation. At the second-level, the estimated coefficients (intercepts and slopes) from the first-level equations became the solutions to two equations, one that modeled student's mean achievement or $\pi 0jk$ and another one that modeled student average learning rate or $\pi 1jk$. Both second level equations included time-invariant student-level variables: grand-mean centered current grade (2, 3, and 4 for mathematics; 1, 2, and 3 for reading; K, 1, and 2 for early literacy); gender (levels: 0 = female and 1 = male); minority status (levels: 0 = minority or and 1 = non-minority or White); and intervention (levels: 0 = absence or 1 = presence of intervention teacher in prior measurement occasions). The current grade variable was considered a time-invariant because the assessment data utilized the latest, 2015-2016 academic year data. The effects of schools, a potential confounding variable, were modeled with the third-level equations. The third-level equations were unconditional or did not include school-context variables.

The equations below depict the specification of the model at each of the three levels. The variance components were specified as random at a student-level. With respect to the school-level, the effects of gender and minority status are assumed to be invariant between schools, while the effects of the current grade and intervention are assumed to be random.

Level-1 Model

$$SCALEDSCijk = \pi 0jk + \pi 1jk*(OCCASIONijk) + eijk$$

Level-2 Model

$$\pi 0jk = \beta 00k + \beta 01k*(CURRENTGjk) + \beta 02k*(GENDER\_Mjk) + \beta 03k*(MINORITYjk) + \beta 04k*(I\_Tjk) + r0jk$$

$$\pi 1jk = \beta 10k + \beta 11k*(CURRENTGjk) + \beta 12k*(GENDER\_Mjk) + \beta 13k*(MINORITYjk) + \beta 14k*(I\_Tjk) + r1jk$$

Level-3 Model

$\beta 00k = \gamma 000 + u00k$
$\beta 01k = \gamma 010 + u01k$
$\beta 02k = \gamma 020$
$\beta 03k = \gamma 030$
$\beta 04k = \gamma 040 + u04k$
$\beta 10k = \gamma 100 + u10k$
$\beta 11k = \gamma 110 + u11k$
$\beta 12k = \gamma 120$
$\beta 13k = \gamma 130$
$\beta 14k = \gamma 140 + u14k$

Mixed Model

SCALEDSCijk = γ000 + γ010*CURRENTGjk + γ020*GENDER_Mjk + γ030*MINORITYjk + γ040*I_Tjk + γ100*OCCASIONijk + γ110*OCCASIONijk*CURRENTGjk + γ120*OCCASIONijk*GENDER_Mjk + γ130*OCCASIONijk*MINORITYjk + γ140*OCCASIONijk*I_Tjk + r0jk + r1jk *OCCASIONijk+ u00k + u01k *CURRENTGjk + u04k *I_Tjk + u10k *OCCASIONijk + u11k *OCCASIONijk*CURRENTGjk + u14k *OCCASIONijk*I_Tjk + eijk

2. Findings: Baseline equivalence was established by examining the fall scores for the STAR Early Literacy assessment for kindergarteners in the study for 2013-2014, 2014-2015, and 2015-2016. A two-level hierarchical model was used to assess the equivalency between the treatment and control cohorts; three separate analyses were performed for the three respective years. The results for the treatment type coefficients for all three years indicated no statistically significant difference between groups: $t(40) = -6.66$, $p = .242$ for Fall 2013; $t(40) = 3.32$, $p = .359$ for Fall 2014; and $t(40) = 0.87$, $p = .777$ for Fall 2015 data. The weighted average absolute value effect size for intervention (Hedges' *g*) was .047, which is considered to be a negligible effect size, so no statistical correction for baseline was used during subsequent data analyses. Results of the baseline analyses indicated group equivalency.

