Abstract Title Page

Title:
Reducing bias and increasing precision by adding either a pretest measure of the study outcome or a nonequivalent comparison group to the basic regression discontinuity design: An example from Education

Authors and Affiliations:
Yang Tang,
Postdoctoral fellow, Institute for Policy Research, Northwestern University

Thomas D. Cook,
Professor of Sociology, Psychology, Education, and Social policy, Institute for Policy Research, Northwestern University;
Senior Fellow, Mathematica Policy Research

Yasemin Kisbu-Sakarya,
Assistant Professor, Department of Psychology, Koc University

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Regression discontinuity design (RD) has been widely used to produce reliable causal estimates. Researchers have validated the accuracy of RD design using within study comparisons (Cook, Shadish & Wong, 2008; Cook & Steiner, 2010; Shadish et al, 2011). Within study comparisons examines the validity of a quasi-experiment by comparing its estimates to trustworthy benchmarks (usually an experiment) with the same treatment group. First developed by Lalonde (1986), it is a rigorous method to evaluate the internal validity of quasi-experiments.

Despite the accuracy of RD design and its increasing popularity, the basic RD design has three major disadvantages. The first disadvantage is that RD has weak statistical power. Previous literature (Goldberger, 1972; Schochet, 2009) shows that RD needs at least 2.75 times of RCT's sample size to achieve the same statistical power. The second disadvantage is that RD can only estimate causal effect at the cutoff, but most times researchers are concerned with average treatment effect across all populations. The third disadvantage is that RD depends on specification of functional forms. Researchers usually specify several different functional forms to test the sensitivity of estimates.

To mitigate these three weaknesses, researchers have proposed many variants of RD (Lohr, 1972; Shadish, Cook and Campbell, 2002). Wing and Cook (2013) employed within study comparisons in the healthcare context and examined the advantages of adding a pretest to a basic RD, thus creating comparative regression discontinuity with the pretest as a comparison function (CRD-Pre). Tang and Cook (under review) derived theoretically the variance formula for CRD-Pre and illustrated how to use this formula for power analysis in designing CRD-Pre. They also discussed the power of comparative regression discontinuity design with a nonequivalent comparison group as the comparison function (CRD-CG). Theoretically, if the assumptions for CRD designs can be met, both CRD-Pre and CRD-CG have greater power than the basic RD and can be as powerful as an experiment.

Purpose / Objective / Research Question / Focus of Study:

By supplementing a basic RD design with a pretest function, we create CRD-Pre; and by adding a non-equivalent comparison group to a basic RD design we create CRD-CG. The purpose of this paper is to test whether these two CRD designs can mitigate the three weaknesses of a basic RD design so as to give more support to functional form estimation, to generalize treatment effects away from the cutoff, and to increase power. This paper raises three questions about CRD-Pre and CRD-CG designs: first, to what extent do they reduce bias and increase precision compared to basic RD? RD is known for producing unbiased estimates at the cutoff that do not generally differ from those of an RCT at the cutoff. We test whether adding comparison functions – whether pretests or a non-equivalent comparison group –still produces valid causal estimates at the cutoff and to what extent the results are more precise than with basic RD.

Second, we ask which is better for reducing bias and increasing precision - CRD-Pre or CRD-CG? The decision to use CRD-Pre or CRD-CG should depend on data availability – whether pretests of the same individuals are at hand or independent comparison cases. But we also want

to examine whether one has clear advantages over the other as well as to discuss the feasibility of each design variant in actual research practice.

Third, to what extent are CRD-Pre and CRD-CG as good as the RCT in producing unbiased and precise estimates, not just at the cutoff, but also – and particularly – away from it so as to reduce the generalization disadvantage of basic RD? RCT produces average treatment effect applicable to the whole population, and no CRD design can be quite as general. But the potential for greater causal generalization than with RD is real. (And sometimes in RCTs researchers test for treatment effects limited to sub-populations, as is the case with CRD). So we compare CRD designs with RCT for accuracy and precision and check if CRD designs can be as accurate and powerful as an RCT above the RD cutoff on the assignment variable.

