How Well Do Structural Demand Models Work? Counterfactual Predictions in School Choice*

Parag A. Pathak and Peng Shi†

July 2017

Abstract

Discrete choice models of demand are widely used as the basis of counterfactual policy simulations, yet their performance out-of-sample is rarely assessed. This paper uses a large-scale policy change in Boston to investigate the performance of discrete choice models of school demand. In 2013, Boston Public Schools significantly altered their school choice plan by changing the set of schools to which applicants can apply. Earlier work estimated discrete choice models of demand to provide guidance about alternatives and influenced the selection of a new plan that dramatically altered choice sets for thousands of participants. To ensure a genuine out-of-sample comparison, Pathak and Shi (2014) reported predictions using discrete-choice demand models prior to the policy change and described prediction targets involving access, travel, and unassigned students. Here, we assess how well these ex ante counterfactual predictions compare to the choices made under the new choice sets. For predictions of equilibrium outcomes, a simpler *ad hoc* alternative performs as well as the discrete choice models we estimated, even though the discrete choice models are better at predicting school-level market shares. Adjusting for the new applicant population, the discrete choice models dominate on both dimensions. Demand models which allow more flexible substitution patterns do not significantly improve predictions. In our context, discrete choice demand models perform well at predicting choices, but the overall quality of predictions of the equilibrium effects of the new policy are undermined by changes in the underlying environment.

*The first report of this project was released as NBER Working Paper 19859 in January 2014. We thank Kamal Chavda, Superintendent Carol Johnson, Peter Sloan, Kim Rice, and Jack Yessayan and the staff at Boston Public Schools for their expertise and for facilitating access to the data used in this study. Josh Angrist, Glenn Ellison, Patrick Kline, Ariel Pakes, and Michael Whinston provided helpful feedback. Financial support provided by the National Science Foundation under grant SES-1426566 and the W.T. Grant Foundation. Pathak is on the scientific advisory board of the Institute for Innovation in Public School Choice.

†Pathak: Massachusetts Institute of Technology and NBER, Department of Economics, email: ppathak@mit.edu and Shi: USC Marshall School of Business, Department of Data Science and Operations, email: pengshi46@gmail.com.

Contents

1 Introduction **3**

2 Background **6**

2.1 School choice in Boston 6

2.2 Policy reform 7

3 Methodology **8**

3.1 Data sources 9

3.2 Choice models 10

3.2.1 Multinomial Logit (MNL) 10

3.2.2 Mixed MNL (MMNL) 12

3.2.3 Lexicographic 13

3.3 Predicting the Applicant Pool 14

3.4 Generating Predictions 16

3.5 Evaluating Predictions 17

3.5.1 Equilibrium Outcomes 17

3.5.2 School Choices by Neighborhood 17

4 Backtesting and Hypotheses **19**

5 Results **21**

5.1 Equilibrium and Choice Forecasts 21

5.2 Decomposing Prediction Errors 22

6 Conclusions **24**

6.1 Unobserved heterogeneity 26

A Computation Appendix **27**

A.1 Specifying the Likelihood Function 27

A.2 Overview of the MCMC procedure 28

A.2.1 Random Walk Metropolis (RWM) 29

A.2.2 Metropolis Within Gibbs (MWG) 30

A.2.3 Hamiltonian Monte Carlo (HMC) 30

A.3 The MCMC Sampler 32

1 Introduction

The aim of developing models capable of quantitatively forecasting the effects of policy changes has been an objective of economic science since at least Hurwicz (1950) and Marschak (1953). Recent years have seen the popularity of design-based research strategies that focus on estimating particular parameters or causal effects. The design-based approach does not immediately allow for ex ante policy evaluations involving changes outside of historical experience. Both Angrist and Pischke (2010) and Heckman (2010) attribute the growth of the design-based research program to skepticism on the value of structural modeling for counterfactual analysis given its heavy reliance on parametric and/or behavioral assumptions.

Though opinions vary on the credibility of structural models, there is one area of consensus: there are relatively few systematic comparisons of the ex ante counterfactual predictions of structural models after policy changes have occurred to predictions made from simpler methods. For instance, Angrist and Pischke (2010) lament:

> "Many new empirical industrial organization studies forecast counterfactual outcomes based on models and simulations, without a clear foundation in experience. [...] At minimum, we'd expect such a judgement to be based on evidence showing that the simulation-based approach delivers reasonably accurate predictions. As it stands, proponents of this work seem to favor it as a matter of principle."[1]

The goal of this paper is to fill this void by evaluating the performance of predictions from discrete-choice models of demand, which underlie much of the work in the new empirical industrial organization, using a large-scale policy change that affected thousands of families in Boston in 2014.

Each year, thousands of Boston's families submit rank order lists of public schools in the city's student assignment plan.[2] In 2013, Boston Public Schools' (BPS) officials, the Mayor, and members of the school committee wanted to change this plan to encourage students to attend schools closer to home, partly motivated by a desire to reduce transportation costs. BPS put forth a number of plans that redraw the boundaries of the city and modify applicant choice sets. After these plans were publicly presented, there was widespread interest in predicting the choices families would make and in comparing the final assignments under these alternatives.[3] The mayor delayed the timeline for selecting a plan after asking us to conduct an analysis (Menino, 2012b). The report we

[1] There is a small literature in the context of merger analysis. Peters (2006) examines the predictive value of airline mergers and find that structural simulation methods yield poor predictions of post-merger ticket prices. Ashenfelter and Hosken (2008) argue that "transparently" identified design-based estimates of the mergers differ markedly from those produced by the structural approach. Nevo and Whinston (2010) describe a handful of examples of counterfactual validations in the context of merger analysis. Einav and Levin (2010) acknowledge that there should be more retrospective analysis of past mergers, but question whether information from particular mergers can be extrapolated for previous ones.

[2] The Boston assignment system has been subject to a number of theoretical studies including Abdulkadiroğlu and Sönmez (2003), Abdulkadiroğlu, Pathak, Roth, and Sönmez (2005), Pathak and Sönmez (2008), and Dur, Kominers, Pathak, and Sönmez (2016).

[3] For more details, see the materials available at http://bostonschoolchoice.org and press accounts by Goldstein (2012) and Handy (2012). BPS communications reported more than 1,850 residents offered feedback on the plans. For specific reactions to proposed plans, see Vaznis and Andersen (2012) and Burge (2012).

issued, Pathak and Shi (2013), used historical choices to predict the choices that participants would rank in the choice sets in the different proposals. Our report played a significant role in helping administrators and the public to evaluate tradeoffs among the proposals and ultimately to select a new plan. By January 2014, families throughout Boston ranked schools under these new choice sets. For a representative applicant, 9.4 school choices stayed the same, 16 choices were removed, and 2.5 choices were added. Figure 1 summarizes the timeline of the reform.

In this paper, we use submitted rank-order list data from new applicants in the first year of the new assignment plan to evaluate the accuracy of predictions in Pathak and Shi (2014), the updated predictions issued right before the reform. Specifically, we ask two questions:

1. Policy question: How well do structural models perform in predicting equilibrium outcomes important for the school choice context?