Statistically significant differences between students who had a NURTURES teacher and those who did not were found on each of the scales—Early Literacy, Mathematics, and Reading. Test statistics and a technical report can be found in the Appendix. Learning from a NURTURES teacher was associated with student net gains of 11.2 points to a student's STAR Early Literacy spring score, 21.8 points to a student's STAR Mathematics spring score, and 47.9 points to a student's STAR Reading spring score compared to students who had never had a NURTURES teacher. In addition, the 47.9 points in STAR Reading translated to an effect size of 0.29, a level considered *substantively important* by the What Works Clearinghouse evidence standards (U.S. Department of Education, 2013). Based upon these findings, it can be concluded that the NURTURES teacher professional development and Math Science Partnership reached its goal of improving student learning outcomes in the participating high needs schools. All results of the student outcome analyses can be found in the Appendix.

V. Conclusion

A comparison of method/instrument, results, and findings are provided in Table 6.

Table 6: NURTURES outcomes and results

Teacher Outcome		
Method	**Results**	**Findings**
Observations of participants teaching science coded with the EQUIP	Statistically significant increases on EQUIP score (all 4 scales) between pre and post intervention ($p < .001$).	Participant teaching practices align with *A Framework for K-12 Science Education* including exploration, discussion, and visible thinking.
Preschool Teachers Attitudes & Beliefs about Science (P-TABS)	Statistically significant increase in comfort teaching science after participating in NURTURES.	NURTURES teachers have more confidence teaching science as a result of participation in the MSP.
Teacher Beliefs about Effective Science Teaching (TBEST)	Statistically significant gains in NURTURES teachers' alignment of instruction with cognitive learning theory.	NURTURES teachers have a better understanding and acceptance of cognitive learning theory as a result of participating in the MSP.
Family Science Outcome		
Community science event activity observations using revised project developed observation rubric	Family interaction scores were at an acceptable level of balanced to slightly dominated by parent or child. Discourse was predominantly observed at question/short answer; however there was evidence of more complex levels of discourse. Parents made good use of Event Guide suggestions to incorporate Talking Tips into conversations.	Quality of family engagement was medium to high. Observations revealed that the more a family participated in the events, the better the quality of engagement. Events were well attended indicating that this aspect of the MSP was successful in engaging families in science activities and exploration.
Project Goal (Student Outcome)		
STAR achievement tests	Students who had a NURTURES teacher scored higher on posttests in early literacy, reading, and mathematics on the STAR K-12	Student learning in NURTURES teachers is more advanced than students in non-NURTURES classrooms indicating successful overall goal attainment.

VI. References

Banilower, E. R., Boyd, S.E., Pasley, J. D., & Weiss, I. R. (2006) Lessons from a decade of mathematics and science reform: A capstone report for the local systemic change through Teacher Enhancement Initiative. Report prepared for the National Science Foundation. Accessed online August 17, 2015 at: http://www.horizonresearch.com/pdmathsci/htdocs/reports/capstone.pdf.

Brown, R. S., & Coughlin, E. (2007). The Predictive Validity of Selected Benchmark Assessments Used in the Mid-Atlantic Region. Issues & Answers. REL 2007-No. 017. Regional Educational Laboratory Mid-Atlantic.Corbin, J. & Strauss, A. (2015). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (4th ed.). Thousand Oaks, CA: Sage

Gajda, R., & Koliba, C. (2007). Evaluating the imperative of intra organizational collaboration: A school improvement perspective. *American Journal of Evaluation*, *28*(1), 26–44.

Kaderavek, J. N., North, T., Rotshtein, R., Dao, H., Liber, N., Milewski, G., Molitor, S. C., & Czerniak, C. M. (2015). SCIIENCE: the creation and pilot implementation of an NGSS-based instrument to evaluate early childhood science teaching. *Studies in Educational Evaluation, 45*, 27–36.

Michaels, S., Shouse, A. W., & Schweingruber, H. A. (2007). *Ready, set, SCIENCE!: Putting research to work in K-8 science classrooms.* Washington, DC: National Academies Press.

