ABSTRACT

Title of Thesis: EFFECT OF INSTRUCTIONAL CONSULTATION ON ACADEMIC ACHIEVEMENT IN THIRD THROUGH FIFTH GRADE

Kristi S. Maslak, Master of Arts, 2011

Thesis directed by: Associate Professor William Strein
Department of Counseling and Personnel Services

The present study evaluated the effect of Instructional Consultation (Rosenfield, 1995) on the academic achievement of third through fifth grade students. Students whom teachers did (n = 201) and did not (n = 8119) select as the focus of consultation were balanced on their estimated propensity to be selected using logistic regression of observed covariates. Multilevel modeling compared students in the two treatment conditions on teacher assigned grades and standardized measures of reading and math, net of prior achievement. A small, but statistically significant negative effect of the program (d = -.13) was found for standardized measures of math. No significant differences were found on the other outcome measures. Limitations include model misspecification, missing data, and treatment diffusion.

EFFECT OF INSTRUCTIONAL CONSULTATION ON ACADEMIC ACHIEVEMENT IN THIRD THROUGH FIFTH GRADE

by

Kristi S. Maslak

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College park in partial fulfillment
of the requirements for the degree of
Master of Arts
2011

Advisory Committee:

Associate Professor William Strein, Chair
Professor Sylvia Rosenfield
Assistant Professor Jeffrey Harring

Table of Contents

List of Tables

Chapter 1: Rationale and Overview of Literature

Current practice goals for the specialty of school psychology are twofold: (a) to improve the academic and social-emotional development of all students, and (b) to build capacity within educational systems to foster development and prevent dysfunction (Yssledyke, et al., 2006). Consultation with classroom teachers, specialists, or administrators, rather than directly intervening with individual students, provides the most efficient means through which school psychologists can achieve both goals (Bradley-Johnson & Dean, 2000; Ehrhardt-Padgett, Hatzichristou, Kitson, & Myers, 2004; Gutkin & Curtis, 1999). Through consultation, school psychologists can help school professionals to apply the knowledge and skills needed to address and prevent academic and social-emotional difficulties among the students with whom they interact.

Partly due to variability across the models currently driving the practice of consultation and the research methods used to evaluate its effectiveness, the evidence base for consultation in the schools has been characterized as "promising but underdeveloped" (Erchul & Sheridan, 2008, p. 3). Most evidence in support of consultation in the schools centers on the application of behavior models of consultation that address student behavior problems and use experimental or single-subject designs that minimize the plausibility of threats to causal inference (Sheridan, Welch, & Orme, 1996). However, research on the effect of other models of consultation or their effect on academic achievement is less common, and when conducted, these studies often apply research methods that allow credible, alternative causal explanations to remain.

Instructional Consultation Model

Instructional Consultation (Rosenfield, 1995) is a consultee-centered model of consultation that aims to improve student academic performance, decrease overall referrals and disproportionate minority referrals to special education, and to enhance teachers' instructional practices through a multi-stage problem solving collaboration between the teacher and a trained instructional consultant. According to Rosenfield (1995; 2005), student learning in the classroom results from an interaction among the student's prior knowledge, the task demands, and the instruction delivered. When a student fails to meet teacher expectations for learning, Rosenfield's Instructional Consultation (IC) model assumes an ecological mismatch, namely an incongruous relationship among elements of this three-part instructional triangle (Gravois, Rosenfield, & Gickling, 1999). Therefore, identifying the relational mismatch and creating balance among the student's knowledge, task demands, and instruction are the focus of consultation within the IC model.

The process of IC described by Rosenfield (1995; 2005) includes five stages: contracting, problem identification and analysis, intervention planning, intervention implementation and evaluation, and closure. At contracting, the instructional consultant responds to the teacher's request for assistance, explains the assumptions of IC, and describes the collaborative, data-based process. During problem identification and analysis, the teacher and instructional consultant operationally define the presenting problem within the context of the instructional triangle (Gravois et al., 1999), use Instructional Assessment (Gravois & Gickling, 2008) to establish a baseline measure of the student's current level of performance, and clarify performance goals. Throughout

the intervention planning and implementation stages, the teacher and instructional consultant pool knowledge about research-based instructional practices to design and implement interventions, regularly collect data to monitor student progress, evaluate intervention effectiveness, and modify operationally defined problems or interventions if needed. During the final stage, closure, the teacher and instructional consultant agree to end their current case because stated goals are successfully attained or because both agree that a referral for additional support services, such as special education, is warranted.

Rosenfield and Gravois (1996) developed a multidisciplinary team model (IC Teams) to support and sustain the delivery of IC in schools. Within a school, the IC Team is composed of general educators, special educators, school administrators, school psychologists, and school social workers who are trained in the process of IC. According to Rosenfield and Gravois (1996; 1999), IC Teams differ from other problem solving team models in that the relationship between an individual team member and the teacher requesting assistance, rather than the team, operates as the primary forum for problem solving. Therefore, team members assume the role of case managers, and the team functions as a resource for targeted problem-solving and team member training. Through this case management approach to problem solving teams, Rosenfield, Silva, and Gravois (2008) suggest that IC Teams expands the capacity of schools to address the needs of students and staff, thereby diffusing and enhancing IC's hypothesized treatment effects.

Instructional Consultation Research

Until recently, the research literature on IC and IC Teams has relied on quasi-experimental methods to evaluate the effect of the program on special education referral patterns, with limited studies of the program's effect on student academic achievement.

Based on data from three separate pre-post studies, Gravois and Rosenfield (2002) conclude that IC Teams reduces the number of special education referrals and placements, and that fewer referrals and placements are made through the IC Teams process than concurrently operating pre-referral teams. In another pre-post study, Gravois and Rosenfield (2006) conclude that IC Teams decreases the risk and odds of minority special education placement compared to non-IC Teams schools. However, this research did not control several threats to internal validity common in pre-post quasi-experimental studies, namely history, maturation, selection, or interactions of selection with other threats. Using an interrupted time-series design, which is more robust in controlling threats to validity from history and maturation, but not selection, Newman (2007) did not find differences in special education referral patterns between IC Teams and non-IC Teams schools. Because these studies did not control for possible systematic differences between groups due to the method of treatment assignment, evidence for an effect of IC Teams on special education referral patterns is inconclusive.

Two studies have specifically considered the effect of IC Teams on student academic achievement. Using a pre-post design, school-system-developed criterion-referenced measures of reading achievement, and a small sample size ($N = 37$), Levinsohn (2000) compared the differential effect on second grade reading achievement when students were served through an IC Team or a Student Support Team (SST). Although the goal of both teams was to facilitate problem-solving and address student reading difficulties, the context of problem-solving differed with IC Teams focusing on case management and SST utilizing team meetings. Levinsohn found that all students made pre-post gains in reading achievement, but gains did not differ between the IC

Team and SST conditions after controlling for prior student achievement. However, Levinsohn's power to detect an effect was limited by the small sample size, and findings are not likely to generalize beyond second grade students due to sample restriction.

Using a larger sample size ($N = 5942$) of fourth and fifth grade students in 28 schools and multilevel modeling to account for the nesting of students and classrooms within schools, Silva (2007) compared scores on state-wide, standardized, criterion-referenced measures of reading achievement between students attending IC Teams schools and students attending schools in which IC Teams were not implemented. While Silva did not find an effect for attending an IC Team school on reading scores on students, a significant positive effect was found on average classroom reading levels. This finding suggests that IC Teams may have a positive effect on reading achievement at the aggregated classroom level, rather than the individual level. However, comparing the coefficients and standard errors of the multilevel models that included only level-one predictors with models that included both level-one and level-two predictors suggests the presence of multicolinearity and raises questions about the validity of the inferences that can be made from the findings. Moreover, the treatment and no-treatment schools were grossly nonequivalent on several variables that are likely related to reading achievement, including percent of students receiving free and reduced meals, second language learners, and ethnic minority (non-white, non-Asian) status. Because between-school differences for these variables were not controlled at the school level, threats to validity from selection and interactions of selection with other threats remain plausible.

Selection bias and its interactions with other threats to validity are salient in all five of the previously described studies on IC Teams. Therefore, systematic variation

between conditions remains a plausible explanation of the findings. A four-year, randomized-control-trial of the effect of IC Teams (Rosenfield & Gottfredson, 2004) has recently come to a close. Analysis of the data from this large scale study in which schools were randomly assigned to conditions has examined effects of the program on students and teachers during the final year of implementation, net of baseline performance, using both intent-to-treat-schools and intent-to-treat-students models.