2. Scientific question: How accurate are discrete choice models in predicting school demand?

The major innovation of our research design is that predictions were made prior to the policy change. Figure 2 presents the research design. The benefit of committing to forecasts before the policy change is that there is no scope for modifying forecasts once realized data are in hand. In this respect, our exercise contrasts with other validations of structural models that use social experiments as a validation tool for structural models (see, e.g., Wise (1985) and Todd and Wolpin (2006)). This format allows us to report unexpected outcomes of the forecasting exercise and therefore provide a genuine out-of-sample assessment of discrete choice models for our setting. Moreover, Pathak and Shi (2014) report forecasts from a simpler "non-structural" model. This reference point allows us to consider the relative performance of discrete choice methods to a non-discrete choice alternative.

The prediction targets we focus on were of central interest to policymakers and the public discussion. For students in each of Boston's 14 neighborhoods, we predict their chances of getting into high performing schools, their average distance to their assigned school, and the number of applicants who would be unassigned after the first round of assignment. Each target was part of earlier reports (Pathak and Shi, 2013). The metrics we use to evaluate the accuracy of choice models include how well they predict the individual choices and how well they predict the distribution of choices for each neighborhood.

The discrete choice methods we consider are the multinomial logit (MNL) model and the mixed MNL model, which allows for random coefficients. In a pioneering contribution, McFadden and co-authors used the MNL model to study the impact of BART, a new fixed-rail rapid transit system in San Francisco Area (McFadden, 2001). They collected data on the travel behavior of a sample of individuals in 1972, prior to the introduction of BART, and estimated models were used to predict the behavior of the same individuals in 1975 after BART began. Multinomial logit models were estimated using pre-BART commuter data at the realized attributes of the alternatives (auto alone, carpool, bus). McFadden, Talvitie, and Associates (1977) provide a detailed account of the performance of these models. McFadden (2001) reports that

> "our overall forecasts for BART were quite accurate, particularly in comparison to the official

1973 forecast [...]. We were lucky to be so accurate, given the standard errors of our forecasts, but even discounting luck, our study provided strong evidence that disaggregate RUM-based models could outperform conventional methods."

Based in part on the BART experience, random utility models are widely employed in travel analysis and many other areas of economics involving consumer choice (McFadden, 2001). More than three decades have progressed since the BART analysis and there has been considerable progress in demand modeling during this time. For instance, McFadden (2001) states that the methods used to account for substitution between modes of transportation were inferior to the modeling methods used today. In addition to serving as one of the only post-BART validations of discrete choice models we are aware of, our work evaluates the importance of accounting more flexible substitution patterns in counterfactual prediction.

Our exercise has the potential to provide unusually compelling evidence on the forecasting performance of structural demand models and their value in counterfactual analysis. First, the forecasts are based on flexible models of demand exploiting the revealed preferences that families expressed historically. The data not only includes student's top choices, but their entire submitted rank order list of schools. Rank order list data can potentially reveal richer information about substitution patterns between choices (see, e.g., Berry, Levinsohn, and Pakes (2004)). Moreover, our dataset includes a large number of observables, including information on student characteristics and exact geographic location. Furthermore, Boston's choice plan has been in existence for more than two decades, meaning that there is a wealth of knowledge and shared experience about the system. The current strategy-proof system, in place since 2005, eliminates the need for participants to be strategic about their choices and the advice BPS provides participants in their school guide reflects this feature.[4] Our exercise should be particularly informative on substitution patterns since the policy change predominantly involves a change in choice menus and relatively small changes in the characteristics of choices. As preference data from school assignment plans become more widely available, there are also a growing number of papers that estimate models of school demand, and our results will speak to their reliability as a policy planning tool for school districts.[5] Finally, since discrete choice models of demand are a building-block for many structural models that examine more involved situations involving dynamics and strategic interactions, the relatively simple counterfactual environment should allow us to easily compare our predictions and understand reasons for their performance with these additional complications.

On the other hand, the premise of our exercise, and other forecasts based on discrete choice models of demand is that preferences are stable, can be estimated adequately, and can be used

[4]For instance, the 2012 School Guide states: "List your school choices in your true order of preference. If you list a popular school first, you won't hurt your chances of getting your second choice school if you don't get your first choice.

[5]The growing literature estimating school demand includes: Abdulkadiroğlu, Agarwal, and Pathak (2013), Abdulkadiroğlu, Pathak, Schellenberg, and Walters (2017), Agarwal and Somaini (2014), Burgess, Greaves, Vignoles, and Wilson (2015), Calsamiglia, Fu, and Guell (2017), Glazerman and Dotter (2016), Hastings, Kane, and Staiger (2009), He (2012), Hwang (2015), Kapor, Neilson, and Zimmerman (2017), Harris and Larsen (2015), Ruijs and Oosterbeek (2012), and Walters (2013).

to extrapolate to different environments. Along with a change in the choice set in Boston's new plan, the district also presented the choice set in a different way, which may induce a change in underlying preferences. In a field experiment, Hastings and Weinstein (2008) provide evidence that choice behavior in Charlotte's school choice plan can easily be swayed by informational cues. In other contexts, there is similar evidence that information interventions that give people the same information that is already available in a simpler format change choice behavior, and are evidence of "comparison frictions" (Kling, Mullainathan, Shafir, Vermuelen, and Wrobel, 2012). If these features of choice dominate decision-making, then they may interfere with the reliability of any forecasts based on historical revealed preferences. Indeed, if the details of how counterfactuals are presented matter more than what we can learn from past data using demand models, it may motivate a re-evaluation of the use of demand models for analyzing counterfactuals in our context.

The rest of this paper is structured as follows. Section 2 provides additional details on the Boston student assignment plan and events leading up to the policy change in 2014. Section 3 describes our prediction methodology and the choice models we estimate. Section 4 reviews hypothesis motivated by back-tests of our framework using data from from two years prior to the policy change to make predictions for the year prior to the reform. Section 5 compares our predictions to the the realized outcomes in the first year of the new policy. It also decomposes sources of prediction error by examining changes to the set of participants or underlying stability of the demand model. Section 6 concludes and offers directions for future work.

2 Background

2.1 School choice in Boston

Boston Public Schools is home to one of the nation's most well-known school choice plans, which initially evolved out of a court-ordered busing plan in the 1970s. From 1988 to 2013, the city was divided into the North, West, and East Zone for elementary school admissions. There are about 25 elementary schools in each zone. Students residing in a zone are allowed to rank any school choice in the zone as well as any school within a 1 mile walk-zone of their residence and a handful of city-wide schools on their application form. At each school, students are prioritized as follows: continuing students (who are already assigned to the school at an earlier grade) have the highest priority, followed by students who have an older sibling at the school, followed by other students. Within each group, for half of the program seats, students residing in the walk-zone obtain priority (but this priority does not apply to the other half of the school seats). A single lottery number draw serves as a tie-breaker.[6]

Since 2005, after students submit their choices, they are processed through the a version of Gale and Shapley (1962)'s celebrated student-proposing deferred acceptance algorithm. This algorithm takes as input the submitted preference rankings of students, as well as the priorities of students,

[6]Dur, Kominers, Pathak, and Sönmez (2016) present additional details on Boston's implementation of this algorithm.

and outputs an assignment. The algorithm works as follows:

1. Each student applies to his first choice school. Each school ranks applicants by their priority, rejecting the lowest-ranked students in excess of its capacity. The rest of applicants are provisionally admitted: they are not rejected at this step but may be rejected in later steps.