This paper uses a within study comparison to examine the validity of CRD designs relative to RD design and RCT. We use data from the 3-year-olds in the Head Start Impact Study and estimate RCT causal effects as the benchmarks. We then construct synthetic RD design from the experiment data. We create CRD designs by adding pretest data of the 3-year-olds as the pretest function. For the non-equivalent comparison group we compare 3-years olds in Head Start with 4 year olds not in it. Then we test the performance of CRD designs using three outcome variables in mathematics, literacy and socio-emotional development.

Significance / Novelty of study:
Although a well-implemented basic RD design can produce reliable and unbiased causal estimates, researchers have to tolerate its three weaknesses and their consequences. They are forced to explore many different functional forms or bandwidths, to spend more money and energy to increase sample size, and after all this, they can still only estimate treatment effects for the narrow range of population at the RD cutoff. This paper uses within study comparison methodology to validate whether CRD designs can mitigate all three weaknesses and to estimate the degree to which they do so in Education. This paper demonstrates that (a) CRD design can mitigate all three weaknesses of the basic RD design and can produce accurate and precise estimates, not only at the RD cutoff, but also above it; and that (b) the construction of CRD design is straightforward and easy to apply in practice. The bottom line of the paper is that education researchers should use CRD designs in place of RD in the many real-world contexts where data can be added from a pretest or non-equivalent comparison group.

Statistical, Measurement, or Econometric Model:
This paper conducts six within study comparisons using three outcome variables and two assignment variables. For each within study comparison, we first estimate the RCT's average treatment effect with and without covariates. Next we estimate the RCT's local effects both at and above the RD cutoff. These local RCT effects serve as the benchmarks against which to evaluate CRD effects at these same points. Third, we create a synthetic RD design by removing the treated cases below the RD cutoff and the untreated cases above the cutoff. We then plot the data using local linear methods and observe the functional forms. We estimate the RD causal effect at the cutoff using linear regression method in linear, quadratic and cubic orders. Fourth, we add either pretest data to create CRD-Pre or we add a non-equivalent comparison group to create CRD-CG. We also plot these data using local linear methods and examine whether the three untreated segments are parallel. We then use linear regression to estimate the CRD causal

effect at and above the RD cutoff. For each design, we use bootstrap methods to estimate standard errors of the estimates. Finally, we compare the estimates of RD and CRD designs and their standard errors to each other and also to the RCT benchmarks.

Usefulness / Applicability of Method:
Using data from the national Head Start Impact Study, this study provides empirical evidence of the accuracy and precision of CRD designs. It tests six within study comparisons formed by combining three outcomes (mathematics, literacy and socio-emotional development) over two assignment variables – the date of data collection and synthetic pretest scores. Each within study comparison compares a CRD design (either CRD-Pre or CRD-CG) to a basic RD design and to an RCT.

Data Collection and Analysis:
Head Start Impact Study. Mandated by Congress, this study covered 23 states, 84 randomly-selected agencies and 383 randomly-selected Head Start centers and a total of 4,442 children. It randomly assigned newly entering 3- and 4-year old Head Start applications to either treatment (Head Start) or control status (not enrolled in Head Start but able to receive services from other sources). Children were tested before the treatment in 2002 and again every spring from 2003 to 2006. We use the 2003 outcome data of the 3-year-olds for the main analysis, the 2002 pretest data of the same 3-year-olds to create CRD-Pre and the 2003 outcome data of the 4-year-old nonequivalent comparison group to create CRD-CG.

Findings / Results:
CRD-Pre For *estimates at the cutoff,* Figure 1 plots biases of CRD-Pre and RD estimates relative to each outcome's and the RCT benchmark. The X-axis labels each outcome, and the smaller the bias is from the benchmark the more accurate the estimate is. The biases of CRD-Pre estimates for the three outcomes are small relative to the RCT, ranging from 0.01 to 0.05. The corresponding biases in the basic RD biases range from 0.01 to 0.13. It seems, then, that CRD-Pre slightly improves on RD but both do well in the main task to which RD is traditionally addressed – generating unbiased causal inferences at the cutoff.