The intent-to-treat-schools model considers the schools that were randomized to treatment and control conditions during the baseline year and the students within those schools during the final year of implementation. Multilevel modeling of the intent-to-treat-schools model did not find any effects of IC Teams on standardized measures of reading or math achievement, and the effect on teacher assigned grades was mixed such that significant positive effects were found for grades in reading and math among third grade students, while significant negative effects were found for teacher assigned grades in reading among fourth grade students (Bruckman, Vu, & Vaganek, 2010).

Analysis using the intent-to-treat-students model considers the students in the final year of implementation who attended treatment and control schools during the baseline year in which schools were randomly assigned. This analytical approach found small, but statistically significantly negative effects of IC Teams on standardized measures of reading among third and fourth grade students, and on math among third and fifth grade students (Bruckman et al., 2010). Furthermore, a statistically significant negative effect of IC Teams was found for teacher assigned grades in reading among fourth grade students. However, intent-to-treat-students models tend to slightly overestimate effects because differential attrition may introduce bias.

Regarding the effects of IC Teams on teachers, multilevel modeling of both intent-to-treat-teachers and intent-to-treat-schools models found significant positive effects of the program on teacher efficacy for general education teachers, and on collaboration for other educators (Experimental Evaluation of Instructional Consultation Teams, 2010b). Additional analysis did not find an interaction between levels of teacher use of consultation and teacher efficacy or collaboration (Experimental Evaluation of Instructional Consultation Teams, 2010a).

The average percentage of general education teachers within each school who sought IC Teams support ranged from 19% (*SD* = 12) during Year 1 Intervention (2006-07) to 48% (*SD* = 16) during Year 3 Intervention (2008-09) (Berger et al., 2010). With these low and variable levels of IC Teams use, IC was not likely to diffuse sufficiently within and across schools to yield effects on the population of students that could be measured using the intent-to-treat models. However, this level of use may be sufficient for an evaluation of IC that considers the effect of the program on the students who were the specific focus of the teacher consultation. Because the school, and not the student, was the level of random assignment and unit of comparison in the randomized-control-trial, any analysis of this data that compares students who were and were not the specific focus of teacher consultation must apply quasi-experimental research methods.

Selection Bias in Quasi-Experiments

The fundamental advantage of randomized experiments over all other research designs resides in the random assignment of units to conditions. When units are randomly assigned to conditions, initial differences between groups are attributed to chance and bias is diminished; therefore, differential outcomes are likely due to treatment

effects (Shadish, Cook, & Campbell, 2002). When units have not been randomly assigned to conditions, as is the case with quasi-experiments, selection bias, or systematic differences between conditions resulting from treatment assignment, is a possible threat to the validity of causal inference.

Although it is possible, selection bias may not always remain a plausible threat to causal inference. One common method for reducing the plausibility of selection bias in quasi-experiments is matching. Matching involves equally distributing units with similar scores on a matching variable between treatment and control conditions. When scores are balanced across conditions, the matching variable no longer provides a plausible explanation for differential treatment outcomes. However, the number of required unit combinations increases exponentially with each matching variable considered, and simple matching procedures are not useful with a large number of matching variables.

Propensity Score Analysis

A modern statistical procedure, propensity score analysis, has been applied to quasi-experiments in fields such as medicine and psychiatry (Perkins, Tu, Underhill, Zhou, & Murray, 2000; Vanderweele, 2006), community mental health (Hodges & Grunwald, 2005; Ye & Kastukas, 2009), and education (Condron, 2008; Has-Vaughn, 2006; Hong & Yu, 20-08; Wu, West & Hughes, 2008) as a means to match subjects on a large number of selection and outcome variables.

Propensity score analysis uses observed covariates to estimate each subject's propensity for treatment. Specifically, a propensity score is the conditional probability that participant *i* will receive treatment $(Z_i = 1)$ as opposed to not receiving treatment $(Z_i = 0)$ given an observed covariate vector, $\mathbf{x}_i$, such that $e(\mathbf{x}_i) = \Pr(Z_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i)$

(Rosenbaum & Rubin, 1983; 1984). Because treatment assignment, *Z,* is a dichotomous variable with a binomial distribution, treatment propensity can be estimated using the following equation:

$$\text{logit}(Z) = \alpha + \sum_{k=1}^{K} \beta_k X_k \quad (1)$$

where *k* indexes observed covariates from 1 to K, and β_k is the regression parameter for variable X_k.

When participants are balanced on estimated propensity and observed covariates that determine the propensity score covary with selection and outcome variables, conditions can be compared with equivalent expectations on observed covariates and the covariates no longer pose a selection threat to causal inference (Rosenbaum & Rubin, 1983). Indeed, when a randomized experiment and a quasi-experiment were compared with and without the use of propensity score analysis, Luellen, Shadish, and Clark (2005) found that 73-90% of the observed bias between the randomized and quasi-experiments were reduced when propensity score analysis was applied.

Research Question

The purpose of this study is to evaluate the effect of IC on academic achievement using propensity score analysis to reduce the plausibility of threats to validity in the quasi-experiment from selection bias. Specifically, after Year 1 Intervention (2006-07), did students who were the focus of teacher consultation in IC Team schools receive higher end-of-year grades and standardized achievement scores, net of prior achievement, than other students in IC Team schools who were not the focus of consultation, but were balanced on estimated propensity to be the focus?

Chapter 2: Method

Participants

Data were collected from 45 public elementary schools within a suburban county in the mid-Atlantic region of the United States as part of a four-year experimental evaluation of IC Teams (Rosenfield & Gottfredson, 2004) conducted between the 2005-06 and 2008-09 school years. Of the schools, 11 had been implementing IC Teams for one to three years prior to the experiment and were not included in the experimental evaluation. The remaining 34 schools were matched on a risk composite, and schools from each matched pair were randomly assigned to treatment and control conditions such that 17 schools were assigned to each condition. As expected, post randomization checks found that treatment and control schools were equivalent on expectation for measured variables. The 11 non-experimental schools had higher proportions of students who were ethnic minorities, limited English proficient, and qualified for free and reduced meals (Bruckman et al., 2010; Silva, 2007; Vu et al., 2009).

The sample for this study includes third through fifth grade students ($N = 8320$) and their teachers ($N = 374$) within the 28 schools implementing IC teams during the 2006-07 academic year (see Table 1). Kindergarten through second grade students were not sampled because a different grading rubric was applied and the students did not participate in annual standardized assessments of academic achievement. Classroom teachers self-selected to receive support from the IC Team and selected students ($n = 201$) to be the focus of consultation. The teacher-selected students were identified by their unique student identification code recorded on the case tracking forms maintained in each school implementing IC Teams.

Table 1
Sample Characteristics

Teachers (N = 374)	%	Students (N = 8320)	%
Sex		Sex	
Female	89	Female	48
Male	11	Male	52
Ethnicity		Ethnicity	
Advantaged	87	Advantaged	44
Disadvantaged	13	Disadvantaged	56
Education		Grade Level	
Bachelor's degree	49	3rd grade	34
Master's degree	51	4th grade	33
Years Teaching		5th grade	33
1 year or less	7	Services	
2 to 5 years	32	Free/Reduced Meal	39
6 to 10 years	26	Special Education	12
11 to 20 years	19	ESOL	21
More than 20 years	16	IC Teams	2
Age		Age	
30 years or younger	31	Old for grade	18
31 to 40 years	22	Young for grade	2
41 to 50 years	17		
51 years or older	30		

Note. ESOL = English as a Second or Other Language. All percentages were rounded to the nearest integer.

Measures

Data for this study were measured during Pre-intervention Baseline (2005-06) and Year 1 Intervention (2006-07). Measures included school district maintained records for student and teacher demographic information, student grades, and standardized student achievement test scores. Additional measures included two teacher surveys that were administered as part of the experimental evaluation of IC Teams (Vu et al., 2009).

Demographics.

School district records provided measures of student and teacher characteristics, student enrollment status, and student services received. Characteristics of students and teachers were measured during Year 1 Intervention (2006-07) and included gender, ethnicity, and date of birth. Because Caucasian and Asian students are less likely to be

referred to interventions, such as special education, than are African American or Hispanic students (Artiles, Klinger, & Tate, 2006; O'Conner & Fernandez, 2006; Reid & Knight, 2006), ethnicity was recoded to provide a dichotomous measure of Caucasian/Asian ethnicity. Measures of student and teacher age were derived by subtracting date of birth from the date of the first day of school. Student measures of old for grade and young for grade were derived by comparing each student's age of entry with grade level age expectations. District criteria for the allowable age of entry to Kindergarten suggest an age of entry to the third, fourth, and fifth grades as 8, 9, and 10 years, respectively. Students whose age exceeded expectations were identified as old for grade, and those whose age did not reach expectations were identified as young for grade.