Strickler-Eppard, L. (2016). A detailed analysis of family utilization of science activity packs (Unpublished doctoral dissertation). The University of Toledo, Toledo, OH.

U.S. Department of Education's Institute of Education Sciences (IES). (2013). *What Works Clearinghouse™ Procedures Handbook*.

VII. APPENDIX

Student Learning Outcome Statistics

Results

STAR Early Literacy

The student mean achievement expressed as a γ_{000} (third-level equation intercept coefficient) for the STAR Early Literacy model was 650.18 (see Table 1). This coefficient represented an average, predicted Winter 2015 score for a minority, female 1st grade student who had never had a [program] teacher. This model predicted outcome was affected in a statistically significant way by the following student-level that the demographic variables: *current grade, gender, and minority status*. As expected, students' scale scores increased by 101.43 with an increase in *current grade* expressed as the γ_{010} coefficient (i.e., moving from grade one to grade two) when controlling for the effects of *gender*, *minority status* and *intervention*. The effect of *gender* (the γ_{020} coefficient) on mean achievement status was statistically significant with female students outscoring male students by an average of 14.73 units. Also, a statistically significant effect for *minority status* (the γ_{030} coefficient) on mean achievement was observed, with non-minority students scoring, on average, an additional 15.61 units in comparison to minority students. This final effect, however, has to be interpreted cautiously in the absence of student's socio-economic status information.

Table 1

Summary of Three-Level Exploratory Model for STAR Early Literacy Achievement

Fixed Effect	B	*SE* B	*t*-ratio	df	*p*
Model for average status, π_0					
Model for mean-status of 1st grade minority female who did not have intervention teacher, β_{00}					
Average mean status, γ_{000}	650.18	7.35	88.43	41	<0.001
Model for current grade, β_{01}					
Current grade, γ_{010}	102.43	2.03	50.37	41	<0.001
Model for gender, β_{02}					
Gender, γ_{020}	-14.73	2.34	-6.29	5982	<0.001
Model for minority status, β_{03}					
Minority status, γ_{030}	15.61	3.07	5.09	5982	<0.001
Model for intervention, β_{04}					
Intervention, γ_{040}	11.24	4.80	2.34	41	.024
Model for learning rates, π_1					

Model for learning rates of 1st grade minority female who did not have intervention teacher, $\beta 1_0$					
Average learning rate, γ_{100}	68.15	1.70	39.98	41	<0.001
Model for current grade, β_{11}					
Current grade, γ_{110}	-19.26	2.00	-9.62	41	<0.001
Model for gender, β_{12}					
Gender, γ_{120}	0.80	1.55	0.52	8989	.606
Model for minority status, β_{13}					
Minority status, γ_{130}	0.13	1.31	0.10	8989	.919
Model for intervention, β_{14}					
Intervention, γ_{140}	-2.39	2.14	-1.12	41	.270

Most importantly, the *intervention* variable had a statistically significant impact on students' scores (see the γ_{040} coefficient). Adding a [program] teacher to a student's academic history was associated with an average increase of 11.24 units in mean student achievement, controlling for the effects of the *current grade*, *gender* and *minority status* variables. This effect size (Hedges' *g*) was 0.119, which is to be interpreted as a treatment group having, on average, 0.119 higher scores in standard deviation units as compared to the scores of the control cohort and is to be interpreted a small effect size.

This model also provided information about the associated changes in student mean achievement score from one testing occasion to another, or a learning rate expressed as the γ_{100} coefficient (in Table 1). The learning rate for a minority, 1st grade female student who had never had a [program] teacher was 68.15 units. No student-level variables, with the exception of *current grade* (see the γ_{110} coefficient) had a statistically significant effect on the learning rate over this relatively short assessment time. Overall, students in lower grades experienced 19.26 units faster learning than students in higher grades over testing occasions (see the γ_{140} coefficient), when controlling for the effects of *gender*, *minority status* and *intervention*. As the reliability of the estimate of the mean learning rate was low (see below), these results should be interpreted cautiously.

