Comparing Charter & Traditional Public Schools Using Propensity Score Analysis

Jason M. Bryer

University at Albany, SUNY

jbryer.github.com

jason@bryer.org

Author Note

Paper presented at the 2012 AERA Annual Meeting on April, 15, 2012 in Vancouver, British Columbia, Canada. This paper represents the results of a pilot study done using the 2007 NAEP for a dissertation that will utilize the 2009 NAEP. Only a subset of the figures and tables are provided here. The full set of tables and figures can be obtained at http://github.com/jbryer/Dissertation.

Abstract

The use of propensity score analysis (Rosenbaum & Rubin, 1983) has gained increasing popularity for the estimation of causal effects within observational studies. However, its use in situations where data is multilevel, or clustered, is limited (Arpino & Mealli, 2008; Hong & Raudenbush, 2006; Thommes & West, 2011). This study will introduce the `multilevelPSA` (Bryer, 2011) package for R that provides functions for estimating propensity scores for large datasets using logistic regression and conditional inference trees. Furthermore, a set of graphical functions that extends the framework of visualizing propensity score analysis introduced by Helmreich and Pruzek (2009b) to multilevel analysis will be discussed. An application for estimating the effects of charter schools as compared to traditional public schools on reading and mathematics in grades 4 and 8 is provided.

Keywords: charter schools, propensity score analysis, multilevel analysis

Comparing Charter & Traditional Public Schools Using Propensity Score Analysis

The concept of school choice within the United States is not new. Private schools have been educating students since the founding of the United States. However, in 1988, Ray Budde proposed an alternate approach to school choice that has grown to be known as charter schools (Kolderie, 2005). Unlike their private school counterparts, charter schools receive public funding, but they are relieved of many of the bureaucratic and regulatory constraints public schools adhere to, but are still held accountable for student performance. Despite claims by charter school advocates that charter schools are performing as well if not better than the public school counterparts (see e.g. Allen, Consolettie, & Kerwin, 2009), studies provide mixed results with regard to charter school performance (see e.g. Braun, Jenkins, & Grigg, 2006; Center for Research on Education Outcomes, 2009; Hubbard & Kulkarni, 2009). Ultimately, there is agreement that more research is necessary to address the question of whether charter schools provide substantially better academic experiences for students. This study will investigate the question of whether students who attend charter schools outperform their public school counterparts on two key academic domains: reading and mathematics.

At the center of the charter school debate is an issue that has concerned thinkers for centuries, the question of how to support causal inferences. In this context we aim to assess the effect of a treatment on a measurable outcome using observational data. Traditionally, randomized experiments have been considered the “gold standard” for studying questions of causality. In the context of this study, and many other questions in education randomization is either impractical or unethical. This means that observational studies are often essential to support inferences about whether treatments influence outcomes.

However, lacking randomization, observational studies nearly always invite selection bias. Charter schools are, by definition, schools of choice. In observational data contexts, simple comparisons of two groups such as public and charter schools cannot help but ignore the inherent and systematic differences between the two groups. However, with an

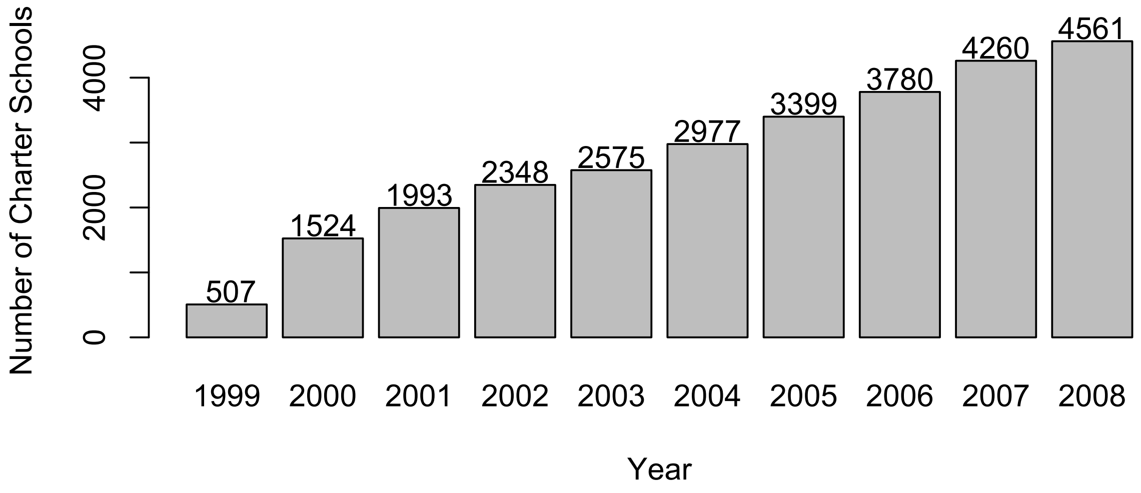

Figure 1. Charter School Growth 1999-2008

appropriately designed observational study, and an appropriate analysis, the effects of the selection bias can be taken into account in a way that simple comparisons are replaced by adjusted comparisons of groups. This is done utilizing a class of statistical procedures introduced by Rosenbaum and Rubin (1983) called propensity score analysis.

