A Multi-level Investigation of Teacher Instructional Practices and the Use of the Responsive Classroom Curriculum

Final Report to The Wing Institute, August 12, 2010

Benjamin Solomon, M.Ed., Suzanne Klein, M.Ed., Amanda Marcotte, Ph.D., and John Hintze, Ph.D.

A Multi-level Investigation of Teacher Instructional Practices and the Use of the Responsive Classroom Curriculum

Overview of the Current Study

This executive summary serves as the final report for the study, "A Multi-level Investigation of Teacher Instructional Practices and The Use of Responsive Classroom", funded by the Wing Institute of Oakland, CA. We utilized longitudinal modeling across multiple school districts in New England to test whether the professional development sequence for teachers, Responsive Classroom (RC), as used in a typical setting, is effective in increasing student achievement over the course of one academic year. Results indicated that RC is an effective program based on teacher perceptions of student math, reading, and motivation growth. However a central tenant of the program, "the first six weeks", was shown not to be an effective practice in maximizing instructional time.

The Need for Tier I Prevention in the Classroom

Considerable attention has been paid recently to the utility of behavioral prevention in elementary schools. Whether it be internalizing problems such as depression or anxiety, or external problems, such as oppositional behavior and delinquency, prevalence studies have converged to show that the American educational system needs cost-effective solutions that prevent problems early in a student's school trajectory. For example, Gottfredson, Czeh, Cantor, Crosse, and Westat (2000) surveyed teachers from 6,451 schools nationally. They reported that 21% of teachers overheard threatening remarks between students in the 1997 – 1998 school year. In middle schools, 16.2% of students surveyed reported that they damaged school property during the same year. Fourty-one percent of students reported physically attacking or verbally threatening another student at least once during the year. At the elementary level, 34.2% of schools reported at least one case in the past year when physical aggression was severe enough to warrant calling law enforcement.

In comparison, Nansel et al. (2001) used the National Institute of Child Health and Human Development database to analyze bullying behaviors. Twenty-six percent of students reported bullying “once or twice” during the year. Ten and four tenths percent of students reported bullying peers weekly. Finally, a seminal report by Satcher in 2001 stated that 20.9% of school-age children had some type of clinically significant mental disorder. Additionally, rates of aggravated assault and robbery involving youths escalated nearly 70.0% from 1980 to 1999, despite other forms of criminal activity decreasing during the same period.

Rotter (1982) suggested that one primary causal mechanism for childhood pathology is the reactive adaptation of children to highly disordered environments, including peer behavior. As children actively construct their perception of behavioral and contextual interactions, vulnerabilities may arise when the environment may have reinforced negative or delinquent behavior. Such a theory suggests that early control of the environment, across all school settings, is critical to healthy development. This was complemented by Bronfenbrenner and Morris’ (2006) Ecological Systems Theory, which stated that human development is moderated by overlapping systems of different influence and size. Stability and positive interactions across individual systems, and combinations of systems, accumulate over time and will define future environment-person interactions. The classroom ecology is a complex, highly influential moderator on a student's behavioral development.

A hostile or unpredictable school environment would not contain itself to behavioral difficulties. Students who engaged in delinquent behavior would likely interfere with their own academic growth and the academic development of other students. For example, Mcintosh, Chard, Boland, and Horner (2006) found that behavioral and academic problems tended to covary together for individual students. Caprara et al. (2000) found similar results in an Italian sample by using social skills to predict academic outcomes for students six years out. Social skills, measured in kindergarten, were a better predictor of academic performance at sixth grade then academic performance measured in kindergarten. These studies are just a small fraction of the body of literature converging to show that behavioral difficulties and academic problems have a reciprocal relationship. This suggests that effective behavioral prevention, which includes strong classroom management, may serve a dual function of optimizing student behavioral development and allowing effective instruction to occur in the classroom. Diperna, Volpe, and Elliott (2001; 2005) called behavioral skills that serve as a prerequisite for achievement "academic enablers"; skills that can be taught and are necessary for academic growth, but are independent of academic instruction.

Fueled by this empirical support, many competing programs have emerged that claim to promote healthy development of social skills for students, effective classroom management, and positive school climate. Examples of such programs are as simple and low cost as the

“Good Behavior Game” (Barrish, Saunders, & Wold, 1969) or as comprehensive and technical as the Olweus Bullying Prevention Program (Olweus, 1997). These two model programs have a substantial body of evidence behind them demonstrating their effectiveness across a wide variety of contexts, however many other programs lack such robust evidence.

The emergence of school-based behaviorally orientated prevention programs have caused a reaction from the research and practicing community to establish quality indicators of programming. How do we know if a prevention program works? For what populations? Are outcome studies trustworthy? How much evidence is required before a program should be disseminated? Given the shrinking budgets of many public institutions, these are not trivial questions. Several organizations have offered comprehensive rubrics to evaluate the quality of prevention programming. The Wing Institute (2010), in their Road Map for Evidence-Based Practice, stated that different standards exist for programs to be deemed to have efficacy, be deemed effective, or ready for implementation.

Efficacy studies were defined as whether a program, under highly controlled conditions, demonstrate significant change. When a program is demonstrated to maintain such change in the natural environment of the target population, the program gains effectiveness. In both cases, constant replication is necessary to inform applied use. When practitioners unaffiliated with the developers of the program have shown to be able to implement the program with fidelity, it is ready for implementation (The Wing Institute, 2010). The process can be summed as a linear trajectory towards generalizability. Since one can never account for all contextual variables that may moderate the effectiveness of an intervention, The Wing Institute also stressed continued monitoring of an implemented intervention (The Wing Institute, 2001). This data can then be used to inform future replication studies.

It is in this framework of generalizability that the current study was proposed. RC is a comprehensive program for teachers whose developers, The Northeast Foundation for Children (NEFC), suggest is effective by borrowing best practice from a wide variety of teaching methodology (Northeast Foundation for Children, 2009). Specifically, RC training facilitates teacher development of social skills, methods for correcting undesirable behavior, introduction of academic materials and lessons, and a focus on student motivation through increased choice in the classroom, presentation of work, and goal setting (Northeast Foundation for Children, 2003). While the details of RC are beyond the scope of this report, in brief summary, RC is typically taught through traditional professional development sessions outside of school, or through a hired RC consultant who trains and provides feedback to teachers within a school. Examples of RC strategies include the use of "logical consequences" to correct student behavior, use of a structured morning meeting to enhance the classroom community, and an emphasis on personal student goals to increase motivation in the classroom. One unique tenant of RC is that teachers should spend the first six weeks of school primarily devoted to behavioral correction and teaching classroom routines, as opposed to engaging primarily in academic instruction. The theory behind this is that the social curriculum is important and should be established first (Northeast Foundation for Children, 2003).