For *estimates above the cutoff*. Figure 2 plots the CRD-Pre biases from the RCT benchmarks above the cutoff. The biases of CRD-Pre are small and range from 0.02 to 0.07 (all below 0.10). Thus, CRD-Pre provides unbiased estimates above the cutoff and over a broader range than basic RD. (Basic RD causal effects away from the cutoff were not estimated since they are generally considered to be untrustworthy).
(please insert figures 1-2 here)

Figure 3 plots two ratios: (a) The ratio of the CRD-Pre standard errors to the adjusted s.e. of the RCT *at the cutoff* (highlighted in red); (b) the ratio of the RD standard errors to the same adjusted s.e. of the RCT is highlighted in blue. The smaller these ratios are, the more precise is the estimate and the more powerful the design. The ratio values of CRD-Pre to RCT range from 1.15 to 1.52; those for RD are larger and range from 1.40 to 1.94. It is obvious to the eye that standard errors of CRD-Pre estimates are much smaller at the cutoff.

Figure 3 also plots the ratios of CRD-Pre standard errors *above the cutoff* to those of the RCT benchmark above the cutoff (in green). The ratios range from 0.89 to 1.02, indicating that CRD-Pre is at least as powerful as RCT above the cutoff.
(please insert figures 3 here)

CRD-CG. For estimates *at the cutoff*, Figure 4 plots the biases for CRD-CG and RD estimates compared to the RCT benchmarks. The biases of CRD-CG for the three outcomes are small and range from 0.02 to 0.08, compared to 0.02 to 0.05 in the basic RD. All the biases for CRD-CG and RD are below 0.10. Therefore, CRD-CG produces unbiased estimates at the cutoff, as does the basic RD.

For *estimates above the cutoff*, Figure 5 plots the CRD-CG biases relative to the RCT benchmarks. The biases of CRD-CG are small and range from 0.01 to 0.04 (all below 0.10). Thus, CRD-CG provides unbiased estimates above the cutoff while RD cannot generalize causal effects away from the cutoff.
(please insert figures 4-5 here)

Figure 6 plots the ratio of CRD-CG standard errors to the adjusted s.e. of the RCT at the cutoff (highlighted in red) and also the ratio of RD standard errors to the same adjusted s.e. of the RCT (highlighted in blue). Ratios of CRD-CG to RCT range from 1.43 to 1.56, and are smaller than those for RD that range from 1.97 to 2.12. Thus CRD-CG is more powerful at the cutoff than a basic RD design.

Figure 6 also plots the ratio of CRD-CG standard errors *above the cutoff* to the adjusted s.e. of RCT benchmarks above the cutoff (in green). The ratios range from 0.93 to 1.00, indicating that CRD-CG is at least as powerful as the RCT above the cutoff.
(please insert figures 6 here)

Conclusions:
We draw two conclusions. First, both the CRD-Pre and CRD-CG designs produced unbiased estimates at the cutoff, just as a basic RD does also; but the standard errors of CRD estimates were lower, being only 0.7 to 0.8 times those of the basic RD. Thus, at the cutoff, CRD designs are as unbiased as the basic RD but are more precise. Above the cutoff, both CRD designs produced unbiased estimates. The standard errors of CRD estimates were low, only 0.9 to 1.0 times those of RCT. Thus, CRD designs were unbiased and at least as powerful as RCT above the cutoff. In contrast, RD cannot generalize treatment effect above the cutoff. Thus CRD is strongly recommended to replace RD whenever possible.

Second, CRD-Pre and CRD-CG reduced bias equally but the power for each CRD design depends on the parameter values. The stronger the pretest-posttest correlation, the larger the power for CRD-Pre; the larger the number of comparison cases, the larger the power for CRD-CG. Researchers may construct either CRD design depending on what data is available - pretest data or a large sample of untreated cases.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Cook, T. D., Campbell, D. T., & Day, A. (1979). *Quasi-experimentation: Design & analysis issues for field settings* (pp. 19-21). Boston: Houghton Mifflin.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of policy analysis and management*, *27*(4), 724-750.

Cook, T. D., & Steiner, P. M. (2010). Case matching and the reduction of selection bias in quasi-experiments: the relative importance of pretest measures of outcome, of unreliable measurement, and of mode of data analysis. *Psychological Methods*, *15*(1), 56.