Propensity score analysis has seen considerably increased use in the social sciences within the last few years Thoemmes and Kim (2011). However, its use in situations where multilevel, or clustered data are of interest, have been limited (Arpino & Mealli, 2008). Using data from the 2007 National Assessment of Educational Progress (NAEP) for mathematics and reading at grades four and eight, estimates of the differences between charter and public schools will be calculated at two levels, namely state and national. Given the variability of charter schools laws across states, it is important to consider the impact of clustering. Analysis will be conducted using the multilevelPSA package in R (R Development Core Team, 2008). Specifically, propensity scores will be estimated within each state and these will be used for matching or stratification of students within each state. Comparisons of specific students, or groups of students, will in all cases be within states. Effects will then aggregated to provide state and national effect estimates.

Method

As with all propensity score analyses, it is preferable to utilize multiple methods for estimating propensity scores (see e.g. Stuart, 2010). Most of the studies conducted using PSA involve analysis in two phases where phase one involves the calculation of propensity scores or matching for both treatment and control units of analysis; and phase two involves the comparison of those two groups. However, there is little research with regard to situations where data is multilevel. As such, this study will be organized as such:

1. *Propensity score analysis using stratification.* This method ignores state assignment as a clustering variable. Under this broader method three statistical methods for stratification will be used:

   (a) Full logistic regression. This method will estimate propensity scores using logistic regression with all available covariates.

   (b) Logistic regression with step AIC. The `stepAIC` in the `MASS` package (Venables & Ripley, 2002) will select the best logistic model based upon the Akaike Information Criterion (Akaike, 1974). In this case the "best" first order interaction terms will be added to the main effect terms in a.

   (c) Conditional inference trees, based on all covariates; missing data will also be accommodated with the tree-based methods.

2. *Propensity score matching.* This method implicitly accounts for clustering. That is, the method used will find matches between treated and control units that first match exactly on state, ethnicity, and gender, then finds a best match based upon the propensity scores estimated using logistic regression. As suggested by Stuart (2010), multiple matched sets will be formed using (charter-to-traditional public school students):

   (a) One-to-one.

   (b) One-to-five.

   (c) One-to-ten.

   A dependent sample analysis will be performed on the resulting matched pairs (Austin, 2011).

3. *Multilevel propensity score analysis.* This method will utilize the same stratification methods as described in method one above, namely:

   (a) Full logistic regression.

   (b) Logistic regression with step AIC.

   (c) Conditional inference trees.

   However, where this method differs from method one is that separate models will be estimated for each state separately. Results from each state are then aggregated to provide an overall, national estimate of the differences between charter and traditional public school. Moreover, this method provides an approach whereby differences between meaningful subgroups (states in this study) can be explored, especially with the use of graphics as described below.

Visualizing Multilevel PSA

Given the large amount of data that needs to be summarized, the use of graphics will be an integral component of representing the results. The multilevelPSA1 package in R provides a number of graphing functions that extend the framework introduced by Helmreich and Pruzek (2009a) for multilevel PSA. Figure 2 represents a multilevel PSA assessment plot. In this graphic, the x-axis corresponds public school grade 4 NAEP scores and the y-axis corresponds to charter school grade 4 NAEP scores. Each colored circle is a state with its size corresponding the number of students within each state. Each state is projected to the lower left, parallel to the unit line, such that a tick mark is placed on the line with slope -1. These tick marks represent the distribution of differences between charter and public schools across states. Differences are aggregated (and weighted by size) across states. For grade 4 math, the overall adjusted mean for charter school students is 236 and the overall adjusted mean for public school student is 237 and represented by the horizontal and vertical blue lines, respectively. The dashed blue line parallel to the unit line corresponds to the overall adjusted mean difference and likewise, the dashed green lines correspond to the confidence interval. Lastly, rug plots along the right and top edges of the graphic correspond to the distribution of each state's overall mean charter and public school NAEP scores, respectively.

Figure 2 provides a more nuanced depiction of the differences both between and states. Similar to the multilevel PSA assessment plot, each blue dot corresponds to a state and is sized relative to the number of students within each state. The light gray dots correspond to each strata within each state. The graphic also provides confidence intervals for each state as well as the overall adjusted mean difference (the vertical blue line) and confidence interval (the vertical green lines).

