Readme file for "Using School Choice Lotteries to Test Measures of School Effectiveness"

David J. Deming, Harvard University and NBER

The data for this paper were made available to me through collaboration between Charlotte-Mecklenburg Schools (CMS) and the Center for Education Policy Research (CEPR) at Harvard University. The analysis was conducted using raw CMS data files that were prepared, cleaned and sent by CMS staff to analysts at CEPR.

To access the data, you must fill out an application for conducting outside research and submit it to CMS along with IRB approval for the project from your university. More information about the application process, including the necessary forms and contacts, can be found here:

http://www.cms.k12.nc.us/cmsdepartments/accountability/cfre/Pages/OutsideResearchProjects.aspx

This project brought together CMS data from several sources. Students are linked longitudinally and across data sources using a unique student ID which is present in all the data files listed below.

1) **Main CMS enrollment file** – this includes yearly information on school enrollment and demographics for students in grades K through 12.
2) **End-of-grade (EOG) test score file** – this includes yearly information on students' state-standardized math and reading scores from grades 3 through 8.
3) **2002-2003 lottery file** – this file includes each student's actual choices (including specialized programs within a school), their priority grouping in each lottery (based on sibling, neighborhood and other preferences), and their randomly generated lottery numbers. One can use this information to reconstruct the lottery assignments and verify that they complied with the stated procedure for randomization. This file was also used in Deming, Hastings, Kane and Staiger (2014).

I have prepared a de-identified data file for public use. The file "cms_VAManalysis.dta" and accompanying .do file can be used to replicate most of the results from the paper, although not exactly (more detail on that below). With access to the original data, the code can be used to replicate the paper exactly. This public use file differs from the actual data in the following respects:

1) To protect the confidentiality of individual students, I have removed any identifying information (including the linking ID and any geographic indicators).
2) I have removed the information about the lottery that allows one to reconstruct the randomization using preferences and priority groupings. Again, this is to protect student confidentiality. The variable "lottery_FE" is a generated ID that provides the correct unit of randomization used in the paper.
3) I have created school and school-by-grade IDs that are scrambled versions of the actual IDs. The school ID you find in the file is consistent across years and grades, but it is based on random perturbations (and thus, a reordering) of the actual school IDs.

4) To further guard against the identification of individual students, I have removed gender, race and free/reduced price lunch eligibility from the data. The concern is that data across schools and years, combined with small cell sizes (for example, if a school is only 5 percent Latino), allows for the identification of individual students.

Because of these restrictions, the results from the publicly available program and data differ slightly from the published version of the paper. Moreover, access to the restricted use data is required to estimate VAMs that make use of the demographic information. Models that control for prior test scores only can still be estimated.