Due to its focus on teacher training, the program is considered to work at the universal level of prevention, intended for all students regardless of risk status (Nelson et al., 2009).

Training costs vary widely, from $24 for a basic textbook on the morning meeting, $700 for a week-long training for a single teacher or a professional development kit, to $15,000 to host an on-site week long session (Northeast Foundation for Children, 2009).

Extant research on RC includes several program evaluations and published articles. This growing body of evidence on RC has been highly variable. For example, a program evaluation by Elliott in 1999 measured effect sizes for social skill growth in RC and non-RC locations. Effect sizes ranged from .07 to .41 based on who was rating student social skills, the highest ratings coming from teachers. Overall, the experimental RC schools showed significantly more growth in social skills then the non-RC schools. While it also appeared achievement scores were higher in the RC classrooms, this analysis was confounded by a low *n* and significant non-random attrition over time.

Rimm-Kauffman and Chiu (2007), using hierarchical regression, modeled the effectiveness of RC with student report cards, the Student-Teacher Relationship Scale, and the Social Skills Rating System as the primary DV's. The sample included 30 RC teachers and 27 non-RC teachers. Effect sizes of significant results were small, the highest $r^2 = .06$. The far majority of results were non-significant. Furthermore, while acknowledging the hazards of nesting effects, the authors did not control for them. A second study by Rimm-Kauffman, Chiu, and You (2007) tested for fidelity to RC practice after in-school consultation occurred. It was found that there were significant differences in teaching practice after the professional development when both teacher reported fidelity and observed implementation were analyzed. Furthermore, significant results were found for standardized math scores, which is discrepant with the previously mentioned research.

Replication of these past results in the academic and behavioral domains would greatly add to the generalizability of RC. This study also attempted to extend research on RC by specifically testing to see whether the "first six weeks" hypothesis could be both observed and then correlated to student outcomes. By controlling for nesting effects, more valid results can be calculated that account for the confound of shared teacher variance amongst student outcomes. Specifically, this study posed the following research questions:

- Will results of an effectiveness study converge with previous findings on RC?
- Will the effects of *RC* generalize across behavioral constructs not previously measured, including direct observation of teacher instructional practices?
- How does observed initial investment in classroom organization play out as a cost-benefit analysis across the school year?
- Can instructional behavior of the teacher be observed and quantified in a reliable and valid manner and used to answer the above questions?

In regards to these questions, we hypothesize that results of this study will converge with past findings to show that students of RC teachers show more growth in social skills and math then students of non-RC teachers. Furthermore, a large initial investment in classroom organization during the first six weeks of school will have a significant "payout" on overall student growth in measured student outcomes as opposed to teachers that choose to keep teaching practices stable over time. Using a generalizable form of direct observation of teaching practices had rarely been done to evaluate program effectiveness. We hypothesized that using structured, direct observation would unveil the ways in which RC teachers differ from non-RC teachers and the effect of these differences on student outcomes.

METHODS

Participants

Twenty-four teachers and 178 students participated in this study. Teachers were recruited from nine different elementary schools. Teachers ranged from having no exposure to RC to having completed a week long workshop. Ten teachers completed a day-long workshop (which includes reading materials) as their highest level of RC training. Twelve teachers had attended a week-long workshop (also including reading materials). Two teachers had either no RC exposure or had familiarized themselves with an RC textbook only. The majority of teachers had a Masters in Education [n = 16], with the remaining having Bachelors level training [n = 5] or a terminal degree beyond a Masters [n = 1]. Teachers had an average of 11.21 years of teaching experience [SD = .701]. Distribution of grade levels included are presented in Table 1. There were 97 female students and 81 male students.

Table 1
Cumulative Frequency of Grade Levels

Grade	Frequency	Percent	Cumulative Percent
Kindergarten	4	16.7	16.7
1st grade	6	25.0	41.7
2nd grade	6	25.0	66.7
3rd grade	2	8.3	75.0
4th grade	3	12.5	87.5
5th grade	3	12.5	100.0

Six of the schools were located in a district located in Western Massachusetts. The district served a total of 6,072 students taught by 508 teachers. Fourty-nine and seven tenths percent of the student body was considered low-income. The district was majority Caucasian [76.3%]. Standardized test scores for the district are descriptively identified as "high" for language arts and "moderate" for mathematics. In comparison to state scores, the district ranks slightly below average, although it met AYP in the last academic year (2008 – 2009). Seventy-one students and 10 teachers came from this site.

A second location was a single public elementary school located in Western Massachusetts, in a separate district. The school served a total of 410 students, 17.1% of which are considered low income. The district is primarily Caucasian [84.1%]. The school is considered to have "high" performance in both language arts and mathematics on statewide assessment. AYP was met for language arts in the previous academic year, but not for mathematics. Forty-one students and six teachers came from this school.

A third location was a private school located in Eastern Massachusetts. The school served a total of 270 students from Kindergarten through 8th grade. Twenty-four percent of students receive financial aid. The school reported that 85% of students are Caucasian. Fifty-six students and seven teachers were recruited from this school. Standardized test scores were unavailable at the time this report was submitted.

The final site for this study was an urban charter school located in Providence, RI. The school has 246 students. The majority of students were Hispanic [43%], with a sizable minority African-American [31%], and then Caucasian [17%]. The school was below state averages for reading and writing; 18 percentile points and 22 percentile points respectively. The school did make AYP in 2008. Sixty percent of the student body is considered low-income when free and reduced lunch status is used as a proxy. One teacher and eight students came from this site (a second teacher from this site dropped out of the study mid-year for a maternity leave).

Measures

Teaching Observation Tool (TOT)–The TOT is a 30 minute, momentary time sampling observation tool that was hypothesized to be sensitive to the instructional practices of the teacher (see Appendix A). The TOT is based on the work of Gibson and Hasbrouck (2007), who developed a brief observation using a frequency count that categorized teacher behavior as either managing the classroom environment, delivering instruction, or correcting behavior. The observation was intended to be used for graphic performance feedback in consultation, with the ultimate goal of shifting teaching time to small group instruction.