Results

Figure 3 provides a summary of all the results across the four dependent variables. This graphic displays the mean adjusted differences as blue dots with the confidence interval in green. Therefore, any instance where the confidence interval does not cross zero (as represented by the vertical black line) indicates there is a statistical significant difference. Examining Figure 4, we see there is general agreement across the nine different methods used. That is, given the various approaches to adjusting selection bias with the available student information provides consistent conclusions. Specifically, these results suggest there is no statistical difference between charter public school students in grade 4 math and reading. Grade 8 math provides mixed results whereby five of the nine methods

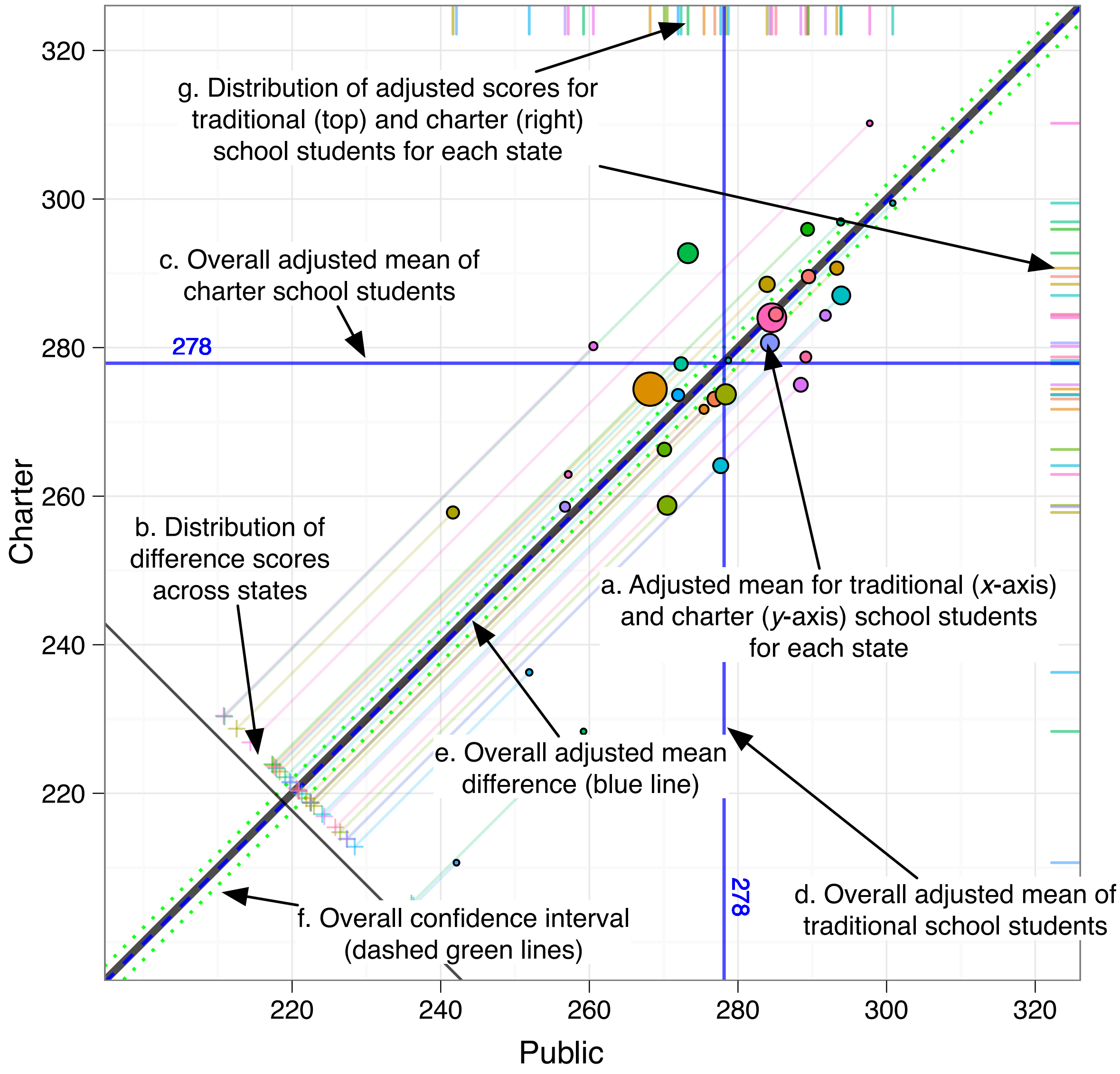

Figure 2. Multilevel PSA Assessment Plot

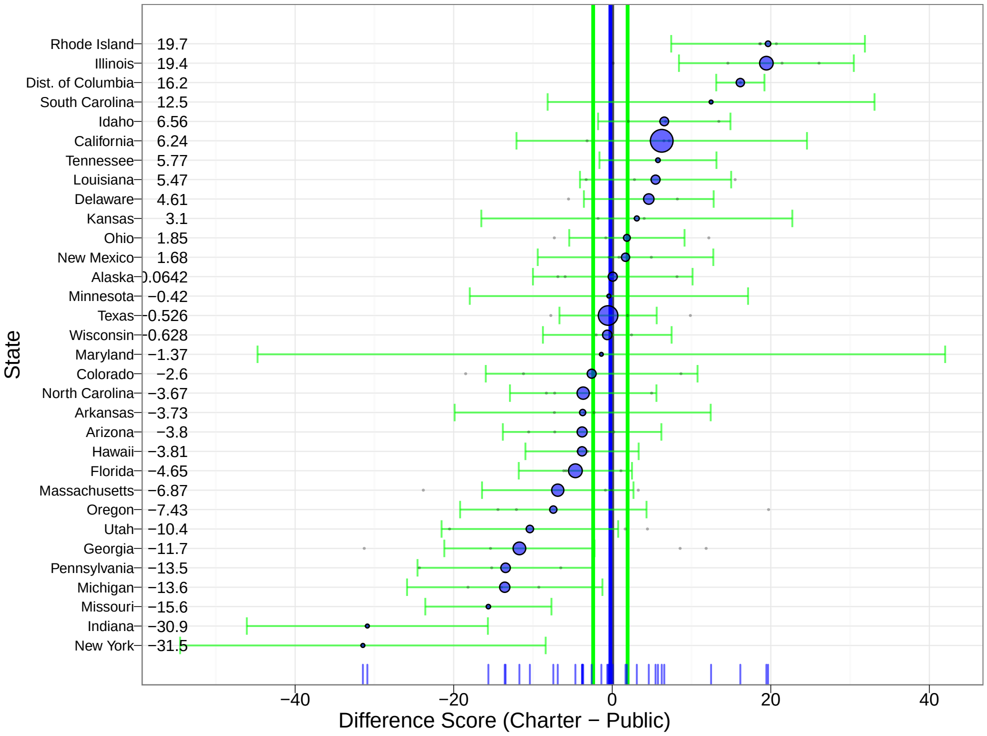

Figure 3. Multilevel PSA Difference Plot: Grade 8 Math

have confidence intervals that do not span zero. For grade 8 reading however, there is a clear statistical difference as all nine methods result in confidence intervals that do not span zero (effect size < .20).

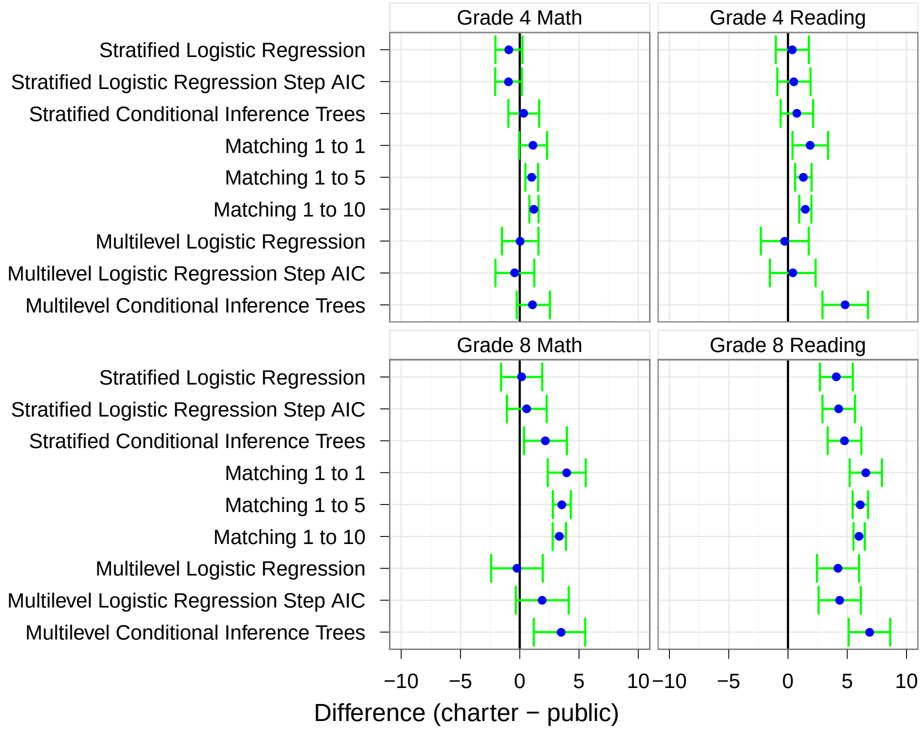

Figure 4. Summary of Overall Results

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Allen, J., Consolettie, A., & Kerwin, K. (2009). *The accountability report: charter schools.* The Center for Education Reform.

