Abstract Title Page
Not included in page count.

Title: Challenges when using the Regression Discontinuity Design in educational evaluations: Lessons from the Transition to Algebra study.

Authors and Affiliations:

Josephine Louie, Education Development Center, Inc.
Christopher Rhoads, University of Connecticut
June Mark, Education Development Center, Inc.

Abstract Body

Limit 4 pages single-spaced.

Problem / Background / Context:

Description of the problem addressed, prior research, and its intellectual context.

Educational evaluations typically involve a partnership between outside evaluation teams and the districts, schools, administrators, teachers, parents and students that are the subject of the evaluation. The different needs of stakeholders that are party to an educational evaluation often lead to compromises Currently, such compromises occur in a context in which recent federal legislation, such as the *No Child Left Behind Act* of 2001 and the *Education Sciences Reform Act (ESRA)* of 2002 has insisted that educational evaluations use rigorous research designs with quantitative outcome measures. In particular, the Institute of Education Sciences (IES) at the U.S. Department of Education has promoted the use of *randomized controlled trials* (RCTs) as the preferred method for educational evaluations that seek to determine the causal effect of interventions (Whitehurst, 2004; Angrist, 2004).

As noted by Cook (2002), school districts and their constituents are often unwilling to participate in randomized experiments. One of the objections is that random assignment of a potentially beneficial educational intervention is inconsistent with the desire to target interventions toward students who are most in need of (and presumably, would most benefit from) the intervention. An alternative research design that can overcome this objection and may be more palatable to practitioners is the regression discontinuity (RD) design. RD designs involve the assignment of treatment to individuals on the basis of a measured value on a so-called "score" variable. Individuals on one side of the "cut score" are assigned to the treatment group and individuals on the other side are assigned to the control group. Measured values on the score variable are often a proxy for the level of need for the intervention. For example, students may be assigned to a summer school program if they achieve below a cutoff score on a standardized assessment.

Additionally, causal inferences from RD designs can be just as credible as inferences from randomized experiments (Lee, 2008). This is because there is often random measurement error contained in the observed values of the score variable. In such cases RD designs can be viewed as "locally randomized" designs around the cutoff score. Measurement error in the score variable implies that individuals with true scores near the cut point are effectively randomized to the treatment and control conditions. Despite the appeal of the RD design as an alternative to RCTs, the literature contains a few warnings for the evaluator hoping that RD will be a magic bullet. First, a RD design will require a larger sample size to produce a treatment effect estimate with the same precision as a RCT. The exact sample size requirements will depend on the distribution of the score variable and the location of the cut point along that distribution. Existing investigations suggest that the sample size in a RD design must be between 2 to 5 times as large as the sample size in a comparable RCT (Schochet, 2009; Cappelleri, Darlington and Trochim, 1994). Second, effect size estimates from RD designs can be highly sensitive to model specification. Evaluators must choose between a non-parametric and a parametric approach to the analysis. Non-parametric analyses can be subject to "boundary bias" and/or can be sensitive to bandwidth choice. Parametric analyses require the correct specification of a functional form relating the score variable to the outcome (Bloom, 2012). Finally, the ability of a RD design to produce a valid estimate of an average treatment effect can depend critically on the distribution of the score variable. As noted above, this distribution will help determine the exact sample size requirements of the design. Additionally, discreteness in the score variable precludes a fully non-

parametric analysis and hinders the ability to determine the functional form of the relationship between the score variable and the outcome (Lee and Card, 2008). Lack of sufficient variation in the score variable can also decrease the precision of treatment effect estimates in parametric analyses and can make it difficult to determine the appropriate functional form.

Purpose / Objective / Research Question / Focus of Research:

Description of the focus of the research.