The TOT used in this study has a 15 second interval with a 3 second observation time. During that time, the observer coded teacher behavior as either “teaching” in small or whole group, “feedback”, “environment” or “behavior”. “Teaching” was defined as teacher-led dialogue with the intention of imparting knowledge. This did not include practicing previously learned skills. “Feedback” was defined as checking and responding to work being done individually or collaboratively by students, separate from a class lesson. “Environment” was defined as language with the intended purpose of managing the classroom, such as directing students to gather supplies or line up at the door, or teaching students routines directly related to classroom operation and unrelated to academic material. “Behavior” was defined as verbal behavior directed at altering a student’s behavior, either preemptively or reactively. “Behavior” could be either positive or negative; teacher actions intended to both increase or decrease the frequency of behavior was included. The observation included a global Likert-style rating of teacher quality for the observer to complete for each observation. It was hypothesized that effective teachers utilize teaching and feedback primarily, and use environmental management minimally and when done so, very effectively. Ineffective teachers would spend more of their time managing behavior and the environment and spend a minority amount of time teaching.

In this study, the TOT was used to test the hypothesis that investment in classroom organization early in the year leads to maximized instructional time later in the year, such that environment and behavior → instruction and feedback. Observers were in the back of the classroom during observations. Observations occurred three times per classroom per data gathering period, for a total of one and one half hours of observation across 360 intervals. In 27.59% of the teacher sample, only two observations could be completed due to limits on available time. A total of 162 observations were completed. Point-by-point interobserver agreement was 86.87% across 19.25% of observations.

Academic Competence Scale (ACES),teacher version – The ACES is an 81 item questionnaire covering seven domains of student functioning. These domains load onto two factors, academic enablers and academic skills (Diperna & Elliott, 2000). It was completed by the general education teacher of the student being evaluated. Academic skills contained the subscales reading, math, and critical thinking. These subscales reflected teacher perception of a student's grade level proficiency on critical academic domains. Questions were based on individual skills such as spelling and vocabulary. Academic enablers included interpersonal skills, engagement, motivation and study skills. Academic enablers were skills that were required antecedents of academic skills. Sample questions from these subscales include "participates in class discussion" for engagement or "works effectively in small group activities" for interpersonal skills. The math, reading, motivation to learn, and interpersonal skills subscales were used in this study to form a comprehensive picture of student achievement while reducing survey length for teachers.

The ACES used a five point Likert-style response option. A one would indicate a skill is "far below" age-based norms, while a five would indicate a skill is "far above" age-based norms (DiPerna & Elliott, 2000). The ACES was nationally normed on a geographically and economically diverse sample of 1000 k-12 age students. The reading subscale has a reliability of $\alpha = .88$, math, $\alpha = .98$, interpersonal skills, $\alpha = .97$ and motivation, $\alpha = .97$. Test-retest reliability was also robust: reading was equal to $r = .95$, math equal to $r = .93$, interpersonal skills equal to r = .81 and motivation, r = .84. The reading subscale strongly correlated to measured reading achievement on the Iowa Test of Basic Skills [r = .80], as did the math subscale [r = .86]. The interpersonal skills subscale had moderate convergent validity with the Social Skills Rating System [r = .50] (Diperna & Elliott, 2000).

The Classroom Practice Measure (CPM) - RC is comprised of a set of skills, each skill used to address different areas of student and classroom functioning. Past research found that teachers often selected certain practices within *RC* and chose not to employ others (Elliott, 1999). Hence, implementation of *RC* is not a dichotomous variable, but rather falls along a continuum based on the individual choices of the teacher. The CPM is a fidelity survey designed to measure these varied practices (Rimm-Kaufman & Chiu, 2007). It includes questions for teachers about their use of *RC* in the following domains: hand signals, classroom opening exercises, classroom rules and consequences, classroom organization, introduction to materials, student choice, student reflection, assessment and parent communication, time-out, and use of a problem solving meeting (2007). The CPM contained no language indicating it is an assessment of *RC* fidelity.

Due to scaling concerns, the CPM’s Likert-style scale was converted to a 3-point scale with possible answers being, “No, this is not my present in my class”, “Yes, this is present in my class, but not in the way described by Responsive Classroom” and “Yes, this is present in class as defined by the developers of Responsive Classroom.” To increase teacher participation, open-ended questions were removed. Included with the CPM was a brief list of demographic questions.

The CPM has excellent reliability and acceptable validity. Reliability was α = .94 (Rimm-Kaufman & Chiu, 2007). In the same study, a sample of 68 *RC* teachers who filled out the CPM showed moderate correlation with two trained observers who went into their classroom [r = .70], demonstrating adequate construct validity. Discriminate validity was established by comparing CPM scores between teachers who had *RC* training and those that did not. Summed scores were significantly different between the two groups [t = 4.86, p ≤ .000] (2007).

Procedure

Permission to observe and survey teachers was secured from school principals in the spring of 2009. In the fall, teachers were recruited through e-mail request and shortly thereafter by mailings. During this time, teachers provided scheduling information of when their reading and math blocks were. Teachers were given the time and told via e-mail at least 24 hours in advance of an observation occurring. Informed consent was secured from all teachers who participated and the school principal.

Teachers were observed for three 30 minute blocks throughout the first six weeks of each sites academic year. Each set of observations were divided amongst at least two days. Shortly thereafter, the CPM and ACES were mailed to teachers. Observations and surveying were then repeated in the final six weeks of the school year in the same fashion. Over the course of the year, two teachers dropped out of the study, one due to a maternity leave and the other switched schools within the district. Observations were completed primarily by the first two authors, however a minority of observations were completed by graduate students from the graduate program of the authors. Participating graduate students completed a three hour training on use of the TOT and were compensated for travel. The training was not completed until 90% agreement amongst trainees was reached when observing a sample video.

A random number generator was used to select target students for the ACES. Teachers then selected the students when they were organized alphabetically by last name. Teachers proceeded on their own time to secure informed consent from parents of these students. If teachers could not get informed consent for a particular student, teachers were instructed to select a replacement student of similar behavior and academic achievement. Three teachers secured informed consent for only four students. Between the fall and the spring, four students dropped out of the study. The primary reason for this was families moving out of district.