Student enrollment status was measured for both Pre-intervention Baseline (2005-06) and Year 1 Intervention (2006-07). Measures included the date of enrollment, grade level in which the student was enrolled, the number of days enrolled, the number of days attended, and if the student was retained at the end of the year. A measure of being new to the district was derived by identifying students who did not have enrollment data for Pre-intervention Baseline (2005-06). Date of enrollment was recoded to provide a dichotomous measure of students who enrolled with the district after the first quarter grading period. A measure of the proportion of days enrolled was derived by dividing the number of days enrolled from the total number of school days according to the school calendar. Similarly, a measure of the proportion of days absent was derived by dividing the number of days attended by the number of days enrolled.

Student services received were measured for both Pre-intervention Baseline (2005-06) and Year 1 Intervention (2006-07). The school district uses a comprehensive

coding system to record a student's limited English proficiency status, qualification for free and reduced meals, and special education codes. This system was simplified to provide four dichotomous measures indicating whether or not a student was considered limited English proficient, received English as a second or other language services, qualified for free and reduced meals, or qualified for special education.

Achievement.

Academic achievement was measured during both Pre-intervention Baseline (2005-06) and Year 1 Intervention (2006-07). Measures included quarterly assigned teacher grades and test scores from state-wide, annually administered, standardized achievement tests. In the academic domains of listening, oral language, art, physical education, music, handwriting, and technology, teachers assigned grades ranging from "N" (not meeting expectations) to "S+' (outstanding). In the academic domains of reading, writing, math, social studies, and science, teachers assigned grades ranging from "F" (failure) to "A" (outstanding). Grades were recoded from nominal to numerical values in the following manner: S+ or A = 4; S, B+, or B = 3; S-, C+, or C = 2; and N, D+, or F = 1. An overall measure of student grade point average (GPA) was derived by averaging the quarterly grades received across all 12 academic domains. Specific domain measures of GPA were derived for reading, writing, math, and listening by averaging the quarterly grades within each domain.

During the spring of each academic year, students in the third through fifth grades were administered standardized, state-wide assessments for reading and math achievement that were aligned with the state's standards of learning (SOL). Students received scale scores ranging from 200-600. The SOLs were developed using Item

Response Theory to equate the scales across years of implementation, but not vertically across grade levels (Virginia Department of Education, 2005). Therefore, SOL scores did not provide an absolute measure of academic achievement. Rather, within each academic domain, SOL scores provided a measure of academic achievement relative to grade-level expectations.

Teacher surveys.

Teachers completed two surveys that were administered online through the school district's intranet each February as part of the four-year experimental evaluation of IC Teams (Rosenfield & Gottfredson, 2004). Response rates for both surveys exceeded 80% for the 2005-06 and 2006-07 school years included in this study (Bruckman et al., 2010; Vu et al., 2009).

The Teacher Self Report (TSR) was a 100-item survey composed of researcher developed items as well as items adapted from Teschannen-Moran and Hoy (2001) and Byrk and Schneider (2003). Mean composites were derived from five-point Likert scale measures of a teacher's sense of efficacy when working with students, perception of collaboration among colleagues, job satisfaction, and instructional practices. Reliabilities for the Teacher Efficacy (α = .94 & .92), Collaboration (α = .88 & .82), Job Satisfaction (α = .92 & .92), and Instructional Practices (α = .90 & 91) composites were high for 2005-06 and 2006-07, respectively (Vu et al., 2009).

Additional TSR items asked teachers about their highest level of education attained, teaching licensure status and type of licensure, years working as a teacher, and years working as a teacher in the current school. Two measures were derived to indicate teachers with a Master's degree or higher, and teachers with a provisional or full

elementary (pre-K to 6) license. Years working as a teacher and years working in the current school were measured in the following manner: 1 = 1 year or less, 2 = 2 to 5 years, 3 = 6 to 10 years, 4 = 11 to 20 years, and 5 = 20+ years.

For each student, teachers completed the Teacher Report on Student Behavior (TRSB). The TRSB was a 36-item survey that measured student behavior in the classroom and teacher perceptions of the student relationship. Two researcher developed items measured the teacher's overall rating of a student's academic progress and classroom behavior using a five point Likert-scale. Remaining items were adapted from the Teacher Observation of Classroom Adaptation-Revised (TOCA-R; Werthamer-Larsson, Kellam, & Wheeler, 1991) and the Student-Teacher Relationship Scale (STRS; Pianta, 2001), which used four-point and five-point Likert scales, respectively. Mean composites for the TOCA-R and STRS items were derived. Reliabilities for the Internalizing Behavior (α = .85 & .85), Externalizing Behavior (α = .90 & .90), Concentration and Readiness to Work (α = .92 & .92), Closeness (α = .86 & .85), and Conflict (α = .86 & .87) composites were high for 2005-06 and 2006-07, respectively (Bruckman, Vu, & Vaganek, 2010).

Missing Data

A total of 43 variables from the student and teacher measures were identified for the propensity score estimation and treatment outcome analyses. However, of the student sample, 63% (n = 5246) were missing values on one or more variables. Furthermore, with missing data for 62% (n = 124) of the students selected as the focus of IC, there was not a significant relationship between treatment selection and missing data (r = .004). Because most statistical methods and software packages, including those planned for this

study, assume complete sample case data, several approaches for handling the problem of missing data were considered.

First considered was listwise deletion, whereby cases with one or more missing values are excluded from analyses. When the probability of missing data for any given variable is unrelated to the value of that variable and all other variables in the analysis, a condition known as missing completely at random (MCAR), the cases with complete data are assumed to be a random subsample of the full sample and parameter estimates will be unbiased (Allison, 2002; Schafer & Graham, 2002). However, in this study, data are known to be missing on account of at least two variables identified for the propensity analysis: a) students who were new to the school in 2006-07, and b) students who entered after the first quarter. Therefore, the data are not MCAR. When data are not MCAR, the cases with complete data no longer represent the full sample, and listwise deletion may introduce bias. Moreover, listwise deletion would have substantially reduced the size of the sample available for analysis, thereby reducing statistical power and inflating standard errors. As such, listwise deletion was not applied.

The remaining approaches that were considered for handling missing data required imputation, or the process of using observed data to fill in missing values and build complete case data (Allison, 2002). Single imputation methods build one complete set of case data through multiple regression, maximum likelihood (ML) estimation, or the expectation maximization (EM) algorithm, for example. Of the single imputation methods, ML estimation yields relatively unbiased parameter estimates when sample sizes are large, but it requires specialized software. Although multiple regression can be implemented with general use software, it requires large sample sizes and data that are

MCAR to yield unbiased estimates. The EM algorithm can be implemented with general use software and data need not be MCAR. Instead, the EM algorithm assumes that data are missing at random (MAR), such that the probability of missing data for a given variable is unrelated to the value of that variable after controlling for other variables in the analysis. While it is possible that data in this study are MAR, testing that assumption is not possible because missing values are unknown.

Multiple imputation (MI) is another method that can be implemented with general use software and assumes that data are MAR, but unlike the EM algorithm, which may yield standard errors that are biased downward, MI introduces random variance that adjusts standard errors upward and reduces bias (Allison, 2002). Therefore, missing data were imputed using MI, where missing values on *x* are predicted from known values on other variables such that the equation for imputing missing values on *x* given known values on *y* is as follows:

$$\tilde{x} = a + by_i + s_{x.y}u_i \quad (2)$$

where $s_{x.y}u_i$ is a random draw from the residual distribution of *x* for the i^{th} participant. For a single data file with missing values, different draws of $s_{x.y}u_i$ generate multiple sets of data in which missing values were imputed. When using MI data files, analyses are completed for each set of data, and results are pooled.

With moderate amounts of missing values, five imputed sets are sufficient to stabilize *p*-values and standard errors (Allison, 2009). Although data are missing for the majority of participants in this study, the mean number of missing values per participant was moderate (M = 5.55, SD = 6.22). Therefore, five imputed data sets were generated.