STAR Mathematics

The model predicted mean achievement of 493.66 expressed as the γ_{000} coefficient (see Table 2) for the STAR Mathematics model is to be interpreted as a Winter 2015 scores for a minority, female, 3rd grade student who had never had a [program] teacher. Three of the four student-level variables had a statistically significant effect on the mean measure. The effect of *gender* on a student mean achievement status was not statistically significant (see the γ_{020} coefficient). However, students' scale scores increased by 85.96 units with an increase in *current grade* (i.e., moving from grade three to grade four) when controlling for the effects of *gender*, *minority status* and *intervention* (see the γ_{010} coefficient). A statistically significant effect for *minority status* on mean achievement was observed, with non-minority students scoring, on average, an additional 21.11 units in comparison to minority students (see the γ_{030} coefficient). This effect, again, should be interpreted cautiously in the absence of student's socio-economic status information.

Most importantly, the intervention variable has a statistically significant impact on students' mean achievement on the STAR Mathematics assessment (see the γ_{040} coefficient). An average increase of 21.75 units was associated with adding a [program] teacher to a student's academic history, controlling for the effects of the *current grade*, *gender* and *minority status* variables. This effect size (Hedges' *g*) was calculated as 0.179.

Analogously, with respect to the assessment of a student's learning rate, the average slope coefficient for a minority, 3rd grade female student who had never had a [program] teacher was 47.52 units (see the γ_{100} coefficient in Table 2). No student-level variables, with the exception of *current grade,* had a statistically significant effect on the learning rate over this relatively short assessment time. On average, students in higher grades increase their scores at 6.80 units slower than students in lower grades, when controlling for the effects of *gender*, *minority status* and *intervention* (see the γ_{110} coefficient).

Table 2

Summary of Three-Level Exploratory Model for STAR Mathematics Achievement

Fixed Effect	B	*SE* B	*t*-ratio	df	*p*
Model for average status, π_0					
Model for mean-status of 3rd grade minority female with no intervention β_{00}					
Average mean status, γ_{000}	493.66	5.26	93.79	40	<.001
Model for current grade, β_{01}					
Current grade, γ_{010}	85.96	2.01	42.87	40	<.001
Model for gender, β_{02}					
Gender, γ_{020}	2.86	2.24	1.28	5537	.190
Model for minority status, β_{03}					
Minority status, γ_{030}	21.11	3.07	6.88	5537	<.001
Model for cumulative intervention, β_{04}					
Cumulative intervention, γ_{040}	21.75	3.48	6.25	40	<.001
Model for learning rates, π_1					
Model for learning rates of 3rd grade minority female with no intervention, β_{10}					
Average learning rate, γ_{100}	47.52	1.78	26.75	40	<.001
Model for current grade, β_{11}					
Current grade, γ_{110}	-6.80	1.58	-4.32	40	<.001
Model for gender, , β_{12}					
Gender, γ_{120}	2.31	0.90	2.58	5537	.031
Model for minority status, β_{13}					

Minority status, γ_{130}	0.99	1.46	0.68	5537	.499
Model for cumulative intervention, β_{14}					
Cumulative intervention, γ_{140}	0.35	1.90	-0.19	40	.853

STAR Reading

The predicted mean achievement of 301.46 (the γ_{000} coefficient) represented a Winter 2015 score for a minority, female student between grades two and three who had never had a [program] teacher, as seen in Table 3 which summarizes the regression coefficients for the STAR Mathematics mean achievement model. The examination of the student-level variables included in the model demonstrated statistically significant effects for all of the second-level variables. Students' scale scores increased by 80.37 with an increase in *current grade* (i.e., moving from grade three to grade four) when controlling for the effects of *gender*, *minority status* and *intervention* (see the γ_{010} coefficient).