RESULTS

Descriptive data for the TOT are presented in Table 2 and descriptive data for the ACES is presented in Table 3. CPM scores had a mean of 53.43 and standard deviation of 13.03. CPM scores showed evidence of moderate negative skew (more teachers reported implementing RC with high fidelity then with moderate or low fidelity).

Table 2
Average Scores of TOT Dimensions Across Time

	Pre-Test M (SD)	Post-Test M (SD)
Instruction	76.92 (20.23)	82.67 (18.96)
Feedback	14.79 (19.33)	21.87 (22.56)
Environment	25.80 (16.46)	22.24 (11.85)
Behavior	8.12 (7.02)	5.76 (7.38)

Table 3
Comparison of ACES scores across sample locations

	Pre-Test M (SD)				Post-Test M (SD)			
	Site I	Site II	Site III	Site IV	Site I	Site II	Site III	Site IV
Reading	31.25 (14.50)	32.77 (8.64)	34.09 (12.00)	35.97 (11.39)	32.50 (12.92)	35.00 (9.76)	40.27 (8.32)	36.23 (9.95)
Math	24.88 (7.61)	20.82 (4.04)	23.09 (3.65)	22.91 (5.52)	27.63 (9.62)	25.27 (6.39)	29.68 (6.06)	27.11 (7.42)
Interper-sonal Skills	41.25 (6.69)	41.44 (8.24)	43.27 (5.78)	41.43 (7.09)	40.75 (7.23)	42.47 (7.49)	42.72 (6.01)	41.86 (7.26)
Motivation	38.00 (13.16)	37.28 (9.69)	42.13 (6.77)	39.95 (9.81)	42.13 (12.11)	38.97 (10.49)	43.79 (7.65)	42.47 (9.56)

Teacher level results

To answer the question of how teachers who endorsed the use of RC in their classroom change their behavior in the beginning and end of the year, a correlational analysis was done across dimensions of the TOT in the fall and spring and the CPM. Results of this analysis are presented in Table 4. These are partial correlations; summed average quality rating of the TOT from the fall and spring served as a control.

Table 4
Partial Correlation Matrix of Teacher Variables

Controlling for total quality rating across observations									
	1	2	3	4	5	6	7	8	9
1. RC	-	.021	.336	.034	.102	.205	.141	-.112	-.076
2. PreIns	.021	-	.031	-.443*	-.499**	-.006	.495**	-.096	-.352*
3. PreFeed	.336	.031	-	.319	-.052	.225	.135	-.011	-.204
4. PreEnv	.034	-.443*	.319	-	-.368*	.143	-.253	.206	-.264
5. PreBeh	.102	-.499**	-.052	-.363*	-	-.051	-.390*	-.058	.479*
6. PostIns	.205	-.006	.225	.143	-.051	-	.249	-.754**	-.403*
7. PostFeed	.141	.495**	.135	-.253	-.390*	.249	-	-.106	-.403*
8. PostEnv	-.112	-.096	-.011	.206	-.058	-.754**	-.106	-	.049
9. PostBeh	-.076	-.352*	-.204	-.264	.479*	-.403*	-.403*	.049	-

Note. n = 25
Note. RC = CPM score; PreIns = Fall TOT instruction time; PreFeed = Fall TOT feedback time; PreEnv = fall TOT environment time; PreBeh = fall TOT behavior time; PostIns = spring TOT instruction; PostFeed = spring TOT feedback time; PostEnv = spring TOT environment time; PostBeh = spring TOT behavior time
*significant at the .05 level
** significant at the .01 level

Correlational analysis failed to demonstrate that fidelity to RC practice could be definitively linked to certain patterns of teaching behavior. There were significant relationships across time and dimensions of the TOT. There was a consistent negative correlation between the amount of instructional time and time spent managing the environment and correcting behavior. Unexpectedly, there was also a negative relationship between time spent managing the environment and behavioral correction. Looking at temporal variability, there was a negative relationship between time spent correcting behavior in the fall and spring levels of giving feedback to students [$r = -.395$, $p = .028$]. Fall behavioral correction also predicted spring behavioral correction [$r = .463$, $p = .011$]. Time devoted to instruction in the fall predicted spring levels of behavioral correction [$r = -.360$, $p = .042$].

Student level results

To address the question of the effect of RC and teaching behavior on student outcomes, a multilevel regression was done using difference scores (spring – fall) of the TOT dimensions as the predictors of ACES difference scores. Multilevel modeling was appropriate over linear regression as the ICC was large; ranging from .33 for the motivation subscale to .51 for reading. HLM 6.08 (Scientific Software International, 2009) was used to calculate results for all models. It was hypothesized that teachers with the largest difference scores of Environment (decreases over time) and Instruction (increases over time) conform to the "first six weeks" hypothesis and will have the largest degree of positive student change. Major assumptions of regression were fulfilled. However, despite a low colinearity statistic, ranging from VIF = 1.04 to VIF = 3.02, multicolinearity might have been a problem since the IV's, all coming from the same source, tended to vary in proportion to one another. To reduce multicolinearity, the Behavior variable of the TOT was removed. Behavior was noted to be problematic due to its rapid, short frequency, that may not have been appropriate for momentary time sampling. Additionally, it was relatively infrequent in comparison to other variables. All IV's were grand-mean centered to further reduce multicolinearity and aid in interpretation.

In the following model, variables were defined as:

Level I (student)

$ACES = \beta_0 + \beta_1 PREACES_1 + e$

In this equation, Y is the difference across time of a given ACES domain, β_0 is the intercept and β_1 equals the matching prescore of Y .