As Allison (2002) recommends, the imputation model was built from the pool of dependent and independent variables for estimating propensity and treatment effects. Measures from Pre-intervention Baseline (2005-06) and Year 1 Intervention (2006-07) were included to improve model fit. Imputations for continuous variables were constrained within allowable maximum and minimum values. Furthermore, multiple category variables were dummy coded. The dummy variable with the largest imputed value was assigned as the missing category value. A summary of measures included in the imputation model can be found in Appendix A.

The problem of missing data was further addressed when estimating treatment propensity. With a sample size of $N = 4500$, D'Agostino and Rubin (2000) demonstrated that including missing indicator dummy variables to estimate propensity equates participant expectations on patterns of missing data as well as observed covariates. Therefore, a set of dummy variables was derived for each predictor to indicate cases that had imputed values on that predictor. These dummy variables were then included as predictors when building the treatment propensity model.

Data Analysis

Estimating treatment propensity.

Propensity scores are most commonly estimated using logistic regression due to its advantages over other methods, including classification trees and ensemble methods (Luellen, 2007). First, logistic regression is relatively robust against violations of multivariate normality and homogeneity of variance-covariance matrices. Second, logistic regression can model curvilinear relationships between observed covariates and treatment assignment. Furthermore, Luellen demonstrated through a series of Monte

Carlo experiments that propensity scores derived through logistic regression were less likely to introduce selection bias and yielded a more precise adjusted estimate of treatment effects than other methods. Therefore, logistic regression was the propensity score estimation method chosen for this study.

When building the regression model, decisions for determining the pool of variables were guided by practical and theoretical relationships with treatment assignment rather than parsimony or statistical significance of a single predictor (Luellen et al., 2005; Rubin & Thomas, 1996). Because characteristics of teachers who did and did not seek IC support may differ, the initial pool of variables included teacher demographics, TSR, and TRSB measures. Furthermore characteristics of students whom teachers did and did not choose as the focus of consultation may differ because student academic and behavioral difficulties are a focus of IC, and because students have historically been disproportionately referred to interventions, such as special education, according to student demographic characteristics (Artiles, Klinger, & Tate, 2006; O'Conner & Fernandez, 2006; Reid & Knight, 2006). Therefore, student grades, enrollment, support services received, and demographics were included when building the model. Initial differences between students who were and were not the focus of IC on the predictor variables are summarized in Table 2, and for missing value dummies in Table 3.

The regression model included a pool of 66 variables, with 38 predictor variables and 28 missing value dummy indicators. Variables with the greatest likelihood of contributing to model fit were identified using the backward stepwise logistic regression procedure of SPSS Statistics 17.0 for Windows (IBM, 2008). Because Cramer (1999)

Table 2

Initial Differences on Measures for IC and Not IC Students

Measure	IC		Not IC					
	M	*SD*	*M*	*SD*	*t*	*df*	χ^2	*d*
Student								
Gender (female)	.37	.49	.48	.50			8.59	-.22*
Advantaged Ethnicity	.43	.50	.44	.50			.14	-.02
Limited English Proficient	.21	.41	.27	.44			3.57	-.14†
Old for Grade	.23	.42	.18	.38			3.25	.12†
Young for Grade	.02	.16	.02	.15			.08	.00
Free and Reduced Meals	.35	.48	.39	.49			.89	-.08
Special Education[a]	.13	.34	.13	.34			.02	.00
English as Second Language	.17	.38	.21	.41			2.36	-.10
New to District	.15	.36	.14	.35			.16	.03
Entered after 1st Quarter	.07	.26	.08	.27			.05	-.04
Proportion Days Enrolled[a]	.95	.17	.95	.16	.08	7156		.00
Proportion Days Absent[a]	.04	.04	.03	.03	-.74	7156		.06
Retained at End of Year[a]	.01	.08	.01	.07			.02	.00
Third Grade	.25	.44	.34	.47			6.21	-.20*
Fourth Grade	.54	.50	.33	.47			41.63	.43***
Fifth Grade	.20	.40	.34	.47			15.55	-.32***
Listening[b]	2.87	.69	3.11	.59	5.71	7713		-.37***
Math[b]	2.71	.98	3.05	.82	5.71	7688		-.38***
Reading[b]	2.61	.89	3.03	.80	7.15	7688		-.50***
Writing[b]	2.78	.89	3.15	.80	5.70	7678		-.44***
GPA[a]	2.91	.36	3.10	.35	6.86	7149		-.54***
Global Progress[a]	3.39	1.15	3.85	1.06	4.91	5567		-.42***
Global Behavior[a]	3.85	1.08	4.04	1.02	2.07	5535		-.18*
Concentration[a]	1.71	.73	2.04	.72	5.22	5588		-.46***
Externalizing[a]	.33	.48	.29	.46	-1.00	5588		.09
Internalizing[a]	.61	.48	.57	.51	-.73	5588		.08
Closeness[a]	3.14	.73	3.19	.78	.73	5582		-.07
Conflict[a]	.68	.94	.53	.90	-1.84	5582		.16†

Table 2. (continued)

Measure	IC		Not IC					
	M	*SD*	*M*	*SD*	*t*	*df*	χ^2	*d*
Teacher								
Gender (female)	.81	.40	.89	.31			14.22	-.22***
Advantaged Ethnicity	.92	.28	.87	.34			3.32	-.16†
Age	34.54	9.49	40.29	12.40	6.24	7044		-.52***
Years Teaching	2.58	.82	3.08	1.18	5.85	7214		-.49***
Years at School	2.03	.69	2.30	1.07	3.59	7214		-.30***
Elementary Licensure	1.00	.00	.98	.13			3.32	.22†
Master's Degree or Higher	.48	.50	.51	.50			.81	-.06
Efficacy[a]	4.16	.35	4.19	.45	.70	6030		-.07
Collaboration[a]	3.87	.68	4.13	.69	4.65	6030		-.38***
Job Satisfaction[a]	4.29	.55	4.39	.71	1.89	6030		-.16†
Instructional Practices[a]	3.72	.44	3.84	.52	2.96	6004		-.25*

Note. Chi square is computed for dichotomous variables, and $df = 1$. Effect size, *d*, is calculated as $d = (M_{IC} - M_{No\ IC}) / \sigma_{pooled.}$

[a]Measured during Pre-Intervention Baseline (2005-06). [b]First Quarter Grades Measured during Year 1 Intervention (2006-07).

†p < .10. *p < .05. ***p <.001.

Table 3
Initial Differences on Missing Values for IC and Not IC Students

	IC		Not IC			
Measure	*M*	*SD*	*M*	*SD*	χ^2	*d*
Student Measures						
FARM	.00	.00	.00	.01	.03	.00
Special Education	.15	.36	.14	.35	.16	.03
Proportion Days Enrolled	.15	.36	.14	.35	.16	.03
Proportion Days Absent	.15	.36	.14	.35	.16	.03
Retained at End of Year	.16	.37	.15	.36	.27	.03
Listening	.04	.20	.07	.26	3.31	-.13†
Math	.04	.21	.08	.27	2.82	-.17†
Reading	.04	.21	.08	.27	2.82	-.17†
Writing	.04	.21	.08	.27	3.00	-.17†
GPA	.14	.35	.14	.35	.00	.00
Global Progress	.34	.47	.33	.47	.06	.02
Global Behavior	.34	.48	.33	.47	.07	.02
Concentration	.33	.47	.33	.47	.03	.00
Externalizing	.33	.47	.33	.47	.03	.00
Internalizing	.33	.47	.33	.47	.03	.00
Closeness	.33	.47	.33	.47	.02	.00
Conflict	.33	.47	.33	.47	.02	.00
Teacher Measures						
Gender	.00	.07	.03	.17	4.14	-.23*
Advantaged Ethnicity	.16	.37	.06	.24	32.21	.32***
Age	.08	.28	.15	.36	7.46	-.22*
Years Teaching	.04	.20	.13	.34	15.44	-.53***
Years at School	.04	.20	.13	.34	15.44	-.53***
Elementary Licensure	.10	.30	.20	.40	12.33	-.28***
Level of Education	.05	.22	.14	.35	14.40	-.31***
Efficacy	.15	.36	.28	.45	15.07	-.32***
Collaboration	.15	.36	.28	.45	15.07	-.32***
Job Satisfaction	.15	.36	.28	.45	15.07	-.32***
Instructional Practices	.16	.37	.28	.45	14.51	-.29***

Note. Participants did not have missing values for gender, advantaged ethnicity, limited English proficient, old for grade, young for grade, English as second language, new to district, entered after 1st quarter, and grade level. Effect size, *d*, is calculated as $d = (M_{IC} - M_{No\ IC}) / \sigma_{pooled.}$
†p < .10. *p < .05. ***p <.001.

found that adjusting the logistic regression cut point to match sampling proportions can improve case classification when group sample sizes are grossly unbalanced, as is the case with this study, the default cut point of .50 was changed to .976 to reflect that only 2.4% of participants were the specific focus of IC.