This paper describes the experience of our research team when conducting a RD study to evaluate the impact of a supplemental algebra-readiness curriculum. We illustrate how compromises in the evaluation design and the difficulties mentioned above prevented strong conclusions about the efficacy of the intervention. In particular, restrictions in the sample available for our evaluation and difficulty convincing one of our school district partners to comply with cut-off based assignment of the intervention group led to a smaller sample size than originally anticipated. Discreteness and lack of variation in the score variable made it difficult to determine the correct functional form for our analysis, and estimates of effect size were highly sensitive to different specifications of the functional form. The research team's experience offers lessons about the implementation of RD studies and the need to work closely with practitioner partners to carry out evaluation designs that address both stakeholder concerns and broader goals of generating rigorous evidence to learn what works in education.

Improvement Initiative / Intervention / Program / Practice:

Description of the improvement initiative or related intervention, program, or practice.

The curriculum under study, Transition to Algebra (TTA), targets students who have been identified as underprepared for algebra and was designed primarily for full-year use as a supplemental curriculum in a Grade 9 double-period algebra context. TTA aims to foster algebraic habits of mind in order to help students use reasoning to simplify and make sense of algebraic work. Examples of these habits include: puzzling and persevering to solve problems; identifying and using numerical patterns and structure to approach algebraic tasks; and communicating mathematical ideas clearly and precisely through words, text, symbols, or other representations (Cuoco, Goldenberg, & Mark, 1996, 2010).

Setting:

Description of the research location and partners involved, if applicable.

The TTA curriculum was developed over a period of two years and then evaluated through a formal field test in two Massachusetts high schools in 2011-2012. Both schools, which we will call School A and School B, were located in ethnically diverse cities, with Grade 9-12 populations of over 3,000 and just under 2,000 students, respectively, during the noted academic year. Over 60% of the students in each school were non-white, and approximately 40% or more did not speak English as their first language. Almost 60% of students in both schools qualified for free or reduced-price lunch, compared to a statewide average of 35%.

Both School A and School B offered a supplemental algebra intervention class for Grade 9 students who had fallen in the lowest performance category on the Grade 8 Massachusetts Comprehensive Assessment System (MCAS8) state mathematics exam. Scaled scores on the MCAS8 (which include even numbers only) range from 200 to 280. These scores were used to determine assignment to the intervention and control conditions (i.e., MCAS8 was the "score" variable for our RD study).

After working collaboratively with administrators and teachers at both schools to develop the TTA curriculum materials, the TTA study team asked administrators to assign students to the TTA intervention using a single cutoff score on MCAS8 during the field test year. Due to its approach to course scheduling and a desire to honor students' elective requests, administrators at

School B did not agree to this request. School A did agree, and a cut score of 222 on MCAS8 was set to ensure that: (a) the intervention was provided to students most in need of supplemental algebra instruction and (b) all seats in intervention classrooms were filled.

Population / Participants / Subjects:

Description of the participants in the research: who, how many, key features, or characteristics.

Out of a list of 584 potential Grade 9 students at School A, school administrators asked the study team to make assignments to TTA from a list of only 183 students who had been assigned to particular administrative clusters within the school. Because there were more eligible students than seats available in intervention classrooms, 12 students with MCAS8 scores below 222 were randomly excluded from the TTA intervention group. Not all students from the original list ultimately enrolled in School A, and there was some attrition over the course of the school year. Thus, out of the 183 students on our original list, we were able to collect outcome data from 85 students who had been properly assigned to the TTA intervention (our "core" treatment group); 9 students who were eligible for the intervention but had been randomly excluded (our "core RCT" comparison group) and 27 students who were ineligible for the TTA intervention due to high MCAS8 scores (our "core RD" comparison group). Additionally, the school provided outcome and MCAS8 data for 38 students who were not on the original list and did not receive the TTA intervention. Fifteen of these 38 students had MCAS8 scores that made them eligible for TTA (the "supplemental eligible" group) and 23 had scores that made them ineligible (the "supplemental ineligible" group).

Research Design:

Description of the research design.

The main research design is a RD design. However, because 9 students were randomly excluded from the TTA intervention, there is a small RCT embedded within the basic RD design.

Data Collection and Analysis:

Description of the methods for collecting and analyzing data or use of existing databases.