Level II (teacher)

$\beta_0 = \gamma_{00} + \gamma_{01} QUALITY_{01} + \gamma_{02} PREINSTRUCTION_{02} + \gamma_{03} PREFEEDBACK_{03} + \gamma_{04} PREENVIRONMENT_{04} + \gamma_{05} INSTRUCTION_{05} + \gamma_{06} FEEDBACK_{06} + \gamma_{07} ENVIRONMENT_{07} + \gamma_{08} CPMSCORE_{08} + \mu_0$

$\beta_1 = \gamma_{10} + \gamma_{11} QUALITY_{11} + \gamma_{12} PREINSTRUCTION_{12} + \gamma_{13} PREFEEDBACK_{13} + \gamma_{14} PREENVIRONMENT_{14} + \gamma_{15} INSTRUCTION_{15} + \gamma_{16} FEEDBACK_{16} + \gamma_{17} ENVIRONMENT_{17} + \gamma_{18} CPMSCORE_{18} + \mu_1$

In this equation, γ_1 was the average quality score assigned to observations, which allowed us to control for quality of instruction; γ_2 through γ_4 are the various prescores of the TOT domains, thus creating a pretest covariate; γ_5 through γ_7 are difference scores from the domains of the TOT; finally, γ_8 was score on the CPM. Results of β_0, which represented the teacher level variables that explain student level ACES scores, are presented in Table 4.

Three different models for this analysis were created to model the progressive proportion of variance explained in the level II intercept, defined as τ. Model A was a null model that only included pretest TOT scores as predictors. Model B included CPM scores on Level II in addition to model A. The full model is described above. For reading, model B explained 11.75% more variance then model A; the full model explained 6.45% more variance then model B. For math, model B explained 3.52% more variance then model A; the full model explained 18.52% more variance then model B. For interpersonal skills, model B explained 1.56% more variance then model A; the full model explained 6.13% more variance then model B. Finally, for motivation, model B explained 3.6% more variance then model A and the full model explained 8.27% more variance then model B.

Teachers who devoted more time to feedback in the spring then in the fall had students with higher growth in teacher-rated math [γ_{06} = .58, p = .001]. For each one unit increase in fall to spring feedback scores, one can expect overall math scores on the ACES to increase .58 units. For the DV's of reading, math, and motivation, scores on the CPM were significant, positive predictors. Specifically, a one unit increase on the CPM equates to a .86 unit increase in reading growth over the course of one year [p = .014], a .56 unit increase in math [p = .003], and a .83 unit increase in motivation [$p \leq .000$].

To test the sensitivity of the TOT and investigate teaching practices more broadly, a similar multilevel model was constructed using either the pretest TOT data or the posttest TOT data as predictors. This was done to test whether varying levels of teaching time in the fall and spring predicted student outcomes, ignoring direct shifts in teaching time that were considered in the previous analysis and CPM scores. Level I remained the same as in the previous analysis. Level II was constructed as such:

Level II

$\beta_0 = \gamma_{00} + \gamma_{01}QUALITY_{01} + \gamma_{02}PREINSTRUCTION_{02} + \gamma_{03}PREFEEDBACK_{03} +$

$\gamma_{04}PREENVIRONMENT_{04} + \mu_0$

$\beta_1 = \gamma_{10} + \gamma_{11}QUALITY_{11} + \gamma_{12}PREINSTRUCTION_{12} + \gamma_{13}PREFEEDBACK_{13} +$

$\gamma_{14}PREENVIRONMENT_{14\ 7} + \mu_1$

For analysis of post-test, spring TOT scores were substituted for fall scores. Results are presented in Table 6. Amount of time delivering instruction in the fall and spring was associated with higher growth in teacher-rated reading . Specifically, for every 15 seconds of additional instructional time (one unit on the TOT) during a 30 minute time period in the fall, .91 units were gained on the ACES over the course of the year [$p = .002$]. Amount of time devoted to individualized feedback in the spring was also related to higher slopes of teacher-rated reading improvement [$\gamma_{03} = .630$, $p = .018$]. Higher levels of individualized feedback in the spring was also associated with a greater slope of improvement on teacher-rated math achievement, although it was the smallest effect ($\gamma_{03} = .42$, $p = .002$).

Table 5
Level II Results of Multilevel Model – TOT Difference Scores

	Reading			Math			Interpersonal Skills			Motivation		
R_1^2 of τ											.	
Level II	Coefficient	St. error	t-ratio	Coefficient	St. error	t-ratio	Coefficient	St. error	t-ratio	Coefficient	St. error	t-ratio
γ_0	14.83*	3.14	4.72	11.15*	1.98	5.64	17.00*	2.38	7.16	15.11*	1.79	8.43
γ_1	-1.20	1.23	-0.98	2.26*	0.88	2.58	3.97	1.78	2.23	2.83*	0.96	2.96
γ_2	-0.13	0.72	-0.18	-0.45	-0.26	-1.71	0.61	0.52	1.17	0.05	0.38	0.12
γ_3	0.03	0.43	0.07	0.38	0.24	1.58	0.69	0.30	2.28	0.26	0.26	1.02
γ_4	0.38	0.74	0.51	0.84	0.40	2.09	1.15	0.61	1.88	1.06*	0.42	2.53
γ_5	-0.75	0.59	-1.28	-0.50	0.26	-1.89	0.77	0.46	1.68	-0.04	0.32	-0.14
γ_6	0.36	0.29	1.26	0.58*	0.13	4.57	0.22	0.24	0.91	0.32	0.16	2.00
γ_7	-0.44	1.06	-0.41	0.37	0.48	0.78	1.52	0.73	2.07	0.96	0.54	1.78
γ_8	0.86*	0.30	2.87	0.56*	0.15	3.78	0.28	0.27	1.05	0.83*	0.15	5.53

Note. The Dunn-Bonferroni correction was applied to statistically significant results with academic enablers and academic skills each treated as a family.
Note. Results are weighted by the total number of individual observations done with each teacher.
*significant at the .05 level.