Propensity scores were stratified into five strata, and dummy variables were derived to indicate strata for each participant. Rosenbaum and Rubin (1994) recommend stratifying the propensity score into quintiles, or five strata, so that the propensity score distribution for participants in the treatment and no treatment groups are similar within strata. Furthermore, Cochran (1968) found that approximately 90% of the bias from a single continuous variable can be reduced with five strata.

Validating the propensity model.

Following the procedure described in Luellen et al. (2005), the propensity model was validated to assess whether participants were equated across treatment groups within strata on observed covariates. Each covariate from the pool of predictor variables was subject to analysis as a dependent variable to determine if the covariate differed between groups. An ANCOVA model (2 groups x 5 strata) with all two-way interactions was evaluated, and analyses were re-run after dropping non-significant interaction terms. Statistical significance at $p < .05$ for both main and interaction effects was considered. Because multiple analyses were performed, it was expected that 5% of the results ($n = 3$ variables) would be statistically significant by chance alone.

Evaluating treatment effects.

Students are nested within classroom teachers, and multilevel modeling with the Hierarchical Linear Modeling program (HLM 6.08; described in Raudenbush & Bryk, 2002) was used to evaluate the effect of IC on academic achievement during Year 1 Intervention (2006-07). Individual students comprise Level I, whereas classroom teachers comprise Level II. Analyses were conducted separately for each dependent measure, which were 4th quarter grades and SOL scores in reading and math. One factor

with two levels was whether or not the student was the focus of consultation during Year 1 Intervention (2006-07). A second factor was estimated propensity with five levels to equate participants across groups on observed covariates. Third grade students did not take SOLs during Pre-intervention Baseline (2005-06); therefore, first quarter grades from Year 1 Intervention (2006-07) in the same domain as the dependent measure were used to control for prior achievement when evaluating both grades and SOL scores. Among fourth and fifth grade students, correlations between first quarter grades and prior SOL scores were moderate (r_{math} = .582; r_{read} = .487).

Dichotomous variables were entered in the model uncentered, and prior achievement was entered group mean centered. The homogeneity of student level slopes was tested by entering predictors into the model with their slopes free to vary. Slopes that did not significantly vary between classrooms at $p < .10$ were fixed. Equation 3 describes the mixed model used to evaluate the effect of being the focus of IC on academic achievement when all slopes were left free to vary.

$$Y_{ij} = \gamma_{0j} + \gamma_{1j}X_1 + \gamma_{2j}X_2 + \gamma_{3j}X_3 + \gamma_{4j}X_4 + \gamma_{5j}X_5 + \gamma_{6j}X_6 + r_{ij} + u_{0j} + u_{1j} + u_{2j} + u_{3j} + u_{4j} + u_{5j} + u_{6j} \quad (3)$$

where, Y_{ij} was the measure of academic achievement for the i^{th} student in the j^{th} classroom,

γ_{0j} was the unadjusted mean achievement in classroom j,

γ_{1j} was the effect of being the focus of an IC case,

γ_{2j} to γ_{5j} were the effects of strata,

γ_{6j} was the effect of prior achievement,

X_1 was student treatment assignment,

X_2 to X_5 were student indicators for strata 2 through strata 5,

X_6 was the student measure of prior achievement,

r_{ij} was the residual error for student *i* in the j^{th} classroom, and

u_{1j} to u_{6j} were the residual errors for the j^{th} classroom.

Chapter 3: Results

Treatment Propensity

Of the 66 variables entered into the backward stepwise logistic regression, 29 variables were retained as contributing to model fit in at least one imputed data set (see Table 4). Students in the fourth grade with a close teacher relationship, and who had a teacher with a Master's degree or higher reporting higher than average self-efficacy for teaching and job satisfaction, were more likely to be selected as the focus of consultation. Students who received free and reduced meals or ESOL services, were new to the district, maintained lower than average grades, were rated by teachers as maintaining lower than average concentration, externalizing, or internalizing behaviors, and who had a male or younger than average teachers who were either new to teaching or to teaching at the school and reported lower than average collaboration and good instructional practices, were less likely to be selected as the focus of consultation. Furthermore, students with missing data about their teacher's ethnicity were more likely to be selected as the focus of consultation, while students with missing data about their listening grades, grade point average, teacher's sense of efficacy, and teacher's teaching experience were less likely to be selected. However, these conclusions should be interpreted with caution as the retained model did not predict any students as being the focus of a teacher consultation despite the use of an adjusted cut value (see Table 5).

Although treatment selection was not adequately modeled, the estimated propensity scores could be applied in further analyses to control for the retained variables. Therefore, the propensity scores were stratified into quintiles and validation checks were conducted. Because the model failed to classify any students as being

Table 4

Variables Retained Across Imputations when Estimating Propensity with Backward Stepwise Logistic Regression

	Minimum *p* value				Maximum *p* value			
Variable	β	*SE*	Wald	Exp(β)	β	*SE*	Wald	Exp(β)
Student								
Free and Reduced Meals	-.53	.171	9.51**	.59	-.44	.172	6.61**	.64
English as Second Language	-.69	.212	10.46**	.50	-.62	.211	8.76**	.54
New to District	-2.45	.831	8.70**	.09	-.19	.823	5.60*	.14
Proportion Days Absent[a, c]					-2.47	1.509	2.69	1.14
Fourth Grade	1.09	.156	48.92***	2.99	1.04	.154	45.74***	2.84
Listening[b]	-.35	.124	8.05**	.70	-.24	.130	3.50†	.78
Math[b]	-.24	.113	4.33*	.79	-.20	.114	3.09†	.82
Reading[b]	-.32	.108	9.02**	.72	-.23	.109	4.27*	.80
Writing[b]	-.35	.114	9.59**	.70	-.29	.123	5.42*	.75
GPA[a]	-.77	.279	7.62**	.46	-.52	.309	2.88†	.59
Global Behavior[a, c]					.13	.073	3.04†	1.14
Concentration[a]	-.60	.164	13.20***	.55	-.30	.167	3.11†	.74
Externalizing[a]	-.55	.177	9.59**	.58	-.41	.171	5.83*	.66
Internalizing[a]	-.57	.175	10.49**	.57	-.27	.157	2.96†	.76
Closeness[a]	.21	.112	3.55†	1.23	.19	.105	3.35†	1.21
Teacher								
Gender (female)	-.73	.211	11.83***	.48	-.57	.218	6.83**	.57
Age	-.02	.010	3.37†	.98	-.02	.010	2.94†	.98
Years Teaching	-.52	.078	43.28***	.60	-.30	.107	7.94**	.74
Years at School	-.30	.124	5.74*	.74	-.26	.124	4.49*	.77
Master's Degree or Higher	.57	.166	11.58***	1.76	.42	.165	6.58*	1.53
Efficacy[a]	.62	.222	7.85**	1.86	.43	.217	3.96*	1.54
Collaboration[a]	-.37	.125	8.78**	.69	-.23	.126	3.31†	.77
Job Satisfaction[a]	.27	.127	4.58*	1.31	.23	.125	3.51†	1.25
Instructional Practices[a]	-.59	.175	11.34***	.56	-.33	.176	3.48†	.72

Table 4. (continued)

	Minimum *p* value				Maximum *p* value			
Variable	β	*SE*	Wald	Exp(β)	β	*SE*	Wald	Exp(β)
Missing Indicator								
Listening[b]	-1.86	.455	16.75***	.16	-1.07	.431	6.22*	.34
GPA[a]	-2.15	.836	6.63**	.12	-1.77	.834	4.51*	.17
Teacher Ethnicity	1.59	.232	47.09***	4.92	1.49	.238	39.37***	4.44
Years Teaching	-1.79	.386	21.48***	.17	-1.18	.393	9.01**	.31
Efficacy[a]	-1.02	.210	23.48***	.36	- .83	.217	14.60***	.44

Note. Parameter estimates differed across imputations. *Df* = 1.