2. The rejected students apply to their next most preferred school (if any). Each school considers these new applicants *and* provisionally admitted students, ranks them by their priority, rejecting the lowest-ranking students in excess of capacity. This produces a new admit of provisionally admitted students at each school.

The algorithm terminates when either every student is matched to a school or every unmatched student has been rejected by every school he has ranked. It is a weakly dominant strategy for all participants to rank schools truthfully (Dubins and Freedman, 1981) and (Roth, 1982). Moreover, this algorithm produces a stable assignment (Gale and Shapley, 1962; Abdulkadiroğlu and Sönmez, 2003).

2.2 Policy reform

The policy reform, which took place starting in 2014, affected the set of schools each applicant is allowed to rank. There were two major rationales for the reform. First, a number of stakeholders wished to assign students closer to home, including families who wanted assignments at nearby schools and the district which aspired to reduce busing costs.[7] Second, there was longstanding concerned about inequities between the three zones in access to high-performing schools.

The reform was informed by the Pathak and Shi (2013) report, which used discrete choice modeling and simulations to evaluate various proposed plans. The study was commissioned by a mayoral-appointed city committee, which met for over a year and hosted community meetings to collect feedback and discuss proposals. The plans compared include alternative division of assignment zones proposed by BPS.[8] The methodology of the report inspired BPS to later propose a 10-zone plan, as well as a modified 11-zone plan, and consider other plans from the community. Shi (2015) provides a more detailed account on the role of the counterfactual study in quantifying trade-offs between plans. It's worth noting that prior to the reform in 2013, there were two failed attempts to reform the choice sets of students, in 2003 and 2009. These prior attempts did not have access to a simulation-based study describing the effects of the policy.

Based on the Pathak and Shi (2013) report and other discussions, the Boston school committee chose a plan devised by Shi (2013), which became known as the home-based plan. The plan

[7]This motivation was emphasized by Mayor Menino, who spent the last year of his administration advocating for a "radically different school assignment process—one that puts priority on children attending schools closer to their homes" Menino (2012a).Other districts have also moved closer to home; see, e.g., the discussion about Seattle in Pathak and Sönmez (2013).

[8]The initial plans suggested by BPS include dividing the city into 6, 9, 11, or 23 zones, or assignment based purely on neighborhood. When these plans were publicly unveiled in September 2012, they were met with widespread criticism (see, e.g., Seelye (2012)).

constructs choice sets based on the exact residential address of applicants. It makes uses of a BPS categorization of schools into tiers, which are computed using a combination of standardized test score growth and levels from the Massachusetts Comprehensive Assessment System (MCAS) tests for the past two years. For the 2014 admissions cycle, the tiers were finalized as of January 2013. Under the new plan, every family is allowed to choose from any school within a mile (as the crow flies), along with the two closest Tier 1, the four closest Tier 1 or 2, the six closest Tier 1, 2 or 3 schools, and the three closest "option schools" chosen by BPS. The set of choices also includes the closest early learning center (ELC) and closest school with an advanced work class (AWC) program.[9] Families can easily access their choice set via an online portal, which shows a map of all schools in the choice menu and a summary of their attributes. Figure 3b illustrates how the choice information is presented to participants. Relative to previous years, the online application portfolio lists information on transportation and the reason the choice is in the choice set. The school brochures from previous years did not include this information.

Beside the changes in choice menu, another change under the new plan is that there is no longer walk-zone priority (Dur, Kominers, Pathak, and Sönmez, 2016). Hence, the priorities under the new plan are simply: continuing students first, followed by siblings, followed by other students. Each student receives a random number, which is the same for the student at all schools and breaks ties within each priority group.

The new plan involves large changes in the set of choices an applicant can rank. This fact can be seen in Table 1, which shows that for an average grade K1 student, the reform added about 3 new options, removed 16 options, and kept 9 options in tact. For 16% of grade K1 students who applied in the year before the reform, their top choice was no longer an option if they had applied in the first year of the reform. Conversely, for 7% of grade K1 students who applied in the first year of the reform, their top choice was not offered if they had applied in the year before the reform. The extent of choice set changes for applicants in grade K2 is similar.

3 Methodology

We start with an overview of the methodology. We use data from before the reform to estimate choice models to predict what would happen in the first year of the reform, and compare with the outcome induced by the actual choice data.[10] We focus on round 1 assignment, which takes place

[9]There are a few exceptions to this plan. First, students residing in parts of Roxbury, Mission Hill and Dorchester are allowed to rank the Jackson Mann school. Second, because transportation outside of East Boston requires tunnels travel, students in East Boston are eligible to apply to any school in East Boston. East Boston students have priority over non-East Boston students at East Boston schools. Non-East Boston students have priority over East Boston students for Non-East Boston Schools. Finally, students who are limited English proficient or special needs have additional choices. Limited english proficient students of ELD level 1, 2, and 3 are allowed to apply to any compatible ELL program within their ELL zone, which is a specially constructed six-zone overlay of Boston. Substantially separated special education students do not apply in Round 1, the focus of our investigation.

[10]The outcome induced by the actual choice data may not be identical to the actual round 1 assignment outcome, since we use previous year's program capacities in our computation rather than the actual capacities, since capacities are within the school board's control and thus not variables we need to predict in a typical application of structural modeling.

in January and February of each year and is when over 80% of the assignment takes place. We specify the choice models and the aggregate forecasts before the reform, by posting a pre-analysis plan before the data from the reform became available Pathak and Shi (2014).[11] This protects the integrity of our out-of-sample comparison.

In this section, we describe our data, the choice models, and how we measure the performance of predictions. Additional computational details are included in Appendix A.

3.1 Data sources

For the students, our main data sources are BPS round 1 choice data and enrollment data for the assignment years 2010 through 2014. The 2014 data represent the first year of the reform and is used to evaluate our forecasts. The forecasts are based only on data from 2010 through 2013.

The choice data contains submitted preference rankings of every student who participated in round 1 and demographic information on applicants. The fields included are student identification numbers; English proficiency status and first language; special education or disability status; geocode (a geographic partition of the city into 868 regions); school program to which the student has guaranteed priority (designation for continuing students); student identification numbers of the student's siblings currently enrolled in BPS; lottery number; first 10 choices and priorities to each; school program to which the student was assigned and the student's priority. Using the assigned school and program codes, we infer the capacity available for round 1 assignment for each school program. We assume that this reflects true round 1 capacities.

The enrollment data represents a snapshot of BPS students as of December and contains demographics of assigned students. This data file covers the vast majority of the students in the choice data, and contains student identification numbers, enrolled school and program, grade, geocode, address, gender, race, and languages spoken at home.