Table 6
Level II Results of Multilevel Model – TOT fall and spring scores

	Reading			Math			Interpersonal Skills			Motivation		
Level II	Coefficient	St. error	t-ratio	Coefficient	St. error	t-ratio	Coefficient	St. error	t-ratio	Coefficient	St. error	t-ratio
Fall TOT scores												
γ_0	18.13	3.84	4.73*	7.67	3.30	2.32	17.38	2.78	6.25	14.39	3.15	4.57*
γ_1	-1.28	1.87	-0.69	2.63	2.50	1.06	4.28	1.98	2.16	3.96	1.99	1.99
γ_2	0.91	0.25	3.68*	0.27	0.18	1.45	0.13	0.21	0.61	0.25	0.26	0.97
γ_3	0.31	0.42	0.74	0.24	0.28	0.85	0.69	0.30	2.31	0.29	0.29	0.99
γ_4	0.58	0.34	1.70	0.08	0.32	0.25	0.08	0.26	0.29	-0.06	0.35	-0.18
Spring TOT scores												
γ_0	19.43	4.03	4.82*	10.78	2.66	4.05*	16.94	2.88	5.88	14.78	2.44	6.06
γ_1	0.35	2.13	0.16	3.25	0.66	4.90*	2.72	2.16	1.26	4.48	1.11	4.02*
γ_2	-0.02	0.46	-0.05	-0.08	0.27	-0.29	0.69	0.47	1.47	-0.28	0.23	-0.12
γ_3	0.63	0.24	2.64*	0.42	0.11	3.71*	0.17	0.22	0.78	0.28	0.13	2.10
γ_4	1.09	0.87	1.26	0.96	0.58	1.64	1.21	0.64	1.89	0.66	0.40	1.64

Note. The Dunn-Bonferroni correction was applied to statistically significant results with all academic enablers and academic skills each treated as a family.
Note. Results are weighted by the total number of individual observations done with each teacher.
*significant at the .05 level.

DISCUSSION

This study tested whether an independent evaluation of RC would align with previous results. To add to the body of literature on RC, direct observation was used to test whether teachers who chose to sacrifice instructional time in the beginning of the year to implement and teach classroom routines and establish behavioral norms had a greater amount of instructional time in the spring, and a corresponding higher slope of student growth. As direct observation has rarely been used as a measured outcome in program evaluation, this study also investigated how varying teaching practices relate to student growth trends.

Will results of an effectiveness study converge with previous findings on RC?

Will the effects of RC generalize across behavioral constructs not previously measured, such as direct observation of teacher instructional practices?

Correlational analysis showed that there were no behavioral differences between teachers that reported using RC with high fidelity and those that did not. One would expect that use of RC practice would positively correlate with time devoted to environmental management and behavioral correction in the fall and instruction and feedback in the spring. This was not the case, with no discernable pattern to observed RC teaching practices. It was noted anecdotally in observations that teachers highly trained in RC did spend a large amount of time introducing classroom materials, otherwise known as “guided discovery” (Northeast Foundation for Children, 2003). For example, one teacher used yellow caution tape to carefully control students handling of classroom materials before they had been introduced in regards to their function and proper use. Another teacher used a 30 minute math block just to have students brainstorm ways to use certain tools in a math toolkit, then debriefed the class as to their function. Despite these observations, the TOT was not sensitive to the relative infrequency of these types of behaviors. While it is logical that teachers should be proactive in teaching students the proper use of materials, it appeared RC teachers in this sample tended to do this selectively. Certainly not to such a great magnitude that it significantly reduced instructional time. It is possible that teachers changed their normal course of instruction due to the presence of observers in the room.

While there was no observable differences in the way RC teachers managed their classrooms, there were compelling results at the student level. Students of teachers who endorsed the use of RC had a greater slope of progress in teacher ratings of reading, math, and motivation. Interestingly, this included all student DV's except interpersonal skills. In one sense, this confirmed past research, such as Rimm-Kauffman et al. (2007), who found that use of RC increased math achievement. It also added to the current body of literature by demonstrating a strong relationship between slopes of teacher-rated reading growth, student motivation, and the use of RC practices. Elliott (1999) also found significant results for reading achievement. However, this was in contrast to past research, such as Rimm-Kauffman & Chiu (2007), that found non-significant findings for the use of RC to improve reading scores.

This study's results also conflicted with past findings. Elliot (1999) found in a quasi-experimental between-groups analysis that RC did improve teacher-rated social skills of students. This was not the case for this study. It was difficult to isolate a cause of this discrepancy, particularly considering the largest difference was found through teacher observations in Elliott (1999). The current study also used teacher observations. The survey instruments used were different and only moderately correlated. Furthermore, Elliott's (1999) research was an efficacy study, as it included the district in which RC was developed. It is likely that fidelity to RC practice was higher when the developers of RC participated in the study.

The question remains as to why RC teachers rated their student achievement as developing faster over time then non-RC teachers. The authors chose not to look at subscales within the CPM; reliability and validity of individual subscales was not known. One theory is that the heavy emphasis on reminding students about rules and provided frequent in vivo' correction leads to more time on instruction (despite correlational results) which increases student academic achievement over time. This could be supplemented by the focus on an orderly and easily accessible classroom that minimizes the time needed to find materials and maximizes the teacher's supervision of the class. RC also places heavy emphasis on involving parents in the learning process. This collaboration may result in parents responding more quickly to problems in the classroom. RC's focus on freedom of choice in academic work, personalized goals, and use of positive teacher language may have explained at least some of the variance in teachers' perceptions of changes in student motivation.

How does initial investment in classroom organization play out as a cost-benefit analysis across the school year?

The data suggested that "the first six weeks" theory is not the optimal way to manage a classroom. Granted, there was a substantial amount of variance explained in student level data by the addition of the TOT difference scores; far more than just adding CPM scores as a predictor. However, relationships between time spent managing the environment in the fall and spring level instruction and feedback time were not significant. In fact, conflicting relationships emerged. For example, in the analysis of teaching behavior only, there was a significant negative correlation between instruction time in the fall and spring behavioral corrections. This suggested that strong teachers immediately put emphasis on instructional time. Having students engaged during instructional time reduced the potential for behavior problems over the course of the year. While the RC teacher might spend time carefully introducing materials, they also must move quickly into instruction to maintain an optimal classroom. This was further supported by the observation that teachers who spent a significant amount of time correcting student behavior in the fall continued to do so in spring.

This finding converged with results from student level data. When the difference scores from the TOT were used as predictors of student achievement, there were no significant results for student outcomes. This suggested that the teachers who increased instructional time over the course of the year and decreased environmental and behavioral time had no advantage over teachers who came in strong and consistent with instruction from the beginning of the school year.

Taken together, the findings suggested that training in RC does improve a teacher's ability to motivate their students and improves teacher perceptions of reading and math achievement. RC practices, such as maintenance of an orderly classroom, goal setting, positive, specific teacher language, and parent involvement may combine to result in a significant positive effect for students. However, this study also demonstrated that a heavy emphasis on tasks outside of curricular instruction, such as teaching classroom routines, in the beginning of the school year is not best practice. While modeling and reminding are critical features of superior classroom management, good teachers know how to do this without significantly taking away from instructional time. While morning meetings, heavily emphasized in RC, may be important, practices such as this must not significantly take away from the instructional time students need.