[a]Measured during Pre-Intervention Baseline (2005-06). [b]First quarter grades measured during Year 1 Intervention (2006-07).

[c]Variable retained in fewer than two imputations.

[†]p < .10. *p < .05. **p < .01 ***p < .001.

Table 5

Classification of Students as the Focus of an IC Case

	Predicted		Percent Correct
Observed	Not IC	IC	
Not IC	8119	0	100
IC	201	0	0

Note. Cut value is .976

Table 6

Student Participants and Treatment Propensity Ranges by Strata

	IC		Not IC		Propensity	
Strata	*n*	%	*n*	%	Minimum	Maximum
1	4.0	1.99	1660.0	20.45	.000	.004
2	10.4	5.17	1653.6	20.37	.004	.008
3	16.4	8.16	1647.6	20.29	.008	.015
4	39.2	19.50	1624.8	20.01	.015	.033
5	131.0	65.17	1533.0	18.88	.033	.694

Note. Results are pooled across imputations.

selected for treatment, the distribution of propensity scores and the number of cases within each strata was highly skewed such that 80% of the participants (n = 6656) had less than a 3.3% chance of being selected for IC (see Table 6). When validating the propensity score model, results from five, or 8%, of the validation analyses were statistically significant. Variables that continued to differentiate students who were selected as the focus of consultation from students who were not selected included teacher measures of age, gender, teaching experience, collaboration, and the missing value indicator for advantaged ethnicity. Although the number of significant analyses was greater than was likely to occur through chance alone, the number of measures did not grossly deviate from chance expectations. Therefore, the propensity model was retained without adjustments.

Treatment Effects

To determine if multilevel modeling was necessary, intraclass correlations (ICCs) were calculated separately for each outcome measure by running an unconditional model without predictors. The ICCs indicated that between group variance accounted for 13-17% of the total outcome measure variance, and it was determined that multilevel modeling was appropriate (see Table 7).

Table 7
Intraclass Correlations and Reliabilities for Outcome Measures

Measure	τ	σ^2	ICC	λ
4th Quarter Math	.11	.69	.13	.74
4th Quarter Reading	.12	.64	.16	.77
Math SOL	979.13	5566.12	.15	.76
Reading SOL	1009.47	5042.13	.17	.78

Note. Some students (n = 134) were missing unique teacher identifiers and analyses were run with N = 8186 students. Tau (τ) is the between group variance. Sigma squared (σ^2) is the within group variance. The intraclass correlation, or ICC, is the proportion of total variance accounted for by between group variance and is calculated as $\tau / (\tau + \sigma^2)$. λ is the Lambda reliability.

The effects of being the focus of an IC case on academic achievement are summarized in Tables 8 through 11. With $p < .05$, the effect of being the focus of an IC case on fourth quarter math grades, fourth quarter reading grades, and reading SOL scores was not statistically significant ($p = .27$, .49; and .17, respectively). However, a statistically significant negative effect was found on math SOL scores ($p = .04$). Net of prior achievement and controlling for propensity strata, average math SOL scores were 11.54 points lower for students who were the focus of an IC case. The size of this effect was small ($d = -.13$).

Table 8
Effect of Being the Focus of an IC Case on Math Grades

Fixed Effect	γ	*SE*	*t Ratio*
Intercept, γ_{00}	3.09	.09	36.24***
Focus of IC Case, γ_{01}	-.07	.07	-1.11
Strata 2, γ_{02}	-.11	.04	-3.23*
Strata 3, γ_{03}	-.25	.06	-4.27*
Strata 4, γ_{04}	-.41	.06	-6.98***
Strata 5, γ_{05}	-.57	.06	-5.23*
Prior Math Achievement, γ_{06}	.46	.05	9.92***
Random Effect	Variance Component	*df*	χ^2
Intercept, u_{0j}	.14	26	54.64*
IC Case, u_{1j}	.06	26	40.39*
Strata 2, u_{2j}	.03	26	39.05*
Strata 3, u_{3j}	.07	26	41.58*
Strata 4, u_{4j}	.10	26	38.99*
Strata 5, u_{5j}	.14	26	37.75†
Prior Math Achievement, u_{6j}	.03	26	41.58*
Residual Error, r_{ij}	.44		

†p < .10. *p < .05. ***p <.001.

Table 9

Effect of Being the Focus of an IC Case on Math SOL Scores

Fixed Effect	γ	*SE*	*t Ratio*
Intercept, γ_{00}	496.03	6.00	82.63***
Focus of IC Case, γ_{01}	-11.54	5.54	-2.08*
Strata 2, γ_{02}	-5.66	3.76	-1.50
Strata 3, γ_{03}	-17.61	6.54	-2.69*
Strata 4, γ_{04}	-27.44	6.85	-4.00*
Strata 5, γ_{05}	-41.30	9.90	-4.17*
Prior Math Achievement, γ_{06}	44.92	2.78	16.15***
Random Effect	Variance Component	*df*	χ^2
Intercept, u_{0j}	1443.98	108	269.72***
Strata 2, u_{2j}	334.94	108	139.83*
Strata 3, u_{3j}	711.17	108	163.46*
Strata 4, u_{4j}	856.04	108	157.70*
Strata 5, u_{5j}	799.44	108	151.14*
Prior Math Achievement, u_{6j}	159.06	108	169.60***
Residual Error, r_{ij}	3544.37		

Note. Focus of IC Case did not significantly vary between classrooms. Effect size, *d*, is calculated as 2*t*/*df*.

†p < .10. *p < .05. ***p <.001.

Table 10

Effect of Being the Focus of an IC Case on Reading Grades

Fixed Effect	γ	*SE*	*t Ratio*
Intercept, γ_{00}	3.14	.08	41.86***
Focus of IC Case, γ_{01}	-.05	.07	-.69
Strata 2, γ_{02}	-.13	.05	-2.63*
Strata 3, γ_{03}	-.28	.05	-5.37***
Strata 4, γ_{04}	-.45	.07	-6.51***
Strata 5, γ_{05}	-.66	.07	-9.19***
Prior Reading Achievement, γ_{06}	.33	.03	12.06***
Random Effect	Variance Component	*df*	χ^2
Intercept, u_{0j}	.17	108	226.29***
Strata 2, u_{2j}	.05	108	133.51*
Strata 3, u_{3j}	.11	108	139.48*
Strata 4, u_{4j}	.19	108	168.45***
Strata 5, u_{5j}	.18	108	140.96*
Prior Reading Achievement, u_{6j}	.03	108	169.13***
Residual Error, r_{ij}	.46		

Note. Focus of IC Case did not significantly vary between classrooms.

†p < .10. *p < .05. ***p <.001.

Table 11

Effect of Being the Focus of an IC Case on Reading SOL Scores

Fixed Effect	γ	*SE*	*t Ratio*
Intercept, γ_{00}	478.24	5.17	92.50***
Focus of IC Case, γ_{01}	-7.80	5.63	-1.39
Strata 2, γ_{02}	-5.28	3.72	-1.42
Strata 3, γ_{03}	-15.92	4.47	-3.57*
Strata 4, γ_{04}	-23.18	6.75	-3.44*
Strata 5, γ_{05}	-35.46	8.08	-4.39*
Prior Reading Achievement, γ_{06}	33.40	2.56	13.03***
Random Effect	Variance Component	*df*	χ^2
Intercept, u_{0j}	1553.59	108	267.64***
Strata 2, u_{2j}	469.14	108	144.70*
Strata 3, u_{3j}	869.27	108	147.94*
Strata 4, u_{4j}	1145.03	108	139.03*
Strata 5, u_{5j}	1290.12	108	139.07*
Prior Reading Achievement, u_{6j}	127.35	108	143.97*
Residual Error, r_{ij}	3773.60		

Note. Focus of IC Case did not significantly vary between classrooms.

$^{\dagger}p < .10$. *p < .05. ***p <.001.

Chapter 4: Discussion

Instructional Consultation (Rosenfield, 1995) and its multidisciplinary team-based model of delivery, IC Teams (Rosenfield & Gravois, 1996), aim to improve student academic performance, decrease overall referrals and disproportionate minority referrals to special education, and to enhance teachers' instructional practices. Until recently, research on IC and IC Teams has used quasi-experimental methods that did not adequately address the problem of selection bias to evaluate the effect of the program on special education referral practices (Gravois & Rosenfield, 2002; 2006; Newman, 2007) or student reading achievement (Levinsohn, 2000; Silva, 2007). A randomized-control study of the effect of IC Teams has come to a close; however, levels of program use may not have been sufficient to yield measurable effects on the population of students (Berger et al., 2010).