Arpino, B., & Mealli, F. (2008). *The specification of the propensity score in multilevel observational studies.* Munich Personal RePEc Archive. Retrieved from http://mpra.ub.uni-muenchen.de/17407

Austin, P. C. (2011). Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched smaples. *Statistical in Medicine*, *30*, 1292–1301.

Braun, H., Jenkins, F., & Grigg, W. (2006). *A closer look at charter schools using hierarchical linear modeling*. U.S. Government Printing Office.

Bryer, J. (2011). *Multilevelpsa: multilevel propensity score analysis*. R package version 1.0. Retrieved from http://multilevelpsa.r-forge.r-project.org

Center for Research on Education Outcomes. (2009). *Multiple choice: charter school performance in 16 states*. Stanford, CA: Stanford University.

Helmreich, J. E., & Pruzek, R. M. (2009a). PSAgraphics: an R package to support propensity score analysis. *Journal of Statistical Software*, *29*(6).

Helmreich, J. E., & Pruzek, R. M. (2009b). [computer software]. PSAgraphics: propensity score graphics. Retrieved from http://cran.r-project.org/web/packages/PSAgraphics/index.html

Hubbard, L., & Kulkarni, R. (2009). Charter schools: learning from the past, planning for the future. *Journal of Educational Change*, *10*, 173–189.

Kolderie, T. (2005). Ray budde adn the origins of the charter school.

R Development Core Team. (2008). [computer software]. R: a language and environment for statistical computing. r foundation for statistical computing. R Development Core Team. Vienna, Austria.

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

Stuart, E. A. (2010). Matching methods for causal inference: a review and a look forward. *Statistical Science*, *25*, 1–21.

Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of preonsity score methods in the social sciences. *Multivariate Behavioral Research*, *46*, 90–118.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth). ISBN 0-387-95457-0. New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Table 1

Summary Propensity Score Analysis using Stratification

	Adjusted Mean		x				
	Public	Charter	Diff	ATE	*n*	95% CI	
	Grade 4 Math						
Logistic Regression	237.95	237.38	-0.57	-0.57	146656.00	-1.72	0.57
Logistic Regression Step AIC	237.96	237.35	-0.60	-0.60	146656.00	-1.73	0.52
Conditional Inference Trees	237.93	238.27	0.34	0.34	146638.00	-0.96	1.64
	Grade 4 Reading						
Logistic Regression	218.27	218.96	0.70	0.70	141352.00	-0.71	2.11
Logistic Regression Step AIC	218.26	219.22	0.96	0.96	141352.00	-0.44	2.37
Conditional Inference Trees	218.26	219.01	0.75	0.75	141340.00	-0.62	2.12
	Grade 8 Math						
Logistic Regression	278.77	279.10	0.33	0.33	97563.00	-1.40	2.06
Logistic Regression Step AIC	278.77	279.54	0.77	0.77	97563.00	-0.91	2.46
Conditional Inference Trees	278.72	280.89	2.17	2.17	97521.00	0.36	3.98
	Grade 8 Reading						
Logistic Regression	259.80	262.82	3.02	3.02	105486.00	1.55	4.49
Logistic Regression Step AIC	259.80	262.67	2.87	2.87	105486.00	1.36	4.38
Conditional Inference Trees	259.75	265.39	5.65	5.65	105468.00	4.36	6.93