We match the choice data with the enrollment data to obtain richer demographics information for students in the choice data. We also use the geocode information to match this to 2010 census data, to obtain the median household income of the census block group in which each student resides. We classify students to one of 14 neighborhoods using the geocode[12]

For the schools, we have access to a dataset of school characteristics for each of the five years (2010-2014). The file includes for each year and each school the school code, address, school type, % of students of each race, % of students who are English Language Learners (ELL), % of students who have Special Education (SPED) requirements, and % of students who scored Advanced or Proficient in grades 3, 4 and 5 for MCAS math and English in the previous year. To measure walking distances to school, we use walking distance estimates from Google Maps API, given the address of students and schools.[13]

[11] In the original part I report, we committed to the specification of the mixed MNL model but did not manage to complete the computationally-intensive estimation procedure. We updated the report later with the forecasts for the mixed MNL model in Pathak and Shi (2015).

[12] In internal reporting, BPS classifies students into 16 neighborhoods, but we combined three neighborhoods with few students, Central Boston, Back Bay, and Fenway/Kenmore, into one neighborhood that we call "Downtown."

[13] For students with missing address information, we approximate their location using the centroid of the geocode

3.2 Choice models

The focus of our study is alternative approaches to predict families' school demand. Since we do not have sufficient data about each family's outside options to precisely estimate to the decision to apply, we focus on choices among BPS alternatives, and the approach we use to forecast the application pool is described in Section 3.3.

We consider three choice models:

- Multinomial Logit (MNL): This simple model is widely used and is easy to estimate. The Pathak and Shi (2013) report is based on this demand model.
- Mixed MNL (MMNL): This model is a popular alternative to the MNL model due to its greater flexibility in capturing complex substitution patterns that violate the Independence of Irrelevant Alternatives (IIA) property of MNL models. Mixture models are a substantial development in discrete choice demand modelling following McFadden (1974) (see e.g., Berry, Levinsohn, and Pakes (1995), Nevo (2001), Berry, Levinsohn, and Pakes (2004)).[14]
- Lexicographic: This model serves as our reference non-discrete choice alternative, which is not based on structural estimation. We specify this model based on antecedents in psychology and marketing. The model assumes that all families rank programs based on a simple rule that agrees with intuition and how the new assignment system has been presented.

We now describe the details of each model.

3.2.1 Multinomial Logit (MNL)

The MNL model is as follows. For each student i and each program j, let u_{ij} be the latent utility of student i for program j, and x_{ij} be a K-dimensional vector of characteristics corresponding to the student and the program, such as the student's distance to the program, whether the student has a sibling at the same school, whether the student is an English Language Learner (ELL) and the program is an ELL program, etc. The kth component of this vector is denoted x_{ij}^k. We have

$$u_{ij} = \delta_{s(j)} + \sum_{k=1}^{K} \beta^k x_{ij}^k + \epsilon_{ij}, \tag{1}$$

where $s(j)$ denotes the school containing the program[15], $\delta_{s(j)}$ is a school specific effect, β is a K-dimensional vector of coefficients, and ϵ_{ij} represents unobserved idiosyncratic taste and is assumed to be distributed according to a type-I extreme value distribution, $Gumbel(0, 1)$. Since utility has

listed on the application file.

[14]McFadden and Train (2000) show that any Random Utility Maximization (RUM) consistent demand model can be approximated to arbitrary accuracy by a MMNL model.

[15]Each school may have multiple programs: regular education, specialized program for English language learner, etc. Since students may later transfer between programs within a school, and since earlier tests in Pathak and Shi (2013) did not find significant program fixed effects, we include only a fixed effect for each school instead of for each program.

no scale, we normalize the scale parameter to one. The school effect captures unobserved school characteristics such as safety, reputation, facilities, environment, and teaching quality. The set of parameters to be estimated is (δ, β).

The list of characteristics in x_{ij} was selected in our pre-analysis plan, Pathak and Shi (2014), after comparing several alternatives. The finalized list includes the following features:

- distance: walking distance from the student's home to the school.
- continuing: indicator for whether the student has guaranteed status for the school program.
- sibling: indicator for whether student has sibling at the school.
- ell match: indicator for the student being an English Language Learner (ELL) and the program being specialized for ELL.
- ell language match: indicator for the student being ELL and the program having a language-specific ELL program in the student's first language.
- walk zone: indicator for whether student lives in the school's walk-zone, which is approximately a 1-mile circle around the school.[16] Students in the walk-zone do not receive school busing.

The list of characteristics also includes interaction terms between student characteristics and school characteristics. The student characteristics interacted include the student's race[17] and the medium household income of the student's census block group (estimated using 2010 census data). The school characteristics interacted include distance, as well as the following:

- mcas: the proportion of students at the school who scored Advanced or Proficient in the previous year's MCAS standardized test for math, averaging the proportions for grades 3, 4 and 5. (The MCAS test begins at grade 3. Grade 5 is the highest grade in many elementary schools. We only choose math because it is highly correlated with English, with a correlation of 0.84 in both 2012 and 2013.)
- % white/asian: the proportion of students at the school who are white or asian.

We estimate the parameters (δ, β) by maximum likelihood, using submitted rank-order preferences in previous years. The estimation of MNL models based on ranking data was first proposed in Hausman and Ruud (1987) and is now standard. To quantify uncertainty in the estimation, we

[16] The one mile circle is only approximate because the walk zones in our data are defined by drawing a one-mile circle from the school and including all geocodes that intersect the circle. A family from a geocode on the circle's boundary may be a little further than one-mile from the school, but still in the walk zone.

[17] The race data includes whether the student is Black, Hispanic, Asian, White, or Other. Based on comparing alternatives, we chose in the pre-analysis plan in Pathak and Shi (2014) to only include interaction terms that are statistically in the back-test. This means that White, Asian and Others are grouped together for all three interactions, and Black and Hispanic are grouped together for two of the interactions. The specific set of interaction terms are given in Table A1.

estimate a covariance matrix by taking the inverse of the Hessian of the log likelihood function at the maximum. The estimated parameters and standard errors are reported in Table A1.[18]

3.2.2 Mixed MNL (MMNL)

This mixed MNL model adds random coefficients to the MNL model. Let x_{ij} be a K-dimensional vector of student-program characteristics as before, and let the first L components be those for which we allow random coefficients. The model specifies the latent utility of student i for program j to be

$$u_{ij} = \delta_{s(j)} + \sum_{k=1}^{K} \beta^k x_{ij}^k + \sum_{l=1}^{L} \gamma_i^l x_{ij}^l + \epsilon_{ij},$$
$$\gamma_i \sim \mathcal{N}(0, \Sigma),$$

where δ and β are the vectors of fixed effects and coefficients in the MNL model, and γ_i is a L-dimensional vector of individual coefficients, assumed to be distributed according to a multivariate normal distribution. The mean is zero without loss of generality because it is already captured in β, and the covariance matrix Σ satisfies certain restrictions which we will specify later. The idiosyncratic term, ϵ_{ij} is assumed to be distributed $Gumbel(0, 1)$ as in the MNL model. The set of parameters to be estimated is (δ, β, Σ). The set of features x_{ij} include all the ones in the MNL model, as well as mcas and % white/asian (which are explained in Section 3.2.1).

We allow random coefficients for the following features, which we organize into "blocks," assuming independence across blocks, but allowing arbitrary covariance within each block:

Block	Features
1	ell match
2	walk zone
3	distance, mcas, % white/asian

This formulation allows students to have heterogeneous preferences of going to an ELL program (if applicable), of choosing schools in the one-mile walk zone[19], and of trading off distance, academics, and school demographics. As a result, the covariance matrix Σ satisfies the restriction

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & \Sigma_2 & 0 \\ 0 & 0 & \Sigma_3 \end{pmatrix},$$

where Σ_1, Σ_2 and Σ_3 are 1×1, 1×1 and 3×3 symmetric positive definite matrices.