Outcomes were measured using a modified version of the MCAS test, which we label MCASmod. Figure 1 shows a plot of MCASmod scores vs. MCAS8 scores, separated by the five data groups described above. Separate linear regression lines fit to the five different data groups are displayed. Table 1 shows means, standard deviations and the correlations between MCASmod and MCAS8 for these same 5 groups.

Due to space constraints we only describe analyses for the core data groups. Given the sample size obtained and the discreteness in our score variable, we opted for a parametric analysis of the data. Following recommendations in Jacob, et al. (2012) we began with a "smoothed" plot of MCASmod vs. MCAS8 for the core RD data using binned data. An example for a bin width of 4 is provided as Figure 2. Plots for other bin widths were very similar.

Following a recommendation in Lee and Lemieux (2010) we tested for the goodness of fit of the basic (non-interacted) linear model and found that the corresponding F test could not reject that the linear model was an adequate fit (Table 2). Given the visual evidence to the contrary in Figure 1 we are inclined to view this result as indicating lack of statistical power in the F test.

Because initial explorations gave conflicting indications about the functional form of the MCASmod-MCAS8 relationship, we proceeded to fit models with a variety of functional forms to investigate the sensitivity of our results to different functional form assumptions. Results for just the core RD data are presented in Table 3. Table 4 presents corresponding results utilizing all of the core data. Models were fit including fixed effects of teachers, however, these coefficients are suppressed for ease of presentation.

Findings / Outcomes:

Description of the main findings or outcomes, with specific details.

Figure 1 shows interesting differences between the five groups in our study. The regression slope for the core RD comparison group is almost flat, whereas the slopes for the treatment and RCT core groups show a strong positive association. Additionally, the supplemental groups have different characteristics than the corresponding core groups. We interpret results for core data groups only.

Tables 3 and 4 illustrate our dilemma in making conclusions from our data. Due perhaps to our limited sample size, the coefficient associated with TTA is not statistically significant in any of the models presented. Effect size estimates are highly sensitive to the specified functional form of the model, ranging from -0.131 to 0.829 in Table 3 and from -.002 to 0.219 in Table 3. Based on the Akaike Information Criterion (AIC) we would choose the linear model with a treatment interaction if only the RD data were available. This would lead to an effect size estimate of -0.131. However, when the RCT data are included (Table 4) the AIC suggests a quadratic model and the effect size estimate is effectively zero.

The 9 RCT cases are extremely important in the present case. Had they not been available both a visual inspection of Figure 1 and the AIC values would have led us to choose a linear interaction model. The difference in the slopes of the regression line above and below the cut point would most likely have been interpreted as a treatment effect. The availability of the RCT cases appears to show that the difference in slopes is not an effect of treatment, but rather a natural, pre-treatment feature of the data distribution. Hence, while it is difficult to say anything with assurance given the noise in the data set and the sensitivity of estimates to functional form specifications, we are inclined to take -.002 as the best available estimate of the treatment effect.

Conclusions:

Description of conclusions, recommendations, and limitations, based on findings.

Our experience with the TTA RD evaluation shows the care evaluators must take when opting to use a RD evaluation design. We opted for a RD design believing that it would increase the willingness of our school district partners to participate in our evaluation. However, only one of our two school district partners was willing to abide by a cutoff-based assignment mechanism. The final sample size was far too small to estimate the average treatment effect with a desirable level of precision. We calculated the minimum detectable effect size (MDES) of our study using the sample size and correlation values obtained in our core data. The MDES is 0.65. This figure is far too high given recommendations that educational evaluations be designed to detect effect sizes as low as 0.25 (Bloom, Hill, Rebeck Black and Lipsey, 2008; Schochet, 2008).

Had we been lucky, we might have obtained effect size estimates that were consistent across a variety of different model specifications. We were not so fortunate. Under these conditions it became crucial to choose a single functional form whose estimated treatment effect was the most trustworthy. Unfortunately, the usual methods for making this determination were not terribly fruitful. Binned plots revealed a non-monotonic relationship that does not seem plausible. F tests of the basic linear (non-interacted) specification also gave results that seemed to contradict the visual evidence. The discreteness and lack of variation in our score variable almost certainly impaired our ability to have confidence in a particular model specification.