Goldberger, A. S. (1972). *Selection bias in evaluating treatment effects: Some formal illustrations*. University of Wisconsin--Madison.

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*,*69*(1), 201-209.

Imbens, G., & Kalyanaraman, K. (2011). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, rdr043.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 604-620.

Rubin, D. B. (2008). Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, *103*(484).

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

Shadish, W.R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage learning.

Shadish, W. R., Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random and cutoff-based assignment. *Psychological methods*, *16*(2), 179.

St.Clair, T., Cook, T.D., and Hallberg, K. (2014). Examining the Internal Validity and Statistical Precision of the Comparative Interrupted Time Series Design by Comparison with a Randomized Experiment. *American Journal of Evaluation,* 35 (2014): 311-327.

Tang and Cook (in prep). Statistical Power for the Comparative Regression Discontinuity Design in Education Research.

Wing, C., & Cook, T. D. (2013). Strengthening The Regression Discontinuity Design Using Additional Design Elements: A Within-Study Comparison. *Journal of Policy Analysis and Management*, *32*(4), 853-877.

Appendix B. Tables and Figures

Figure 1 CRD-Pre and RD biases from the RCT benchmarks for causal effects at the cutoff

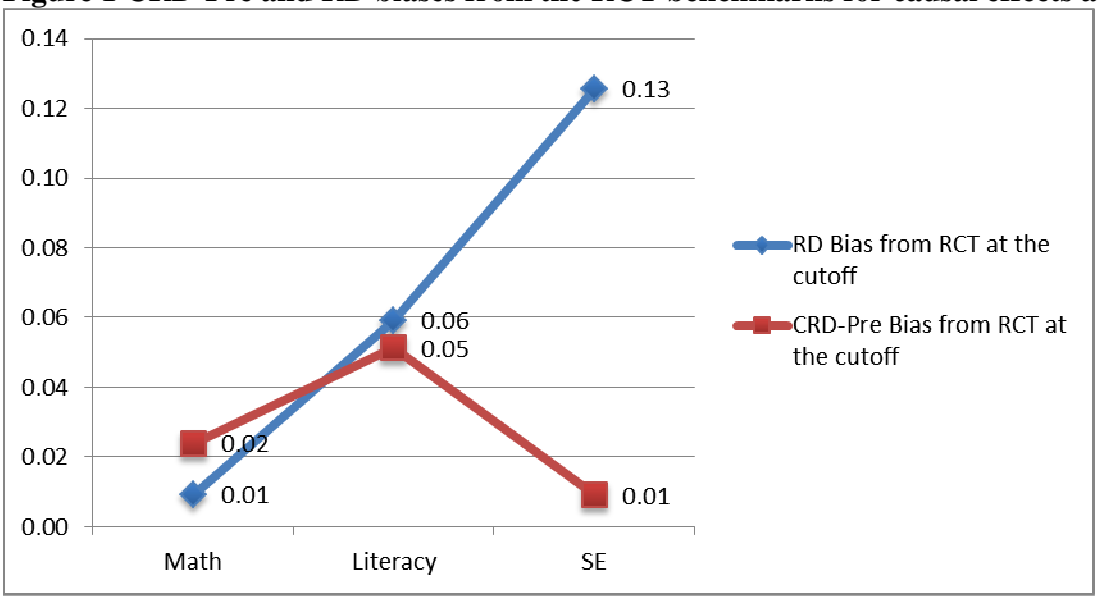

Figure 2 CRD-Pre biases from the RCT benchmarks for causal effects above the cutoff