Can instructional behavior of the teacher be observed and quantified in a reliable and valid manner and used to answer the above questions?

This study introduced the TOT, a measure designed to observe teaching behavior in the classroom. Throughout this study, the TOT was a robust and reliable measure of teaching behavior. IOA was within acceptable levels. Furthermore, the TOT revealed an optimal pattern of teaching behavior over the course of the year that mirrors theory on direct instruction. For example, fall instructional time was a significant predictor of teacher-rated reading achievement. However in the spring, instructional time was no longer a significant predictor. Time spent giving students individualized feedback had taken its place as the key predictor of student reading scores. In this sense, the TOT captured the "I do, we do, you do" theory of direct instruction (Adams & Engelmann, 1996), where control of learning is gradually released to the student over the course of the year.

Fall feedback scores were not a significant predictor of growth in teacher-rated math achievement. However, an increase in feedback (spring feedback scores – fall feedback scores) was a significant predictor of math achievement. Taken together, this suggested that only when teachers first modeled the learning process by using controlled, explicit instruction and gradually shifted into independent work was the use of individualized feedback effective. If feedback occurred too soon, it lost its value as an instructional method. It should be noted that direct instruction and RC are not mutually exclusive. Teachers who endorse the classroom management aspects of RC could simultaneously use direct instruction.

Limitations

The ACES is not a direct measure of student ability in regards to academic achievement. It is a proxy – an opinion – completed by the teacher. Given the logistical constraints of this study, direct assessment was unduly prohibitive. The same concern arose for use of the CPM. While the CPM demonstrated validity in this study and previous research, it is a proxy of actual fidelity to RC.

As with any study based on correlational analysis, there is never certainty regarding causation. While the relationship between RC and student outcome variables were strong, it was possible that a lurking variable created the illusion of a direct relationship. Given that the sample was relatively diverse, we could not think of such a variable. However, that did not preclude one from existing.

While this study added to the generalizability and effectiveness of RC, there were methodological concerns that should be considered. The sample size of teachers was low. Results drawn from only teacher level data are potentially under-powered, increasing the chance of type II error. Previously mentioned, CPM scores were negatively skewed. As the primary measure of RC fidelity, this skewed distribution may have resulted in a loss of overall validity for the study. There were several potential causes for this skew. One, this study took place in an area close to the location of the NEFC. This influence may have resulted in a high concentration of teachers exposed to RC. Second, construct validity may have been threatened by social desirability bias. A third possible cause was that the CPM does not properly differentiate RC from other types of teaching behavior, an issue of discriminate validity, despite previous findings (e.g. Rimm-Kaufman et al., 2007).

Finally, due to logistical limitations, we could not observe teachers for more than a total of 1.5 hours each semester. This potentially increased the chance of spurious results since certain behaviors may have been infrequent and require a much longer observation period to reliably detect. While observations were split across days and occurred across all times of the school day, the brevity of observations remains a concern.

Appendix A

Teaching Observation Tool

Date of Observation: ________________________________ **Time:** ___________________________

Grade Level: _____________ **# of Adults:** _____________ **# of Students:** _____________

Content of the Lesson: Reading Writing Math Social Studies Science Other: _____________

	1	2	3	4	5	6	7	8	9	10
Teaching	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W
Feedback	F	F	F	F	F	F	F	F	F	F
Environment	E	E	E	E	E	E	E	E	E	E
Behavior	B	B	B	B	B	B	B	B	B	B

	11	12	13	14	15	16	17	18	19	20
Teaching	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W
Feedback	F	F	F	F	F	F	F	F	F	F
Environment	E	E	E	E	E	E	E	E	E	E
Behavior	B	B	B	B	B	B	B	B	B	B

	21	22	23	24	25	26	27	28	29	30
Teaching	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W
Feedback	F	F	F	F	F	F	F	F	F	F
Environment	E	E	E	E	E	E	E	E	E	E
Behavior	B	B	B	B	B	B	B	B	B	B

	31	32	33	34	35	36	37	38	39	40
Teaching	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W
Feedback	F	F	F	F	F	F	F	F	F	F
Environment	E	E	E	E	E	E	E	E	E	E
Behavior	B	B	B	B	B	B	B	B	B	B

	41	42	43	44	45	46	47	48	49	50
Teaching	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W	T S / W
Feedback	F	F	F	F	F	F	F	F	F	F
Environment	E	E	E	E	E	E	E	E	E	E
Behavior	B	B	B	B	B	B	B	B	B	B

	51	52	53	54	55	56	57	58	59	60
Teaching	T	T	T	T	T	T	T	T	T	T

	S / W	S / W	S / W	S / W	S / W	S / W	S / W	S / W	S / W	S / W
Feedback	F	F	F	F	F	F	F	F	F	F
Environment	E	E	E	E	E	E	E	E	E	E
Behavior	B	B	B	B	B	B	B	B	B	B

	61	**62**	**63**	**64**	**65**	**66**	**67**	**68**	**69**	**70**
Teaching	T	T	T	T	T	T	T	T	T	T
	S / W	S / W	S / W	S / W	S / W	S / W	S / W	S / W	S / W	S / W
Feedback	F	F	F	F	F	F	F	F	F	F
Environment	E	E	E	E	E	E	E	E	E	E
Behavior	B	B	B	B	B	B	B	B	B	B

	71	**72**	**73**	**74**	**75**	**76**	**77**	**78**	**79**	**80**
Teaching	T	T	T	T	T	T	T	T	T	T
	S / W	S / W	S / W	S / W	S / W	S / W	S / W	S / W	S / W	S / W
Feedback	F	F	F	F	F	F	F	F	F	F
Environment	E	E	E	E	E	E	E	E	E	E
Behavior	B	B	B	B	B	B	B	B	B	B

	81	**82**	**83**	**84**	**85**	**86**	**87**	**88**	**89**	**90**
Teaching	T	T	T	T	T	T	T	T	T	T
	S / W	S / W	S / W	S / W	S / W	S / W	S / W	S / W	S / W	S / W
Feedback	F	F	F	F	F	F	F	F	F	F
Environment	E	E	E	E	E	E	E	E	E	E
Behavior	B	B	B	B	B	B	B	B	B	B