Because the model no longer has closed form log-likelihood functions, and the log-likelihood functions are no longer guaranteed to be globally concave, we fit the model by Markov Chain

[18] The table only reports estimates for β, since there are too many fixed effects to tabulate.

[19] The walk zones are only approximately one mile disc, because they were originally defined using geocodes.

Monte Carlo (MCMC) methods. One technical difficulty in our situation is that we have many school fixed effects. As far as we are aware, the state-of-the-art MCMC techniques for including fixed effects in mixed-logit models are described in Train (2003). The techniques involve adding a layer of Gibbs sampling and simulating the conditional distribution of the fixed effects using the Random Walk Metropolis-Hasting algorithm. However, in our case, there are 75 schools, so this step requires simulating a 75-dimensional distribution, which is prohibitively slow using Random Walk Metropolis (RWM).[20] We therefore speed up computation by using Hamiltonian Monte Carlo (HMC) (Neal, 2011), which incorporates the gradient of the log likelihood function, so can more quickly update the 75-dimensional estimate for fixed effects. We fit the model by using 1,000,000 iterations of MCMC sampling, throwing out the first half as burn-in. To check for the convergence of the estimates, we repeated this six times with independent draws, sometimes with random starting values, and found the results to be nearly identical. Additional details are in Appendix A.

3.2.3 Lexicographic

When evaluating structural choice models, Nevo and Whinston (2010) emphasize that it is important to compare these methods to other possibilities. We therefore consider an alternative motivated by psychological/marketing literature on choice heuristics. We posit that every student ranks the programs in his menu based on the following hierarchy:

Hierarchy	Criteria
1 (most important)	(for continuing students) current program
2	(for continuing students) another program in current school
3	school where sibling attends
4	(for ELL students) ELL program
5	(for ELL students) ELL program in home language
6	better tier school
7	closer walking distance

Students only consider the hierarchy that pertain to them. For example, new applicants do not consider hierarchies 1 or 2, and non-ELL students do not consider hierarchies 4 and 5.

This choice model does not require the estimation of parameters and is therefore "non-structural." However, it is still motivated by past choice behavior and by reasonable expectations of how applicants would choose in the new plan. For instance, the vast majority of continuing students (91%) rank their current program first and we anticipated that pattern to continue under the new choice sets. Similarly, most students who have a sibling at a school rank it first, and one would expect families to continue valuing having siblings attend the same school for ease of transportation and other reasons. Furthermore, from conversation with parents and BPS staff, we learned that many people expect families to simply choose schools in the highest tier first and then break ties within

[20] See Katafygiotis and Zuev (2008) for a geometric argument for why RWM breaks down in high dimensions when the dimensions are correlated.

tier using distance. For ELL students, BPS staff stated that if given sufficient information, families would place a premium on ELL programs since they offer targeted programming, especially language-specific ELL programs in their home language. These considerations led to the hierarchy.

The lexicographic choice model is based on the psychology and marketing literature. It is related to Tversky (1969)'s lexicographic semi-order choice model in which options are rated with respect to a variety of attributes and there is a lexicographic order across attributes. Between two options, an agent first compares the most important attribute, and if there is significant difference, choose the better option according to this attribute; if there is little difference, then the agent goes to the next attribute. This encapsulates the lexicographic choice model above if we define the academic quality of schools based on discrete tiers. Another related choice model is Tversky (1972)'s elimination by aspect. In this model, the agent chooses an option from a set by going through different aspects (discrete attributes) in order of importance and eliminating options that are suboptimal with respect to that aspect. Although the original paper allowed for probabilistic choice of aspects, subsequent papers use a deterministic order of aspects: see, for example, Thorngate (1980), Johnson, Meyer, and Ghose (1989), and Payne, Bettman, and Johnson (1988). The lexicographic rule can also be seen as a special case of elimination by aspects with a deterministic ordering of aspect given by the hierarchy.

We picked this alternative to compare with our discrete choice model given the significant empirical evidence in favor of lexicographic choice models.[21] For example, Slovic (1975) conducts a laboratory experiment involving a choice between two options evaluated on two dimensions, and show that the majority of subjects chose consistently based on the more important dimension. Tversky, Sattah, and Slovic (1988) conduct other laboratory experiments and show that in cases in which a decision is framed as choosing from a set, then a lexicographic rule is often used. (If the same decision is framed in terms of varying a numerical dimension to make the decision maker indifferent between the two options, then subjects are less biased toward the more important dimension; this suggests that framing things as a choice makes lexicographic rules more likely.) In the marketing literature, Drolet and Luce (2004) show that lexicographic rules are more likely when consumers have emotional reasons to avoid making trade-offs. Yee, Dahan, Hauser, and Orlin (2007) study choices of consumers for cell phones and fit a variety of choice model to data. A lexicographic rule by aspect predicts 75% of choices, making it perform as well as other discrete choice procedures.

3.3 Predicting the Applicant Pool

An important driver of the number of unassigned students and the average access to Tier 1 or 2 schools is the number of applicants from each neighborhood. Regardless of how applicants choose schools, a large influx of new applicants from a neighborhood would drive up the number of unassigned from that neighborhood and drive down the average access to top Tier schools from that neighborhood. If we had data on all potential applicants and their non-BPS options, we might

[21] This model has also been axiomatized. Fishburn (1974) surveys the older literature. Kohli and Jedidi (2007) study when lexicographic orders can they be represented by a linear utility function. Manzini and Mariotti (2012) generalize the original Tversky (1969) model to choosing from more than two options.

include the decision to participate as part of the structural model. In the absence of such data, we still need to reflect this uncertainty and to capture any trends in the neighborhood participation patterns. McFadden, Talvitie, and Associates (1977) in Chapter IV.3 describe how similar forecasts were necessary to construct a representative sample of households in the San Francisco Bay Area and they describe a method to produce a contingency table giving the distribution of a selected set of household characteristics, given marginal tabulations available from the Census. Since we have the universe of participants in previous years, we directly observe the joint distribution of household characteristics and exploit this fact for predicting the applicant pool.

For applicant pool forecasts, we consider new and continuing students separately. Continuing students are already in the enrollment data of the previous year, while for new students we need to use previous year's applicants' demographics as proxies. For each grade and neighborhood, we examine where there are time trends in participation; in most neighborhood-grade combinations, there are not. We therefore model each of the grade-neighborhood proportions of aggregate participation as a normally distributed random variable. To estimate the mean and standard deviation, we run a regression with respect to year for each of these 28 time series, and discard all regressions for which the slope has less than 95% significance level. For the neighborhood and grade combinations in which we discard the regression, we forecast next year's proportion using the previous 4 years' sample mean and sample standard deviation. For the neighborhood and grade combinations in which the regression slope has 95% significance, we use the predicted mean and standard error of the regression.[22] The regression lines we kept are for K1 Charlestown and K2 Downtown, for which we detected a steady upward trend in the number of applicants.