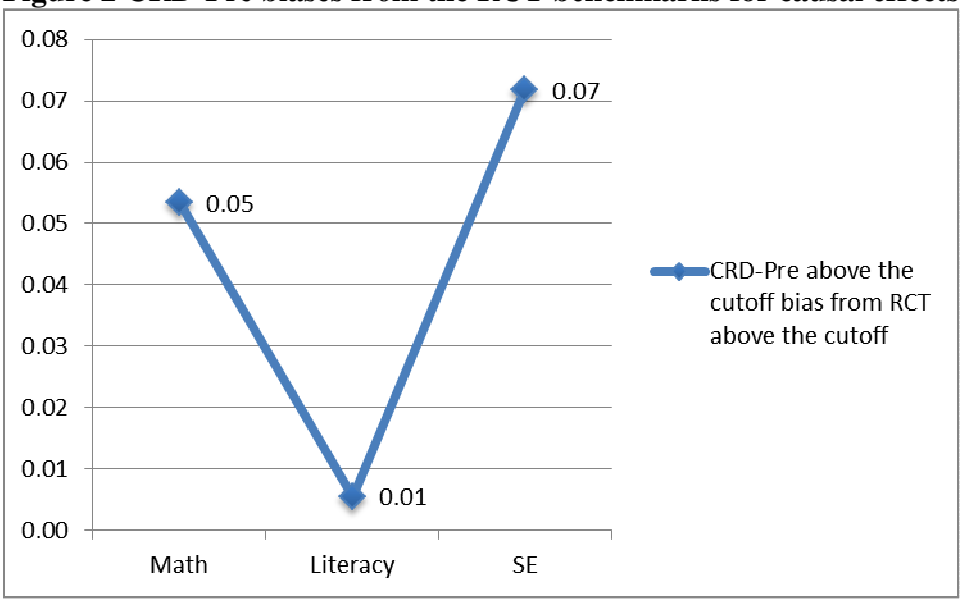

Figure 3 Standard errors of RD, CRD-Pre estimates at the cutoff and CRD-Pre estimates above the cutoff, relative to RCT benchmarks

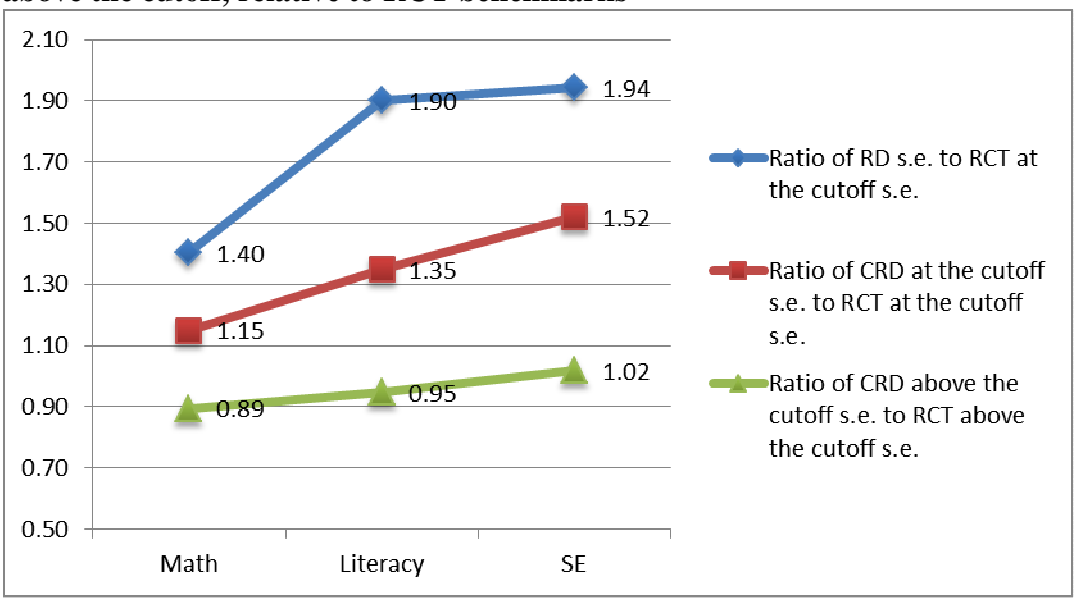

Figure 4 CRD-CG and RD biases from the RCT benchmarks for causal effects at the cutoff