Based on our experience, we recommend that other evaluators take care to make sure that the necessary conditions for a successful RD are in place before undertaking such an evaluation. Without the requisite buy-in from school district partners, a large sample size, and a score variable with desirable properties, it is unlikely that the evaluation will reach strong conclusions.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Angrist, J. (2004). American education research changes tack. *Oxford Review of Economic Policy*, *20*(2), 198–212.

Bloom, H. S. (2012). Modern regression discontinuity analysis. *Journal of Research on Educational Effectiveness*, *5*(1), 43–82.

Bloom, H. S., Hill, C., Rebeck Black, A., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*, 289–328.

Cappelleri, J. C., Darlington, R. B., & Trochim, W. M. K. (1994). Power analysis of cutoff-based randomized clinical trials. *Evaluation Review*, *18*(2), 141–152.

Cook, T. D. (2002). Randomized Experiments in Educational Policy Research: A Critical Examination of the Reasons the Educational Evaluation Community has Offered for not Doing Them. *Educational Evaluation and Policy Analysis*, *24*(3), 175–199. doi:10.3102/01623737024003175

Cuoco, A., Goldenberg, E. P., & Mark, J. (1996). Habits of mind: An organizing principle for a mathematics curriculum. *Journal of Mathematical Behavior*, *15*(4), 375–402.

Cuoco, A., Goldenberg, E. P., & Mark, J. (2010). Organizing a curriculum around mathematical habits of mind. *Mathematics Teacher*, *103*(9), 682–688.

Lee, D. S. (2008). Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics*, *142*(2), 675–697. doi:10.1016/j.jeconom.2007.05.004

Lee, D. S., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, *142*(2), 655–674. doi:10.1016/j.jeconom.2007.05.003

Schochet, P. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, *33*, 62–87.

Schochet, P. (2009). Statistical Power for Regression Discontinuity Designs in Education Evaluations. *Journal of Educational and Behavioral Statistics*, *34*(2), 238–266. doi:10.3102/1076998609332748

Whitehurst, G. J. (2004, April 26). Making education evidence-based: Premises, principles, pragmatics, and politics (IPR Distinguished Public Policy Lecture Series, 2003-2004). Northwestern University, Institute for Policy Research.

Appendix B. Tables and Figures

Not included in page count

Tables

Table 1. Basic descriptive statistics for RD score and outcome variables, by sample subgroup

	MCAS8					MCASmod				r	
	N	Mean	SD	Min	Max	Mean	SD	Min	Max		
Treatment	85	217.1	3.4	206	222	7.5	3.8	0	19	0.364	**
RD core	27	230.1	4.8	224	238	9.7	3.7	4	16	0.015	
Ineligible supp.	23	239.8	6.5	226	250	12.3	3.7	3	18	0.078	
RCT core	9	215.6	5.1	204	220	6.7	3.2	2	11	0.636	
Eligible supp.	15	217.7	3.6	208	222	5.7	3.3	0	10	0.065	

Notes:
The potential score range for the statewide Grade 8 MCAS exam in mathematics (MCAS8) was 200 to 280, even numbers only. The potential score range for the modified MCAS exam (MCASmod) was 0 to 20.

Treatment cases had scores at or below the MCAS8 cutoff of 222 and received the TTA intervention during the study year. RD core cases had scores above the cutoff and did not receive TTA. Ineligible supplemental cases were not part of the study's original enumeration list, had scores above the cutoff, and (properly) did not receive TTA. RCT core cases had scores at or below the cutoff and were randomly excluded from TTA. Eligible supplemental cases were not part of the study's original enumeration list, had scores at or below the cutoff and were therefore eligible for TTA, but did not receive the intervention.