We therefore model the total number of new applicants from each neighborhood in the next year as a product of two independent normals, one representing a BPS wide shock and one a neighborhood shock. The common shock captures the uncertain effect that BPS publicity or policy initiatives have on the propensity for families to apply to BPS round 1. The neighborhood shock captures local population surges or unobserved reasons that affect participation. By using one common shock for all grades, we implicitly assume that different grades exhibit the same reactions to BPS policies, and are trending in the same directions. Pathak and Shi (2014) provide additional details about how this approach performs.

For continuing students, we define the Round 1 continuing ratio for a grade and neighborhood as the proportion of relevant students from the previous year's enrollment data who decide to continue in this year's Round 1. We use the same approach to detect linear trends. However, in this data we failed to find any significant trends in continuing ratios, so we model them as normally distributed random variables, with mean and standard deviation according to the sample mean and sample standard deviation for years 2010-2013.

To simulate the pool applicant pool for next year, we independently draw the total number of new applicants, the proportion of this to allocate to each grade and neighborhood[23], and the continuation

[22]Note that the standard error has one fewer degree of freedom than the sample standard deviation.

[23]Since the total includes grade K0 but we do not estimate proportions for K0, we do not require these draws to add up to 1.

ratio for each grade and neighborhood. Fixing these realizations for one sample of choice data, we draw new applicants by sampling with replacement the previous year's new applicants, but treating them as if they applied in the new year, and we sample the given number of new applicants from each neighborhood and grade. For continuing students, we go through all potentially continuing students and decide whether to include them independently, with probability according to the generated continuation ratio for that grade and neighborhood. We use this method to generate the applicant pool for all our simulations.

3.4 Generating Predictions

To make our predictions, there are four layers of random draws:

1. Population Draw: Draw a pool of applicants according to the steps in Section 3.3. This represents uncertainty in participation rates.

2. Coefficient Draw: For the logit and mixed-logit demand models, draw the coefficients as jointly normal random variables, using the estimated means and covariance matrix. This represents uncertainty in the demand model.

3. Preference Draw: Having fixed a specific demand model and parameters, simulate for each student a complete ranking over his/her personalized menu of options, according to the randomness inherent in the demand model. Truncate this ranking to the first ten choices.

4. Lottery Draw: Generate iid lottery numbers for each student. The lottery numbers are distributed uniformly between zero and one, with lower lottery numbers being better.

After doing these steps once, we have one set of simulated choice data. With these inputs, all outcome metrics are a deterministic function computed by running the BPS assignment algorithm. The reason we truncate to first 10 choices is as follows: currently our choice data from BPS truncates to first 10 choices, although students can rank arbitrarily many. The policy report, Pathak and Shi (2013), provides evidence that assuming everyone ranks 10 choices yields reasonably accurate forecasts.

An alternative approach is to explicitly model outside options and assume that unranked programs are inferior to the outside option. However, we observe in the data that often students end up enrolling to options they did not rank but could have ranked, suggesting that this assumption is invalid. In our interactions with parents and BPS staff, it seems that many families are ranking few options not because they have better outside options, but because they feel confident they would get into the ones they picked and did not bother, or because they do not understand that ranking more options do not harm their chances to top choices.

3.5 Evaluating Predictions

3.5.1 Equilibrium Outcomes

For equilibrium outcomes (Access to quality, unassigned, or distance), let y_h be a random variable that corresponds to the outcome for neighborhood h out of a H total neighborhoods induced by the given choice model. This variable is random due to the randomness in the choice model parameters, in the population, in the idiosyncratic errors, and in the lottery numbers. Then y is a 14-dimensional vector, because there are 14 neighborhoods. The prediction for neighborhood h is

$$\bar{y}_h \equiv \mathbb{E}[y_h].$$

This quantity can be estimated by sampling y many times and taking the average. In this paper, we take many samples so for notational simplicity, we assume that the estimated mean is exactly equal to $\bar{y}$.

Let y^* denote the actual outcome vector, computed using the actual population, actual choices, and actual lottery numbers. The root mean squared error (RMSE), which is our main measure of prediction error of this choice model for this outcome, is defined as

$$\textbf{RMSE} \equiv |\bar{y} - y^*|_2 \equiv \sqrt{\sum_{h=1}^{H} (\bar{y}_h - y_h^*)^2}.$$

As a measure of the inherent uncertainty in the prediction, we define

$$\textbf{Expected RMSE} \equiv \mathbb{E}[|\bar{y} - y|_2],$$

which is also estimated by averaging across many simulations.

Finally, we estimate the 2.5^{th} and 97.5^{th} percentile of each y_h and come up with a 95% confidence interval (C.I.) for that neighborhood. Let this interval be denoted as Y_h. Another estimate of prediction accuracy is:

$$\textbf{\% in 95\% C.I.} = |h : y_h^* \in Y_h|/H.$$

3.5.2 School Choices by Neighborhood

For choice accuracy, the metrics we use are as follows. For student i, let the set of schools in the student's menu be denoted as S_i. For a given choice model, there is a particular distribution of how the student would rank options in this set, which is a distribution over permutations of the set. Let π_i be a random permutation. In other words, π_{i1} is the index of the top choice, π_{i2} the index of the second choice, and so on. Define the set of top k choices as $\Pi_{ik} \equiv \{\pi_{i1}, ..., \pi_{ik}\}$.

Let π_i^* denote the student's actual choice ranking, and r_i the actual number of choices ranked. For notational simplicity, we treat π_i^* as a permutation of S_i, although only the first r_i entries are

meaningful. Denote

$$I_k = \{i : r_i \geq k\},$$

as the set of students that ranked at least k choices.

We measure both the absolute accuracy of the choice model in predicting π_i^* and the accuracy in predicting the distribution of choices. The first evaluation is intuitively interpretable, while the second accounts for the fact that in a random utility model, the true choice distribution is random. That is, even if we had the correct distribution, we would not have perfect accuracy by the first type of metric, but we can have perfect accuracy with respect to the second. For consistency with the rest of the paper, we will formulate everything in terms of prediction error, so the smaller the number the better.

For measuring absolute error, we measure the proportion of top k choices that the choice model can predict. Define the variable

$$z_{ij}^k = P(\{j \in \Pi_{ik}\}).$$

This is the expected proportion of times the option j occurs as a top k choice according to the choice model. For each k and each student i, sort the options in the student's choice menu in decreasing order by z_{ij}^k. Define the set $\bar{\Pi}_{ik}$ as the first k options in this decreasing order. This is best deterministic prediction of the set Π_{ik}^* by the choice model. The following is a measure of absolute prediction error.

$$\textbf{Top k Choices (\% Mistakes)} = \frac{1}{|I_k|} \sum_{i \in I_k} \frac{|\bar{\Pi}_{ik} \backslash \Pi_{ik}^*|}{k}.$$

The term in the expectation is the percentage of predicted top k choices (according to the choice model) that turns out not to be a true top k choice. That is, this is the percentage of the true top k choices that we can predict based on the best deterministic prediction.