	91	**92**	**93**	**94**	**95**	**96**	**97**	**98**	**99**	**100**
Teaching	T	T	T	T	T	T	T	T	T	T
	S / W	S / W	S / W	S / W	S / W	S / W	S / W	S / W	S / W	S / W
Feedback	F	F	F	F	F	F	F	F	F	F
Environment	E	E	E	E	E	E	E	E	E	E
Behavior	B	B	B	B	B	B	B	B	B	B

	101	**102**	**103**	**104**	**105**	**106**	**107**	**108**	**109**	**110**
Teaching	T	T	T	T	T	T	T	T	T	T
	S / W	S / W	S / W	S / W	S / W	S / W	S / W	S / W	S / W	S / W
Feedback	F	F	F	F	F	F	F	F	F	F
Environment	E	E	E	E	E	E	E	E	E	E
Behavior	B	B	B	B	B	B	B	B	B	B

	111	**112**	**113**	**114**	**115**	**116**	**117**	**118**	**119**	**120**
Teaching	T	T	T	T	T	T	T	T	T	T
	S / W	S / W	S / W	S / W	S / W	S / W	S / W	S / W	S / W	S / W
Feedback	F	F	F	F	F	F	F	F	F	F
Environment	E	E	E	E	E	E	E	E	E	E
Behavior	B	B	B	B	B	B	B	B	B	B

	Poor	Weak	Typical	Good	Excellent
Teaching Quality	1	2	3	4	5

Did students Transition? Yes No

If Yes, what did they transition to? ____

____

____

References

Adams, G.L. & Engelmann, Siegfried (1996). Research on Direct Instruction: 25 Years beyond DISTAR. Seattle, WA: Educational Achievement Systems.

Barrish, H.H., Saunders, M., & Wold, M.M. (1969). Good behavior game: Effects of individual contingencies for group consequences on disruptive behavior in a classroom. *Journal of Applied Behavior Analysis*, 2, 119-124.

Bronfenbrenner, U., & Morris, P.A. (2006). The Bioecological Model of Human Development. In W. Damon, & R. M. Lerner (Vol. ed.). *Handbook of Child Psychology: Vol. 1. Theoretical models of human development* (pp. 993-1028). New York: Wiley and Sons.

Caprara, G.V., Barbaranelli, C., Pastorelli, C., Bandura, A., & Zimbardo, P.G. (2000). Prosocial Foundations of Children's Academic Achievement. *Psychological Science, 11*, 302-306.

DiPerna, J.C., & Elliott, S.N. (2000). *Academic competence evaluation scales*. San Antonio: The Psychological Corporation.

DiPerna, J.C., Volpe, R.J., & Elliott, S.N. (2001). A Model of Academic Enablers and Elementary Reading/Language Arts Achievement. *School Psychology Review*, *31*, 298-312.

DiPerna, J.C., Volpe, R.J., & Elliott, S.N. (2005). A model of academic enablers and mathematics achievement in the elementary grades. *Journal of School Psychology*, *43*, 379-392.

Elliott, S.N. (1999). A Multi-Year Evaluation of the Responsive Classroom Approach: Its Effectiveness and Acceptability in Promoting Social and Academic Gain. Springfield, MA: Northeast Foundation for Children & Kensington Avenue Elementary School Staff.

Retrieved January 25th, 2008, from http://www.responsiveclassroom.org/about/research.Html.

Gibson, V., & Hasbrouck, J. (2007). *Differentiated Instruction.* New York: McGraw-Hill.

Gottfredson, G.D., Gottfredson, D.C., Czeh, E.R., Cantor, D., Crosse, S.B., Westat, I.H. (2000). Summary: National study of Delinquency Prevention in Schools (Available from Gottfredson Associates, Inc. 3239B Corporate Court, Ellicott City, Maryland 21042).

HLM for Windows (Version 6.08) [Computer software]. Lincolnwood, IL: Scientific Software International.

McIntosh, K., Chard, D.J., Boland, J.B., & Horner, R.H. (2006). Demonstration of Combined Efforts in School-Wide Academic and Behavioral Systems and Incidence of Reading and Behavior Challenges in Early Elementary Grades. *Journal of Positive Behavior Interventions*, *8*, 146-154.

Nansel, T.R., Overpeck, M., Pilla, R.S., Ruan, J., Simons-Morton, B., & Schedit, P. (2001). Bullying Behaviors Among US Youth: Prevalence and Association With Psychosocial Adjustment. *The Journal of the American Medical Association*, *25*, 2094-2100.

Nelson, J.R., Hurley, K.D., Synhorst, L., Epstein, M.H., Stage, S., & Buckley, J.N. (2009). The Child Outcomes of a Behavior Model. *Exceptional Children*, *76*, 7-30.

Northeast Foundation for Children, Inc. (2003). Responsive Classroom: Level I Workbook (Available from the Northeast Foundation for Children, Inc, 85 Avenue A, Suite 204, Turners Falls, MA 01376).

Northeast Foundation for Children, Inc. (2009). Responsive Classroom. Retrieved October 27th, 2009, from http://www.responsiveclassroom.org/about/aboutrc.html 997.

Olweus, D. (1997). Bully/victim problems among schoolchildren: Knowledge base and an effective intervention program. *Irish Journal of Psychology*, 18, 170-190.

Rimm-Kaufman, S.E., & Chiu, Y. (2007). Promoting Social and Academic Competence in the Classroom: An Intervention Study Examining the Contribution of the Responsive Classroom Approach. *Psychology in the Schools*, *44*, 397-413.

Rimm-Kaufman, S.E., Fan, X., Chiu, Y., & You, W. (2007). The contributions of the Responsive Classroom Approach on children's academic achievement: Results from a three year longitudinal study. *Journal of School Psychology, 45,* 401-421.

Rotter, J.B. (1982). Social Learning Theory. In N.T. Feather (Ed.), *Expectations and actions: Expectancy-Value models in psychology* (pp. 301-323). Hillsdale: Lawrence Erlbaum Associates.

Satcher, D. (2001). *Youth Violence: A report of the Surgeon General.* Washington, DC: U.S. Department of Health and Human Services.

The Wing Institute (2010). Wing Institute Roadmap. Retrieved July 16th, 2010 from the Wing Institute website: http://www.winginstitute.org/.