**p<.01.

Table 2. Test for Assessing Fit of Basic Linear Regression Discontinuity Model

Bin size	Restricted R^2 [a] (R^2_r)	Unrestricted R^2 [b] (R^2_u)	# of bins (K)	n	F[c]	p[d]
8	0.123	0.153	5	112	0.751	0.587
6	0.123	0.145	7	112	0.382	0.911

N = 112.

a The restricted R^2 (R^2_r)comes from a model without bin indicators.

b The unrestricted R^2 (R^2_u)comes from a model with bin indicators (bin size = 8 in top row; bin size = 6 in bottom row).

c The F statistic is calculated as follows: $$F\ statistic = \frac{(R_u^2 - R_r^2)/K}{(1 - R_u^2)/(n - K - 1)}$$

d The p-value was obtained using degrees of freedom K and n-K-1 for the numerator and denominator, respectively.

Table 3. Results for RD core data. Various model specifications.
(standard errors in parentheses)

Variable[c]	(I) Linear		(II) Linear interaction		(III) Quadratic		(IV) Quadratic interaction	
Intercept	6.009	***	7.769	***	7.234	***	4.760	~
	(1.25)		(1.62)		(1.45)		(2.46)	
TTA Status	.593		-.497		-.209		3.138	
	(1.48)		(1.60)		(1.55)		(2.62)	
Grade 8 MCAS c222	.243	**	.039		.222	*	1.164	*
	(0.09)		(0.15)		(0.09)		(0.73)	
TTA * (Grade 8 MCAS c222)			.322				-0.571	
			(0.19)	~			(0.80)	
(Grade 8 MCAS c222)^2					-.010		-.069	
					(0.01)		(0.04)	
TTA * (Grade 8 MCAS c222)^2							.086	~
							(0.05)	
N	112		112		112		112	
R^2	0.23		0.25		0.25		0.27	
AIC	292.60		291.48		291.72		291.98	
SD_y	3.9		3.9		3.9		3.9	
ES	**0.157**		**-0.131**		**-0.055**		**0.829**	
ES upper bound	0.161		-0.127		-0.051		0.834	
ES lower bound	0.152		-0.136		-0.060		0.824	

[c] Variables defined as:

TTA Status - a dichotomous variable indicating whether a student was enrolled in TTA (1=yes, 0=no).

Grade 8 MCAS c222 - Student's Grade 8 MCAS score, centered at 222.

N - Total sample size included in model.

R^2 - Proportion of variation in Modified MCAS scores explained by all the predictor variables in the model.

SD_y - Standard deviation of Modified MCAS scores.

AIC - Akaike Information Criterion.

ES - Effect size, measured as Cohen's d, or TTA Status parameter estimate / pooled SD_y

ES upper and lower bound - Effect size estimates at the upper and lower bounds of a 95% confidence interval.

~p<0.10; *p<0.05; **p<0.01; ***p<0.001.

Table 4. Results for RD and RCT core data. Various model specifications.
(standard errors in parentheses)

Variable[c]	(1) Linear		(II) Linear interaction		(III) Quadratic		(IV) Quadratic interaction	
Intercept	5.948	***	6.273	***	6.806	***	6.816	***
	(0.95)		(0.98)		(1.10)		(1.11)	
TTA Status	.487		0.832		-.009		.770	
	(0.92)		(0.96)		(0.97)		(1.24)	
Grade 8 MCAS c222	.240	*	.185	*	.227	***	.220	**
	(0.06)		(0.08)		(0.06)		(0.08)	
TTA * (Grade 8 MCAS c222)			.174				0.391	
			(0.14)				0.322	
(Grade 8 MCAS c222)^2					-.008		-.009	
					(0.01)		(0.01)	
TTA * (Grade 8 MCAS c222)^2							.029	
N	121		121		121		121	
R^2	0.26		0.27		0.27		0.28	
AIC	307.41		307.62		306.87		309.11	
SD_y	3.8		3.8		3.8		3.8	
ES	**0.128**		**0.219**		**-0.002**		**0.202**	
ES upper bound	0.131		0.222		0.001		0.205	
ES lower bound	0.125		0.216		-0.005		0.199	

[c] Variables defined as

TTA Status - a dichotomous variable indicating whether a student was enrolled in TTA (1=yes, 0=no).

Grade 8 MCAS c222 - Student's Grade 8 MCAS score, centered at 222.