Another measure of absolute prediction error is as follows. Let $\mathbf{u_i}$ be a random utility vector that is induced by the choice model. (This is inversely sorted as the random ranking.) Define the set of comparisons

$$C_i^* = \{(j, l) : j \text{ is ranked before } l \text{ in } \pi^*\}.$$

Comparisons arise if either both j and l are ranked and j is preferred, or if j is ranked and l is unranked. Define

$$\textbf{All Pairwise Comparisons (\% Mistakes)} = \frac{1}{|I_1|} \sum_{i \in I_1} \frac{1}{|C_i^*|} \sum_{(j,l) \in C_i^*} P(u_{ij} < u_{il}).$$

This average is the probability that the choice model mistakenly ranks the less preferred option over the more preferred.

Finally, we measure the error in the distribution of choices. For a fixed k, we define the predicted

top k market share for school j as

$$s_j = \frac{1}{|I_k|} \sum_{i \in I_k} \mathbb{E}[|\text{programs in } \Pi_{ik} \text{ at school } j]/k.$$

Note that previously, j denoted a program, whereas in the above, it denotes a school. Define similarly the actual share s_j^* using Π^* instead of Π. The statistical distance, or total variation. between the actual share distribution and the predicted share is

$$TV \equiv |s^* - s|_{TV} \equiv \frac{1}{2} \sum_j |s_j^* - s_j].$$

This is the minimal amount of movement it is possible to make probability vector s into s^*.

We consider a modified version of the metric based on neighborhoods. Note that both of the above measures of market share can be redefined limiting only to students from a particular neighborhood h (instead of averaging over all students in $|I_k|$). Let the corresponding total variation be TV_h. We will report the following metric:

$$\textbf{Market Shares by Neighborhood (Statistical Distance)} = \frac{1}{H} \sum_{h=1}^{H} TV_h.$$

The final metric we examine is a version of statistical distance that focuses on the joint distribution of two top-ranked schools. For tuple of schools (j, l), define

$$p_{jl} = \frac{1}{|I_2|} \sum_{i \in I_2} P(\{\pi_{i1} \text{ is in school } j \text{ and } \pi_{i2} \text{ is in } l\}).$$

Similarly, we can define p_{jl}^* for the real choice rankings. The metric we use is

$$\textbf{Joint Distribution of Top 2 Choices (Statistical Distance)} = \frac{1}{2} \sum_{jl} |p_{jl} - p_{jl}^*|.$$

4 Backtesting and Hypotheses

To understand what we might expect about the performance of our predictions, we begin with a back-testing exercise. This exercise applies our prediction methodology on data from a year earlier: we use data from two years before the reform (2012) to predict outcomes one year before the reform (2013). Since applicant choice sets did not change between these years, we expect the results from the back-test to provide a best-case scenario for what we might expect following the large change in choice sets in 2014.

Figure 4 reports on back-testing. It shows for each neighborhood the actual access to quality in 2013, as well as the predicted access to quality using each of the three choice models and data from the previous years. For each predicted value, the figure also shows the 95% confidence interval.

Table 2 reports the root mean squared error (RMSE) for each model, computed by taking the square root of the average squared difference of the actual and predicted value for each neighborhood. We also compute the proportion of neighborhoods for which the actual access to quality falls into the predicted confidence interval.

For each grade and for each moment of interest, the MNL and MMNL models exhibit nearly-identical RMSE. Moreover, for every combination except Access to Quality in grade K1, the MNL-based models exhibit smaller RMSE than the Lexicographic model. However, even for the MNL-based models, there are still significant inaccuracies: the actual RMSE is generally larger than the expected RMSE if the model were correct, and the predicted outcome is within the predicted 95% confidence interval (C.I.) only about 70% of the time. For the Lexicographic model, the performance is worse: the actual outcome is inside the 95% C.I. less than 40% of the times on average.

The MNL and MMNL models also outperform the Lexicographic model for individual choice predictions. Table 3 shows how the logit model performance for each grade and whether we consider the top 1, top 2, and top 3 choices. As a benchmark, Table 3 also tabulates the accuracy of uniformly random guessing. As can be seen, while uniformly random guessing predicts the top choice of a student wrong 97% of the time, the Lexicographic model predicts it wrong 63% of the time, which means that it predicts the exact top choice (out of more than 30 options) over 27% of the time. For pairwise comparisons between programs and grade K1, while uniformly random guessing predicts wrongly 50% of the time (by definition since it is the average of multiple coin flips), the Lexicographic model predicts wrongly 30% of the times, while the MNL and MMNL models reduce the percentage of mistakes to 18%. A similar comparison holds for grade K2. In summary, for predicting individual choices, the MNL-based models are indistinguishable, and both outperform the Lexicographic model.

For predicting distribution of choices, Panel B of Table 3 shows that in many cases, the Lexicographic model is not much better than uniformly random guessing, and for the joint distribution of top 2 choices, it can be even worse than guessing. The Lexicographic model does not allow for the possibility that students prefer a further school of the same tier to a closer school, given that the continuing, sibling, and English learning status of the student are the same at both schools. For these metrics, the MNL-based models perform nearly identically to one another, and both outperform random guessing and the Lexicographic model.

As a result of this analysis, we formulate the following hypothesis prior to seeing any post-reform data:

- For equilibrium forecasts, the MNL-based choice models would perform similarly to one another, and both would systematically outperform the Lexicographic model. For all models, the actual prediction error would be significantly larger than the expected error if the model were completely correct.

- For choice forecasts, the same would hold as above. Moreover, the Lexicographic model would reasonably predict individual choices, but would perform badly for the distribution of choices.

5 Results

In this section, we compare the results of our prediction to the actual choices applicants submitted and the resulting equilibrium outcomes. Our main findings are as follows

- For equilibrium forecasts, our hypothesis that the MNL-based models outperform the Lexicographic model is false for grade K2. However, this is due to errors in the applicant pool forecast rather than in the choice model. MNL-based models outperform Lexicographic if we use the actual applicant pool instead of the predicted applicant pool.
- For choice forecasts, all our hypotheses are true. This is because despite the changes in the presentation of choices, the overall choice pattern of students is similar before and after the reform.

We describe the findings for equilibrium and choice forecasts next.

5.1 Equilibrium and Choice Forecasts

The underlying computations of the prediction exercise are illustrated in Figure 5, which shows for each neighborhood the actual access to quality in the first year of the reform (2014), as well as the predicted access to quality according to each of the three choice models based on pre-reform data alone. For each prediction, the figure also shows the 95% confidence interval. Based on this, we can compute for each model the root mean squared error (RMSE), by squaring the difference between the actual and the predicted value for each neighborhood, averaging across the 14 neighborhoods, and taking the square root. We can also count the percentage of neighborhoods for which the actual value falls into the predicted confidence interval.