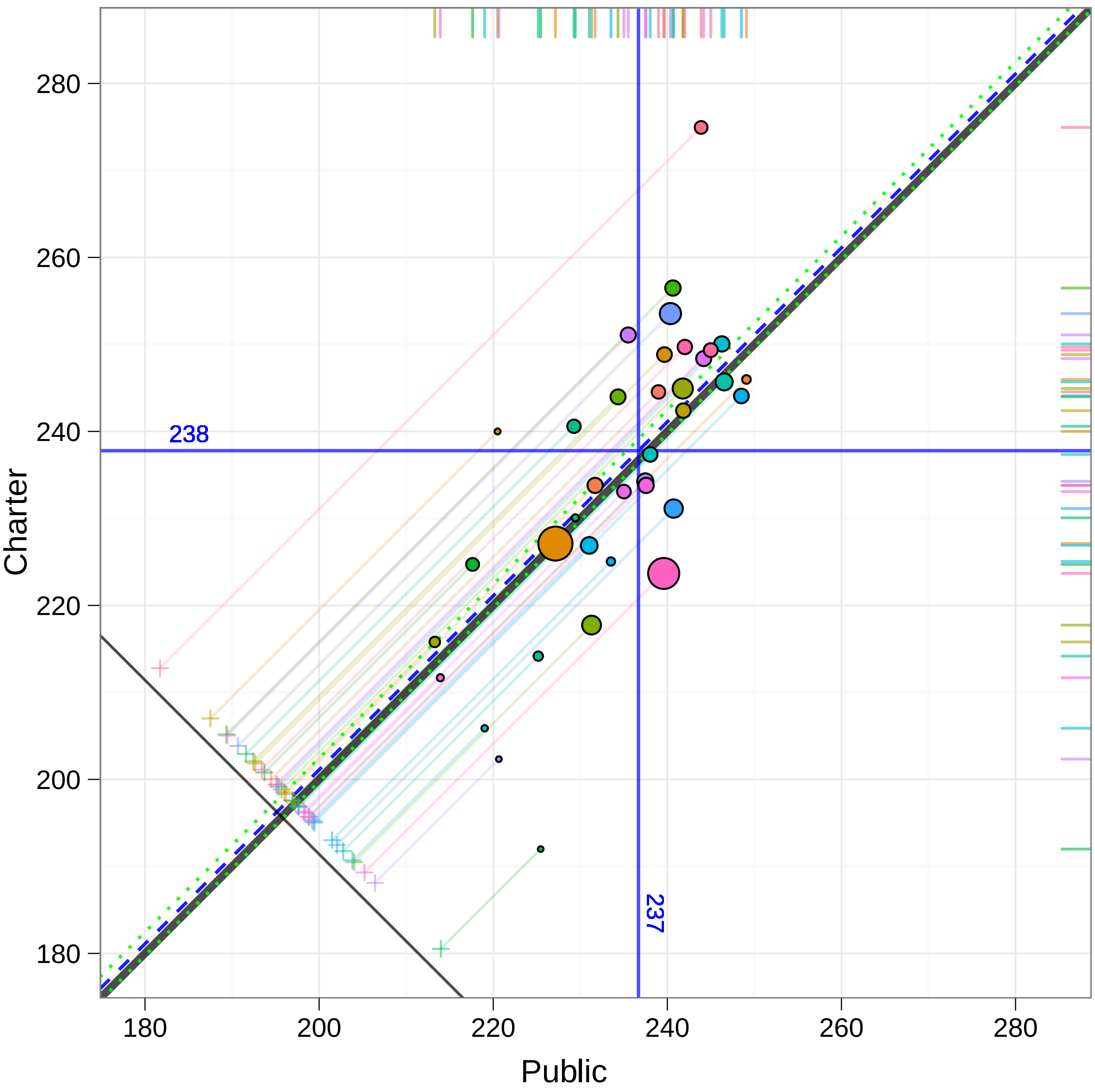

Figure 5. Multilevel PSA Assessment Plot: Grade 4 Math

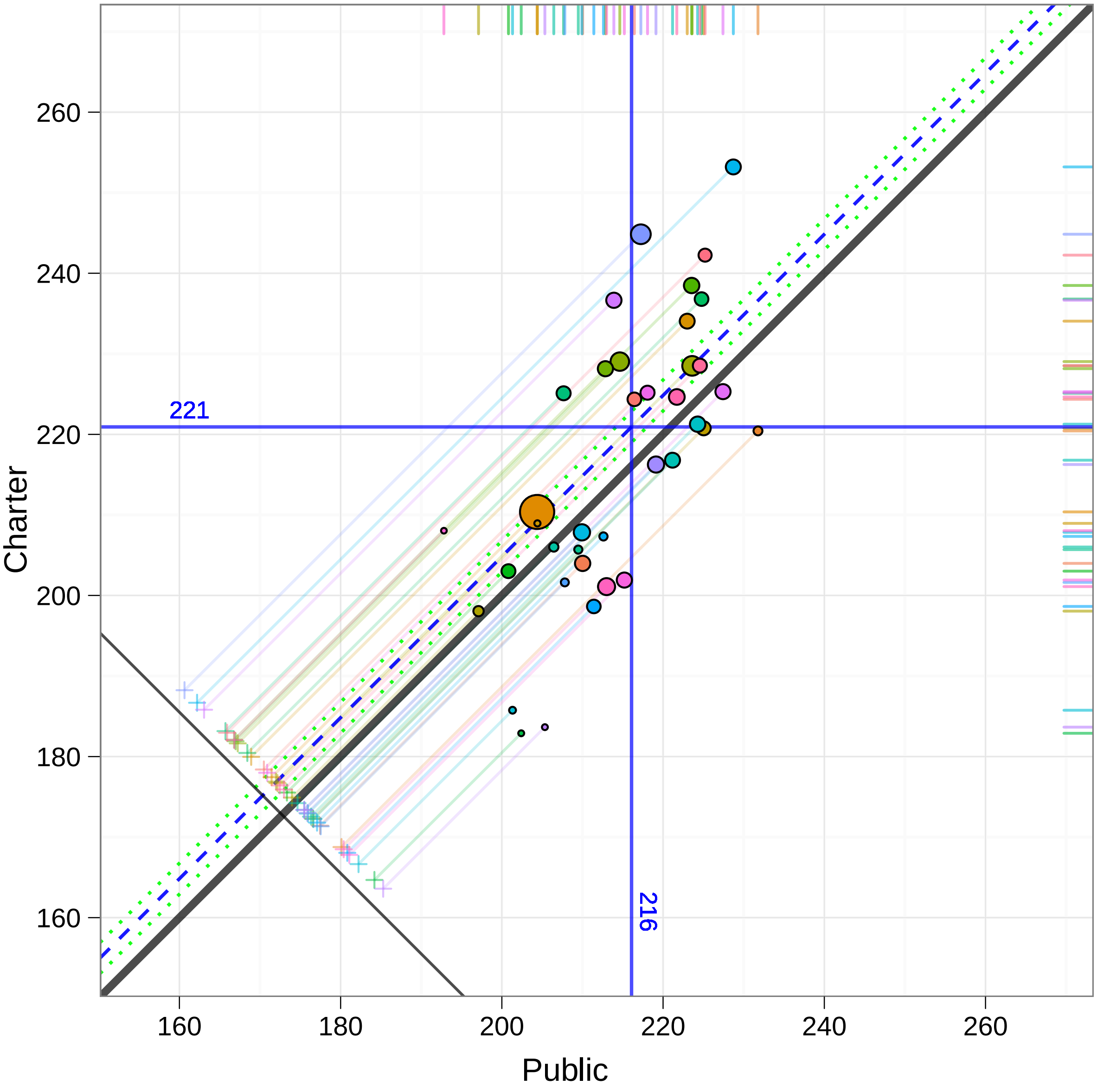