N - Total sample size .

R^2 - Proportion of variation in Modified MCAS scores explained by all the predictor variables in the model.

SD_y - Standard deviation of Modified MCAS scores.

AIC - Akaike Information Criterion.

ES - Effect size, measured as Cohen's d, or TTA Status parameter estimate / pooled SD_y

ES upper and lower bound - Effect size estimates at the upper and lower bounds of a 95% confidence interval.

~p<0.10; *p<0.05; **p<0.01; ***p<0.001.

Figures

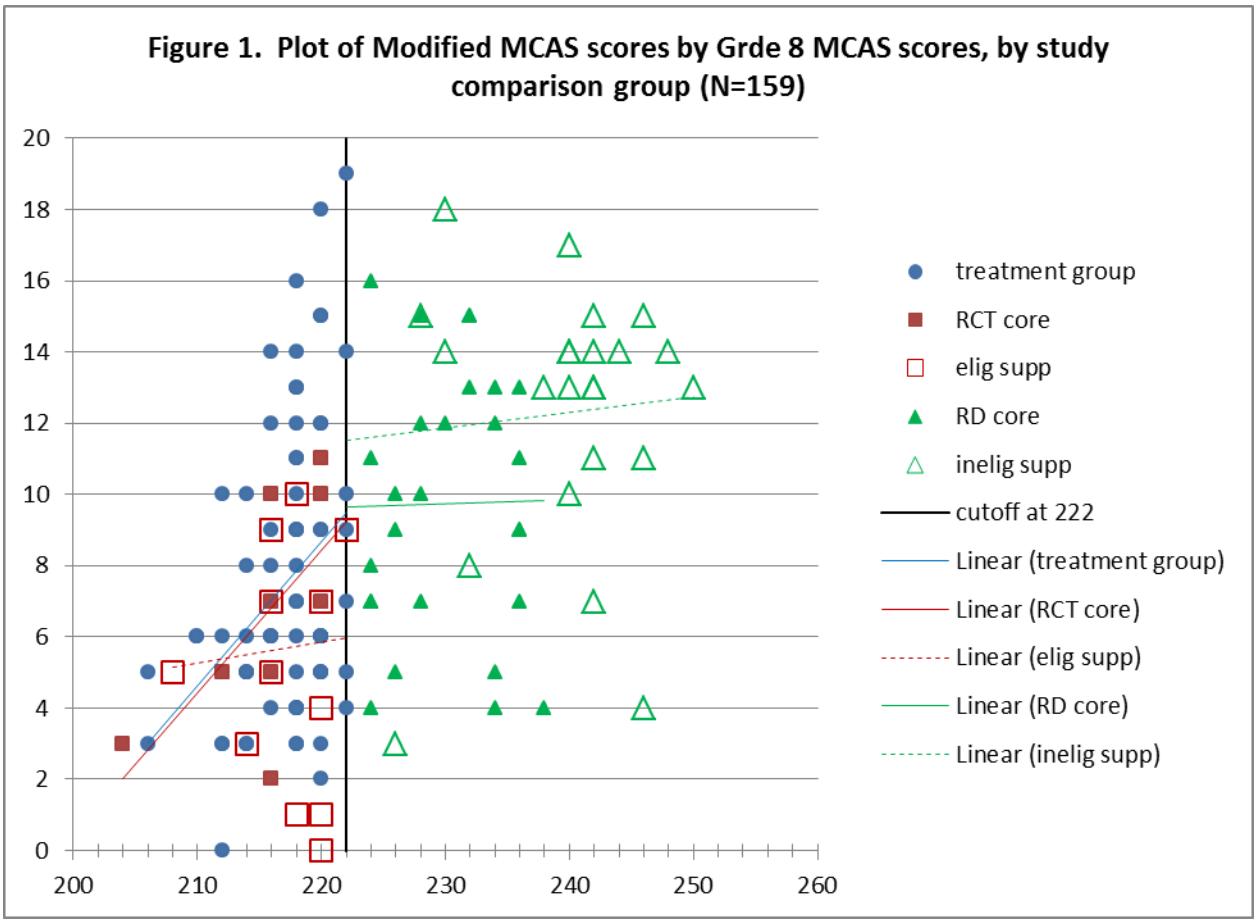
Figure 1. Plot of Modified MCAS scores by Grde 8 MCAS scores, by study comparison group (N=159)
20
18
16
14
12
10
8
6
4
2
0
200
210
220
230
240
250
260
treatment group
RCT core
elig supp
RD core
inelig supp
cutoff at 222
Linear (treatment group)
Linear (RCT core)
Linear (elig supp)
Linear (RD core)
Linear (inelig supp)