The present study evaluated the effect of IC on the students who were the specific focus of a teacher consultation and is the first quasi-experimental study of IC or IC Teams to reduce selection threats to causal inference by applying propensity score analysis. Specifically, this study evaluated the effect of IC on the reading and math achievement in the third through fifth grade by comparing students who were and were not selected as the focus of the consultation, but were balanced in their estimated propensity to have been selected. Results using multilevel modeling did not find statistically significant effects of IC on standardized measures of reading or teacher assigned grades in reading or math. However, a small, but statistically significant negative effect (d = -.13) was found on standardized measures of math.

These findings of no effects or slightly negative effects of the program on academic achievement are consistent with the intent-to-treat-students analyses from the recent randomized-control evaluation of IC Teams (Bruckman et al., 2010). While findings do not suggest that IC has a significant positive effect on academic achievement, the finding of a negative effect on a single measure of math achievement does not suggest that IC interfered with student learning. According to a summary program report (Gravois, Nelson, & Sherry, 2007), the majority of IC cases during Year 1 Intervention (2006-07) addressed student reading, writing, organizational, or behavioral concerns. Only 25% of the IC cases addressed student math concerns. Given the small percentage of cases that provided direct support to teachers and indirect support to students in the area of math, measurable positive effects would not have been likely. Furthermore, the 11-point difference in average math SOL scores between selected and non-selected students represents only 3% of the total possible range of scores. Therefore, this effect, while statistically significant, is not likely to be of practical importance.

Limitations.

Several problems limit the validity of inferences that can be made from the results of this study. First, the treatment propensity estimation model was poorly fit. Despite an adjustment in the cut value, the model did not classify any cases as having been selected as a focus of IC, and the distribution of propensity scores was highly positively skewed. Furthermore, after being balanced on estimated propensity, selected and non-selected students statistically significantly differed on slightly more covariates than would have been expected by chance alone. Although participants were balanced on the observed covariates included in the treatment propensity model when evaluating treatment effects,

treatment selection was not modeled effectively. Therefore, systematic differences between selected and non-selected students remain a plausible explanation for the findings.

Both independent and dependent measures used to estimate treatment propensity and treatment effects had cases with missing values. Although participants were balanced on patterns of missing values and the MI procedure that was used to impute missing data yields parameter estimates and standard errors with less bias than single imputation methods (Allison, 2002), missing data remains a possible problem. First, imputation is less reliable for variables with a high proportion of missing values. While 63% (n = 5246) of the student sample was missing values on one or more variables (M = 5.55, SD = 6.22), approximately 30% of participants were missing values for the teacher survey composites. Due to the high proportion of missing values for the teacher survey composites, including these variables when estimating treatment propensity may have introduced bias and contributed to the poor model fit. Moreover, it is possible that data were not MAR, as is assumed for MI, thereby introducing further bias.

Teachers may have extended their application of the knowledge and skills gained through consultation to improve instructional practices and address additional student concerns, thereby diffusing the effect of the program to non-selected students. If treatment diffusion occurred, the classroom may be a more appropriate unit of analysis than the student. With the classroom as the unit of analysis, it would be expected that classrooms whose teachers sought support of the IC Team would have higher average achievement, net of prior achievement, than classrooms whose teachers did not seek IC Team support. In fact, Silva (2007) did not find effects of attending an IC Team school

on students, but did find a significantly positive effect of being in an IC Team school on average classroom reading achievement.

Finally, the measures of achievement used in this study may not have been sufficiently sensitive to measure change, and prior achievement may not have been sufficiently controlled. The SOLs broadly measure reading and math achievement, but the teacher consultation may have focused on only one of several skills that comprise the SOL score. Furthermore, first quarter domain grades were used as a covariate control when evaluating treatment effects because only fourth and fifth grade students had SOL scores from the previous year. However, for the fourth and fifth grade students, first quarter grades and prior SOL scores were only moderately correlated.

Future Directions.

When random assignment of students to IC and Non-IC conditions is not possible or practical, the utility of applying propensity scores to reduce selection threats relies on effectively modeling the selection process. While it is possible that variables relating to selection were not measured in this study, the treatment propensity model only considered main effects and may not have been sufficiently complex to model the student-teacher dynamics that influenced the selection process. According to Rosenbaum and Rubin (1984), adding interaction or non-linear terms to the propensity model may improve model fit. Replicating the current study with a better specified treatment propensity model would improve the validity of the inferences about the effect of IC on student academic achievement. Furthermore, the pursuit of a better specified treatment propensity model is an appropriate avenue for research independent of an evaluation of treatment effects. A brief review of the literature over the past 10 years did not find any

studies that attempted to quantify the dynamic process of referring students to school intervention teams. Instead, most studies simply described the referred sample or focused exclusively on referral odds based on student demographic characteristics.

The problem of missing data is common in large-scale, school-based research. Imputing values for the teacher survey composites and including those measures when estimating treatment propensity may have introduced bias. Moreover, the pattern of missing data may not have been MAR, as had been assumed. Future research should evaluate the plausibility of these threats to the validity of the findings in this study. If treatment effects are consistent when treatment propensity is modeled with and without the teacher survey composites, then including variables with a high percentage of missing values was not a plausible limitation in this study. Furthermore, listwise deletion was not the chosen option for handling missing data because doing so may have substantially reduced effective sample size, and therefore, statistical power. However, listwise deletion is more robust to violations of the MAR assumption than the EM algorithm or MI (Allison, 2002), and comparing outcomes among listwise deletion, the EM algorithm, and MI data sets should be considered. If treatment propensity models and treatment effects are consistent across methods, then potential violations of the MAR assumption is less a less plausible limitation in this study.

Finally future research that makes use of the data set and methods in this study to evaluate the effect of IC on student academic achievement should consider alternative student samples. First, evaluating the effect of the program exclusively among the fourth and fifth grade students would allow prior SOL scores to be used as controls for prior achievement. Second, the potential problem of treatment diffusion could be evaluated by

replicating the study, but sampling the non-selected students from the 17 schools not implementing IC Teams. Evaluating the effect of IC on academic achievement during Year 2 Intervention (2007-08) and Year 3 Intervention (2008-09) may yield further information about the effect of treatment diffusion and levels of use.

Appendix A

Measures Included in the Imputation Model

Student			Teacher		
	Year			Year	
Measure	2005-06	2006-07	Measure	2005-06	2006-07
Demographic			Demographic		
Gender		x	Gender		x
Advantaged Ethnicity		x	Advantaged Ethnicity		x
Limited English Proficient		x	Age		x
Grade Level		x	TSR		
Old for Grade		x	Years Teaching		x
Young for Grade		x	Years at School		x
Services			Elementary Licensure		x
Free and Reduced Meals		x	Level of Education		x
Special Education	x	x	Efficacy	x	x
English as Second Language		x	Collaboration	x	x
IC Case		x	Job Satisfaction	x	x
Enrollment			Instructional Practices	x	x
New to District in 2006-07		x	TRSB		
Entered after 1st Quarter		x	Global Progress	x	x
Proportion Days Enrolled	x	x	Global Behavior	x	x
Proportion Days Absent	x	x	Concentration	x	x
Retained at End of Year	x	x	Externalizing	x	x
Achievement			Internalizing	x	x
1st Quarter Listening	x	x	Closeness	x	x
1st Quarter Math	x	x	Conflict	x	x
1st Quarter Reading	x	x			
1st Quarter Writing	x	x			
4th Quarter Listening	x	x			
4th Quarter Math	x	x			
4th Quarter Reading	x	x			
4th Quarter Writing	x	x			
Listening GPA	x	x			
Math GPA	x	x			
Reading GPA	x	x			
Writing GPA	x	x			
Overall GPA	x	x			
Math SOL	x	x			
Reading SOL	x	x			

Note. Imputed measures are highlighted. Non-highlighted measures either did not have missing values or were included in the imputation model as a highly correlated predictor. Only one participant had a missing value for Free and Reduced Meals.

References

Artiles, A., Klinger, J., & Tate, W. (2006). Representation of minority students in special education: Complicating traditional explanations. *Educational Researcher, 35*, 3-5.

Bradley-Johnson, S., & Dean, V. (2000). Role change for school psychology: The challenge continues in the new millennium. *Psychology in the Schools, 37*(1), 1-5.