Figure 6. Multilevel PSA Assessment Plot: Grade 4 Reading

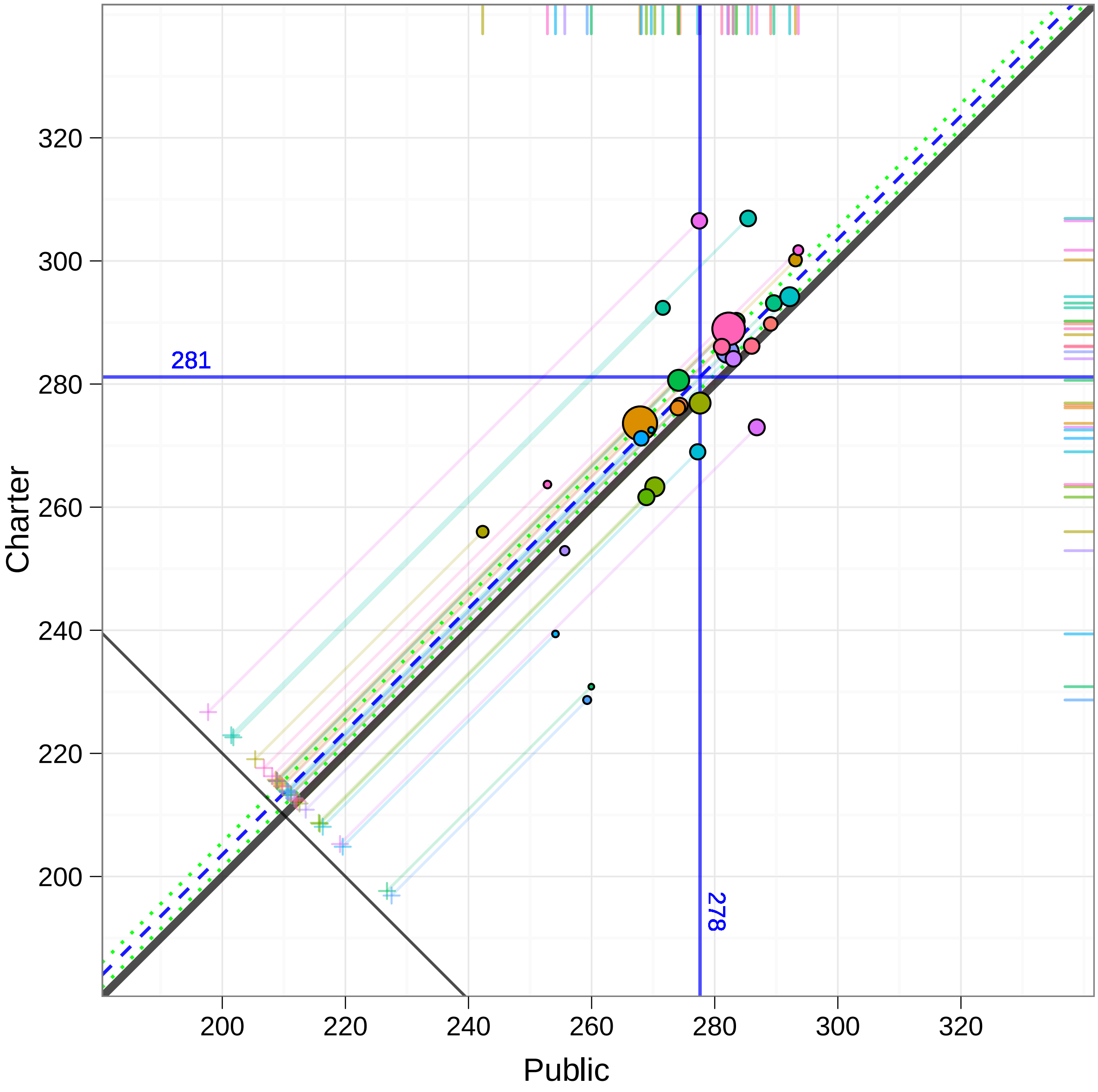

Figure 7. Multilevel