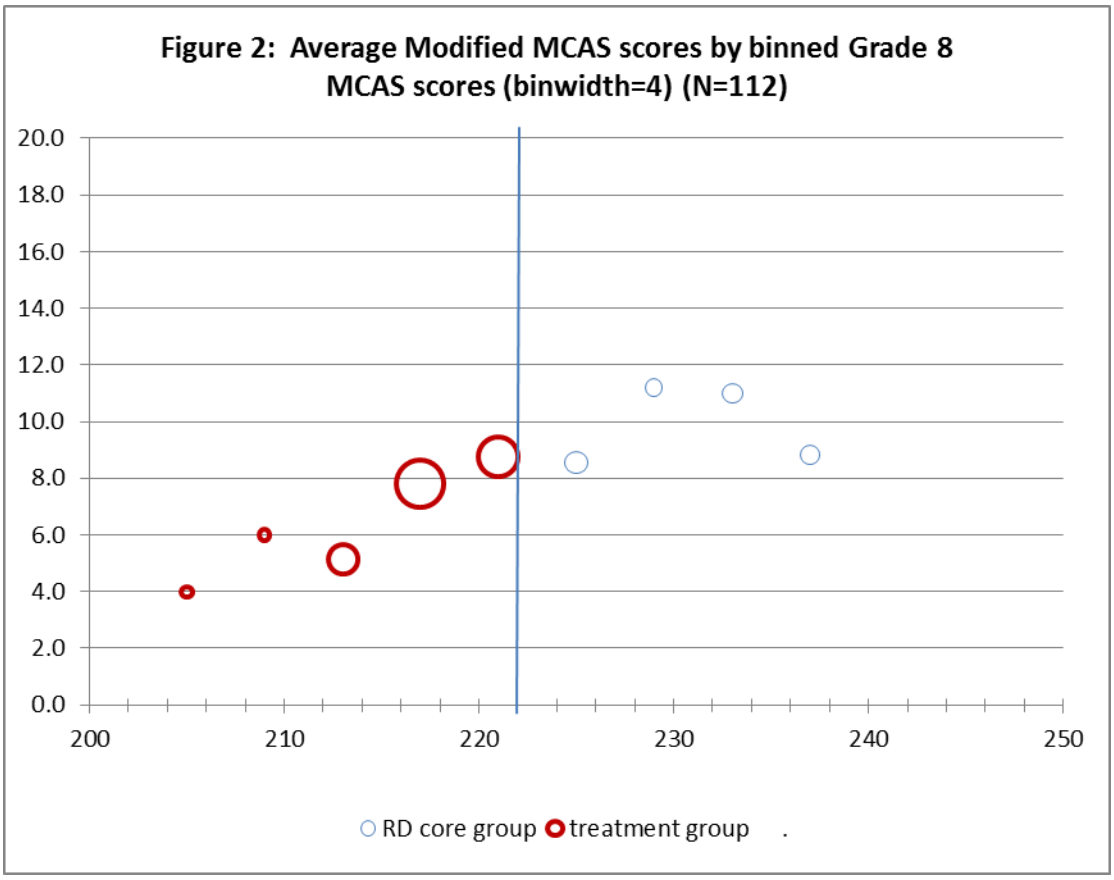
Figure 2: Average Modified MCAS scores by binned Grade 8 MCAS scores (binwidth=4) (N=112)
20.0
18.0
16.0
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
200
210
220
230
240
250
RD core group
treatment group