Table 3

Summary of Three-Level Exploratory Model for STAR Reading Achievement

Fixed Effect	B	*SE* B	*t*-ratio	df	*p*
Model for average status, π_0					
Model for mean-status of minority female who did not have intervention teacher between grades 2 and 3, β_{00}					
Average mean status, γ_{000}	301.46	8.93	33.74	40.00	<0.001
Model for current grade, β_{01}					
Current grade, γ_{010}	80.37	3.29	24.46	40	<0.001
Model for gender, β_{02}					
Gender, γ_{020}	-14.26	3.53	-4.03	4952	<0.001
Model for minority status, β_{03}					
Minority status, γ_{030}	42.42	3.92	10.83	4952	<0.001
Model for cumulative intervention, β_{04}					
Cumulative intervention, γ_{040}	47.85	4.86	9.85	40	<0.001
Model for learning rates, π_1					
Model for learning rates of minority female who did not have intervention teacher between grades 2 and 3, β_{00}					
Average learning rate, γ_{100}	53.06	2.23	23.83	40	<0.001
Model for current grade, β_{11}					
Current grade, γ_{110}	-3.71	1.52	-2.44	40	.019
Model for gender, β_{12}					
Gender, γ_{120}	1.82	1.69	1.08	4952	.282

Model for minority status, β_{13}					
Minority status, γ_{130}	4.88	1.81	2.69	4952	.007
Model for cumulative intervention, β_{14}					
Cumulative intervention, γ_{140}	0.99	2.68	0.37	40	.714

A statistically significant effect for *gender* (see the γ_{020} coefficient) on mean achievement was observed with female students gaining an additional 14.26 units in comparison to male students. A statistically significant effect for *minority status* on mean achievement was present, with non-minority students scoring, on average, an additional 42.42 units in comparison to minority students (see the γ_{030} coefficient). Again, this effect should be interpreted cautiously in the absence of student's socio-economic status information.

Most importantly, the intervention variable has a statistically significant impact on students' mean achievement on the STAR Reading assessment (see the γ_{040} coefficient). An average increase of 47.85 units was calculated as a function of adding a [program] teacher to a student's academic history, controlling for the effects of the *current grade*, *gender* and *minority status* variables. This effect size (Hedges' *g*) was calculated as 0.289, a level considered substantively important by the What Works Clearinghouse (US Department of Education, 2013).

As with the STAR Early Literacy and Mathematics models, this model also provided information about the increase in score from one testing occasion to another, (learning rate). The learning rate for a minority female student who had never had a [program] teacher, see the γ_{100} coefficient in Table 3, was 53.06 units. Most student-level variables had small, statistically significant effects on the learning rate. The effect of *current grade* (see the γ_{110} coefficient) on the learning rate was statistically significant, with students in higher grades learning at 3.71 units slower than students in lower grades. The growth differential for *minority status* (see the γ_{130} coefficient) was also statistically significantly different, with non-minority students making 4.88 unit gains more than non-minority students from one testing occasion to another. Also, the effect of *gender* (see the γ_{120} coefficient), controlling for the effects of *current grade*, *minority status* and *intervention*, was statistically significant, with males outgrowing females by an average of 1.82 units between assessment times.

Exhibit A: *Participating school district demographics*

Characteristic	Monroe ISD Head Start %	Monroe ISD-ECSE %	Monroe Public %	TPS %
Ethnicity				
White	60	89	81	38
African American	8	7	13	41
Hispanic	13	1	6	11
Asian/ Pacific Islander	1	0	0	1
American Indian	1	0	0	0
Multiracial, Non-Hispanic	17	3	0	9
Poverty rate	89	11	45	64
Achievement test data (below grade level)				
Reading	12	78	9	69
Mathematics	24	79	32	65
ELL	3	2	2	11
Disability	17	100	23	4
Homeless/foster care	3	2	1	NA