PSA Assessment Plot: Grade 8 Math

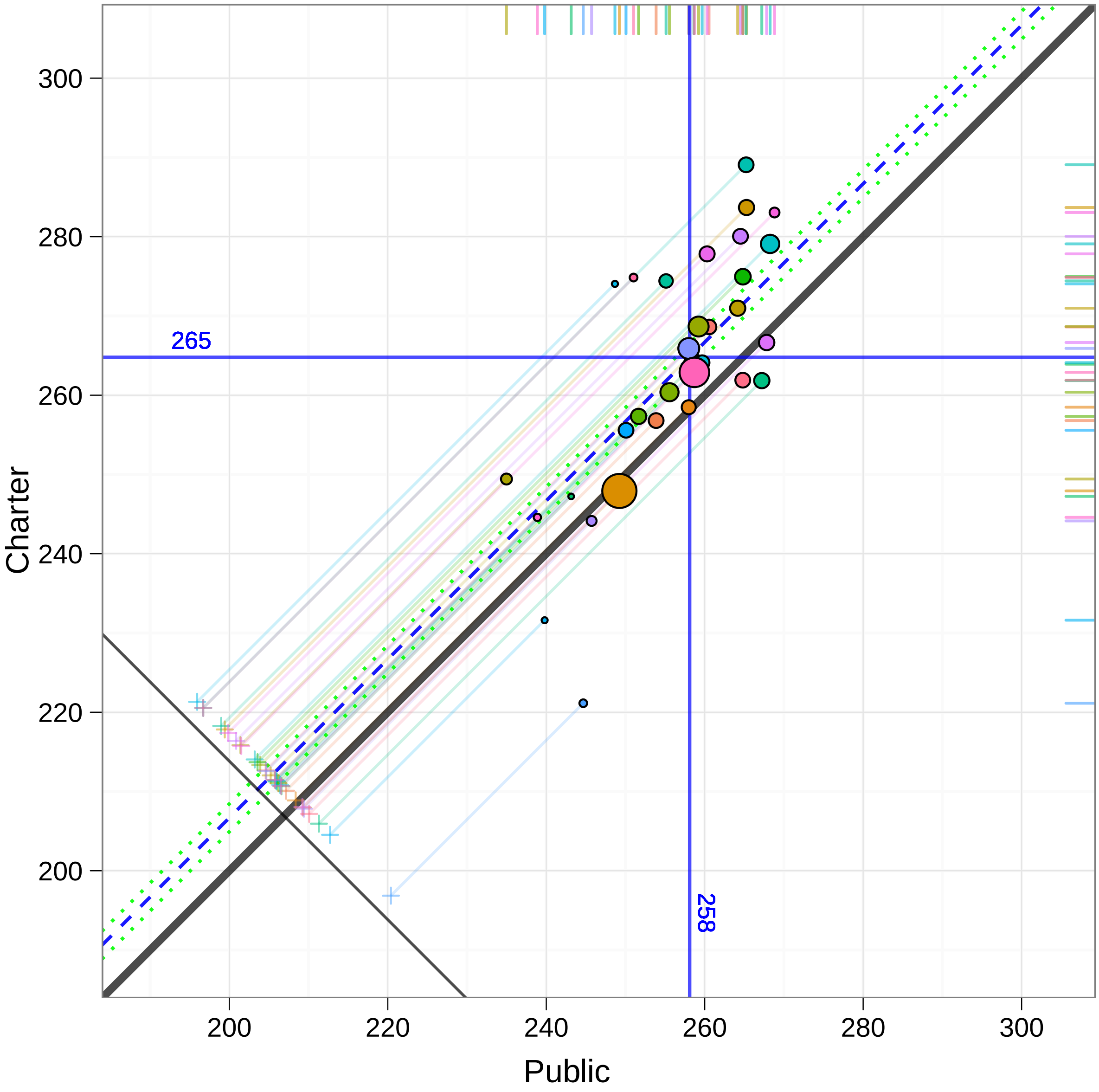

Figure 8. Multilevel PSA Assessment Plot: Grade 8 Reading