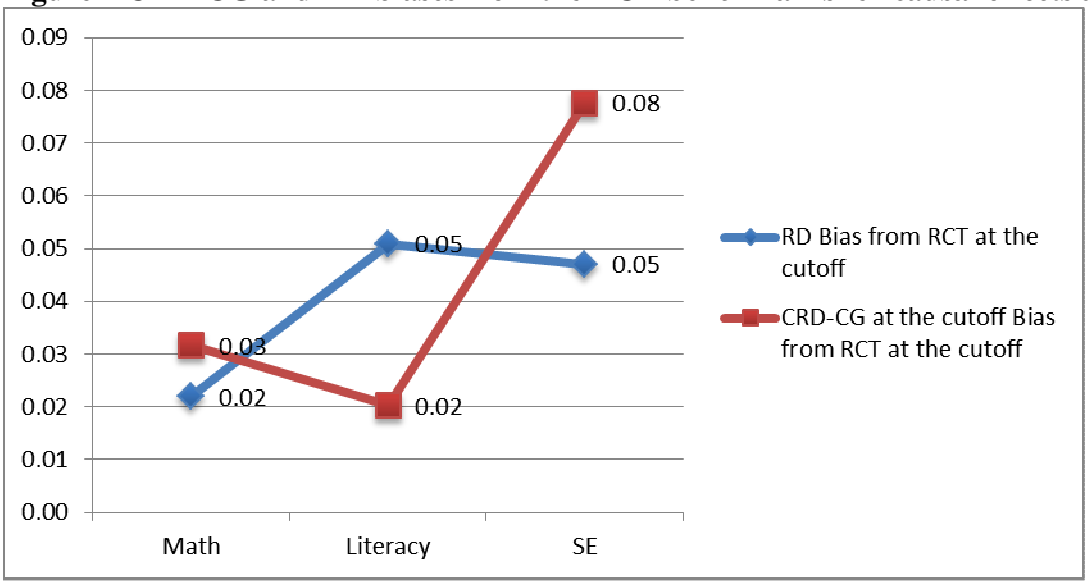

Figure 5 CRD-CG biases from the RCT benchmarks for causal effects above the cutoff

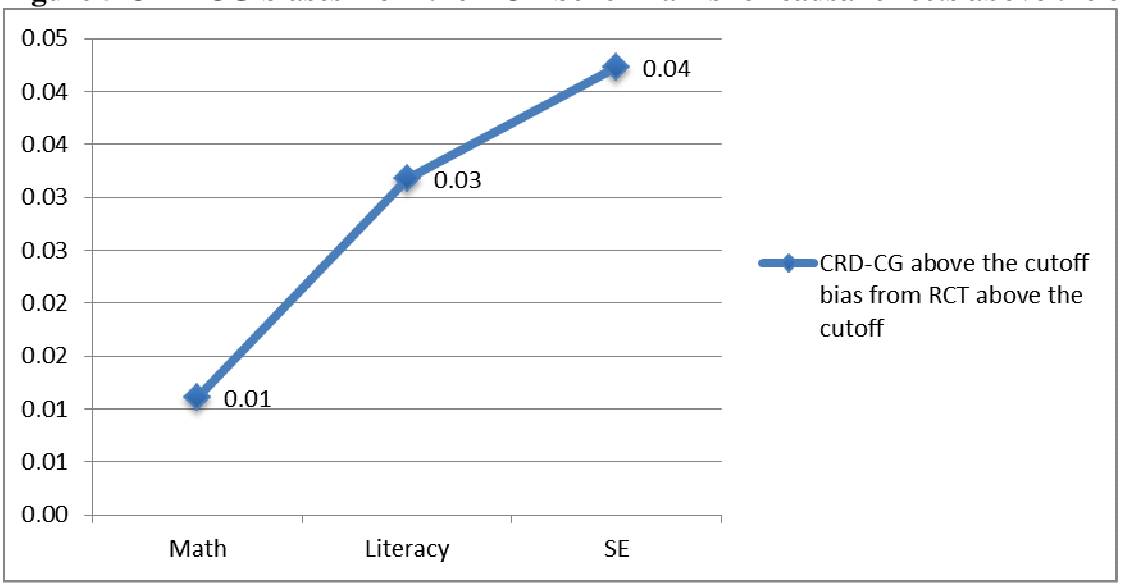

Figure 6 Standard errors of RD, CRD-CG estimates at the cutoff and CRD-CG estimates above the cutoff, relative to RCT benchmarks