Berger, J., Vaganek, M., Yiu, H., Nelson, D., Rosenfield, S., Gravois, T., et al. (2010). Exploratory study of teacher utilization of Instructional Consultation Teams. Unpublished manuscript. University of Maryland at College Park.

Bruckman, K., Vu, P., Vaganek, M., Berger, J., Rosenfield, S., & Gottfredson, G. (2010). The effects of Instructional Consultation Teams on student achievement and teacher ratings. Unpublished manuscript, University of Maryland at College Park.

Bryk, A., & Schneider, B. (2003). Trust in schools: A core resource for school reform. *Educational Leadership, 60,* 40-44.

Condron, D. (2008). An early start: Skill grouping and unequal reading gains in the elementary years. *The Sociological Quarterly, 49*, 363-394.

D'Agostino, R. & Rubin, D. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association, 95*, 749-759.

Ehrhardt-Padgett, G., Hatzichristou, C., Kitson, J., & Meyers, J. (2004). Awakening to a new dawn: Perspectives of the future of school psychology. *School Psychology Review, 33*(1), 105-114.

Erchul, W., & Sheridan, S. (2008). Overview: The state of scientific research in consultation. In W.P. Erchul & S.M. Sheridan (Eds.), *Handbook of research in consultation.* New York: Lawrence Erlbaum Associates.

Experimental Evaluation of Instructional Consultation Teams. (2010a). *Does level of use make a difference for teacher outcomes?* (Research Note 2). College Park, MD: University of Maryland, Department of Counseling and Personnel Services.

Experimental Evaluation of Instructional Consultation Teams. (2010b). *Intent-to-treat-school and intent-to-treat-teacher (maximum-opportunity for exposure) perspectives* (Research Note 1). College Park, MD: University of Maryland, Department of Counseling and Personnel Services.

Gravois, T., & Gickling, E. (2008). Best practices in curriculum-based assessment. In A. Thomas & J. Grimes (Eds.), Best practices in school psychology V (pp. 885-898). Bethesda, MD: National Association of School Psychologists.

Gravois, T., Nelson, D., & Sherry, E. (2007). Prince William County Instructional Consultation Team Consortium: 2006-07 End of Year Progress Report, Prince William County.

Gravois, T., & Rosenfield, S. (2002). A multi-dimensional framework for evaluation of instructional consultation teams. Journal of Applied School Psychology, 19(1), 5-29.

Gravois, T., & Rosenfield, S. (2006). Impact of instructional consultation teams on the disproportionate referral and placement of minority students in special education. Remedial and Special Education, 27(1), 42-52.

Gravois, T., Rosenfield, S., & Gickling, E. (1999). Instructional consultation teams: Training manual. College Park, MD: University of Maryland, Instructional Consultation Lab.

Gutkin, T., & Curtis, M. (1999). School-based consultation theory and practice: The art of science of indirect service delivery. In C. R. Reynolds & T. B. Gutkin (Eds.), *Handbook of school psychology* (3rd ed.). New York: John Wiley.

Hahs-Vaughn, D., & Onwuegbuzie, A. (2006). Estimating and using propensity score analysis with complex samples. *The Journal of Experimental Education, 75*, 31-65.

Hodges, K., & Grunwald, K. (2005). The use of propensity scores to evaluate outcomes for community clinics: Identification of an exceptional home-based program. *The Journal of Behavioral Health Sciences & Research, 32*(3), 294-305.

Hong, G., & Yu, B. (2008). Effects of Kindergarten retention on children's social-emotional development: An application of propensity score method to multivariate, multilevel data. *Developmental Psychology, 44*(2), 407-421.

Levinsohn, M. (2000). Evaluating instructional consultation teams for student reading achievement and special education outcomes. *Dissertation Abstracts International, 62* (01), 128A. (UMI No. 3001440)

Luellen, J. (2007). A comparison of propensity score estimation and adjustment methods on simulated data. *Dissertation Abstracts International, 68*(5-B), 3433. (UMI No. 3263706).

Luellen, J., Shadish, W., & Clark, M. (2005). Propensity scores: An introduction and experimental test. *Evaluation Review, 29*, 530-558.

Newman, D. (2007). *An investigation of the effect of instructional consultation teams on special education placement rate.* Unpublished master's thesis. University of Maryland at College Park.

O'Connor, C., & Fernandez, S. (2006). Race, class, and disproportionality: Reevaluating the relationship between poverty and special education placement. *Educational Researcher, 35*, 6-11.

Perkins, S., Tu, W., Underhill, M., Zhou, X., & Murray, M. (2000). The use of propensity scores in pharmacoepidemiologic research. *Pharmacoepidemiology and drug safety, 9*, 93-101.

Pianta, R. (2001). *STRS Student-Teacher Relationship Scale. Professional manual.* Odessa, FL: Psychological Assessment Resources.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Reid, K., & Knight, M. (2006). Disability justifies exclusion of minority students: A critical history grounded in disability studies. *Educational Researcher, 35*, 18-23.

Rosenbaum, R., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55.

Rosenbaum, R., & Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*, 516-524.

Rosenfield, S. (1995). Instructional consultation: A model for service delivery in the schools. *Journal of Educational and Psychological Consultation, 6*, 297-316.

Rosenfield, S. (2005). Best practices in instructional consultation. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp. 609-623). Bethesda, MD: National Association of School Psychologists.

Rosenfield, S., & Gottfredson, G. (2004). *Evaluating the Efficacy of Instructional Consultation Teams.* Unpublished grant proposal. University of Maryland, Department of Counseling and Personnel Services. Retrieved June 4, 2008 from http://www.icteams.umd.edu/Proposal%20for%20Project.pdf.

Rosenfield, S., & Gravois, T. (1996). *Instructional consultation teams: Collaborating for change*. New York: Guilford.

Rosenfield, S., & Gravois, T. (1999). Working with teams in the school. In C. R. Reynolds & T. B. Gutkin (Eds.), *Handbook of school psychology* (3rd ed., pp. 1025-1040). New York: John Wiley.

Rosenfield, S., Silva, A., & Gravois, T. (2008). Bringing instructional consultation to scale: Research and development of IC and IC Teams. In W.P. Erchul & S.M. Sheridan (Eds.), *Handbook of research in consultation.* New York: Lawrence Erlbaum Associates.

Rubin, D., & Thomas, N. (1996). Matching using propensity scores: Relating theory to practice. *Biometrics, 52,* 249-264.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* New York: Houghton Mifflin Company.

Schafer, J., & Graham, J. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177.

Sheridan, S., Welch, M., & Orme, S. (1996). Is consultation effective? A review of outcome research. *Remedial and Special Education, 17*(6), 341-354.

Silva, A. (2007). *A quasi-experimental evaluation of reading and special education outcomes for English language learners in instructional consultation schools.* Unpublished doctoral dissertation, University of Maryland at College Park.

Tschannen-Moran, M., & Hoy, A. (2001). Teacher efficacy: capturing an elusive construct. *Teaching and Teacher Education, 17*, 783-805.

Vanderweele, T. (2006). The use of propensity score methods in psychiatric research. *International Journal of Methods in Psychiatric Research, 15*(2), 95-103.

Virginia Department of Education. (2005). Virginia Standards of Learning Assessments Technical Report: 2003-2004 Administration. Retrieved December 13, 2009, from http://www.doe.virginia.gov/VDOE/Assessment/home.shtml.

Vu, P., Bruckman, K., Koehler, J., Kaiser, L., Rosenfield, S., Nelson, D., et al. (2009). The effect of Instructional Consultation Teams on teacher beliefs and instructional practices. Unpublished manuscript, University of Maryland at College Park.

Werthamer-Larson, L., Kellam, S., & Wheeler, L. (1991). Effect of first-grade classroom environment on shy behavior, aggressive behavior, and concentration problems. *American Journal of Community Psychology, 19,* 585-602.

Wu, W., West, S., & Hughes, J. (2008). Effect of retention in first grade on children's achievement trajectories over 4 years: A piecewise growth analysis using propensity score matching. *Journal of Educational Psychology, 100*(4), 727-740.

Ye, Y., & Kaskutas, L. (2009). Using propensity scores to adjust for selection bias when assessing the effectiveness of Alcoholics Anonymous in observational studies. *Drug and Alcohol Dependence, 104*, 56-64.

Ysseldyke, J., Burns, M., Dawson, P., Kelley, B., Morrison, D., Ortiz, S., et al. (2006). *School psychology: A blueprint for training and practice III*. Bethesda, MD: National Association of School Psychologists.