Table 4 shows that our hypothesis that the MNL-based models outperform the Lexicographic model is true only for grade K1, but not for grade K2. For grade K1, the MNL-based models exhibit a smaller RMSE than the Lexicographic model, with a significantly higher fraction of predictions being in the 95% confidence interval. This follows the pattern observed in the back-testing exercise. However, for grade K2, the Lexicographic model exhibits similar RMSE as the MNL, with slightly better prediction accuracy for two out of the three moments of interest: access to quality and number of unassigned students. Moreover, for these two moments of interest, the percentage of neighborhoods for which the outcome is within the 95% confidence interval is also higher in the Lexicographic model, compared to the MNL-based models.[24]

The rest of our hypotheses about equilibrium outcomes are confirmed in Table 4. In all cases, the MNL and the mixed MNL (MMNL) models give near-identical results, regardless of whether we consider the RMSE or the fraction of predictions within the 95% confidence interval. The expected RMSE, which measures the level of uncertainty in the prediction, is also similar between the two

[24] This is not due to higher uncertainty in the prediction of the Lexicographic model. As seen in column 5 of Table 4, the expected RMSE is similar in the Lexicographic model as in the MNL-based model.

models. In addition, regardless of the model or the metric, the actual RMSE is higher than the predicted RMSE, which shows that none of the models is completely accurate.

For choice forecasts, the results of the prediction exercise are similar to those reported in the back-test, and thus confirm all of our hypotheses. As seen in Table 5, the MNL and MMNL models exhibit near-identical performance, regardless of the grade and the metric considered. Furthermore, the prediction error is smaller in the MNL-based models than in the Lexicographic model. The amount by which the MNL-based models outperform the Lexicographic model is also much higher for metrics for distribution of choices, rather than for individual choices. This is because as expected from the back-test, the Lexicographic model is reasonably accurate for predicting individual choices, but it is not systematically better than guessing randomly.

5.2 Decomposing Prediction Errors

When we use information on the actual set of applicants and their characteristics but not their choices, the MNL-based choice models systematically outperform the Lexicographic model. However, Table 6 shows that our hypotheses about the relative performance of the various choice models all become true when we use the actual set of applicants and their characteristics, instead of predicting these from pre-reform data by projection and by sampling. The MNL and MMNL models give near-identical performance as expected, and both outperform the Lexicographic model. For access to quality and distance to school, the MNL-based models exhibit significantly lower prediction error in both grades, contrary to the results when we do not control for the applicant pool (Table 4). For the number of students who are unassigned, all of the choice models give similar prediction accuracy, and all the RMSE are much smaller than the RMSE without controlling for the applicant pool (Table 4). Nevertheless, in all cases, the actual RMSE is still larger than the expected RMSE if the choice model were completely correct, consistent with expectations from the back-test.

Our findings suggest that the unexpected poor performance of the MNL-based models in the original prediction exercise are due to poor predictions of the applicant pool, rather than due to the choice models themselves. Table 7 reports on how prediction accuracy changes as we include information from the new dataset. In this table, we reproduce the RMSE of the MNL-based models from Table 4 (the original forecast) and the RMSE under the following assumptions:

- New applicants with Choices Predicted from Demand Model Fit with Old Applicants: Using the actual set of applicants, but choices from demand model fit with old applicants, shown in Table 6.
- New applicants with Choices Predicted from Demand Model Refit with New Applicants: Using both the actual set of applicants and also using choice models estimated from the actual choices.[25]
- New applicants with Choices Predicted from Demand Model Refit with New Applicants and

[25] The new coefficient estimates are in Table A1 and Table A2.

Ranking Length: Using both of the above, as well the actual number of choices ranked by each new applicant.

- New Choices and Sampled New Applicants: Using the predicted number of students from each neighborhood, but sampling students from the actual applicant pool and using the actual choices of these students. In predicting the number of students, we follow the sampling methodology in the original forecasts.

Comparing the prediction error from these differing sets of assumptions shows the following:

1. Estimates of the MNL-based choices models are robust across the reform. Prediction errors are similar regardless of whether we estimate the models using data from before the reform or after the reform. The RMSE in Table 7 for "New Applicants with Choices Predicted from Demand Model Fit with Old Applicants" compared to "New Applicants with Choices Predicted from Demand Model Fit with New Applicants" are very similar.

2. The assumptions on the length of rank-order lists are not of first-order importance. When we control for the actual lengths of submitted rank-order lists, the prediction error only improves for access to quality, but not for distance to school and the number of unassigned students per neighborhood. In comparison, having good predictions of the applicant pool is of first-order importance, as the RMSE improves significantly for every metric when we compare the original forecasts to the version that controls for the actual applicant pool.

3. Much of the overall error in the original forecast is due to predicting the wrong number of students from each neighborhood. This is seen in how large the prediction error is with actual choices for sampled new applicants.

The inconsistent performance of the MNL-based models in the original forecasts is largely due to errors in the predictions of the applicant pool, rather than due to the choice models themselves. (Table A3 provides additional details on the forecast of applicants and applicant demographics compared to the actual demographics.) When we control for the actual set of applicants, the MNL-based models consistently outperform the Lexicographic model. Moreover, the prediction error is similar regardless of whether we estimate the models using pre-reform data or post-reform data. These findings suggest that structural choice models can be effective in predicting counterfactual outcomes, as long as we have accurate forecasts about auxiliary variables such as applicant characteristics, which are inputs to the choice model.

The stability of the MNL-based choice models across the reform is illustrated again in Table 8, in which we compare the performance of these models when we estimate them using pre-reform data and post-reform data. As can be seen, while the prediction error generally reduces when we estimate the model from actual choices, the reduction is relatively small. This provides positive support for the use of such choice models: even if the choice sets change significantly and the presentation of options change, the choice model estimated using past data from old choice sets can be used as a close proxy to a choice model estimated from unobservable future choices.

The stability in choice patterns is not due to the absence of the effect of informational salience, but rather the magnitude of the salience effect is small compared to the overall choice pattern. Figure 6 plots the percentage of top choices that are of Tier 1 in the actual choices, compared to predicted choices from various choice models. As can be seen, while the tier salience effect is clearly present in grade K2, it only shifts top choices toward Tier 1 schools by a few percentages compared to what is predicted by the MNL-based models. On the other hand, the Lexicographic model predicts a huge shift toward Tier 1 schools, which is not supported by the actual choice data.

6 Conclusions

In this paper, we reported on a validation of a structural model of school demand, which was the basis for a major policy change affecting thousands of families in Boston in 2014. Our study design allows for a genuine out-of-sample comparison because forecasts were made prior to the policy change. This approach ensures that our validation is able to answer the question of how well our forecasts perform without the possibility of modifying forecasts after observing the realized outcomes. Since we are able to observe the choices participants made after the policy change, our approach also allows for a controlled decomposition of sources of prediction error. We believe that additional work along these lines is needed to assess the strength and limitations of structural models in other contexts.

We find that structural choice models are reasonably accurate, once we control for changes in the environment outside of the structural model relating to the number and types of participants. The multinomial logit model significantly outperforms an ad-hoc non-structural model, but its performance is comparable to the more flexible mixed multinomial logit model. The limited improvement of the mixed MNL model relative to the MNL was foreshadowed by results from our back-testing procedure and may reflect the rich micro-level data we have on individual characteristics. Judging the predictions based on expectations set from the back-test, the MNL prediction errors are similar to predictions made for years prior to the reform. These two comparisons support the use of MNL choice models for policy analysis. However, these relative performance differences hold important aspects of the environment fixed. Without judging the quality of predictions using the new set of applicants, the performance of all three forecasts is undermined by changes in the set of applicants, so much so that predictions for many grade K2 equlibrium outcomes from the ad-hoc non-structural model outperform those made from the MNL choice model. In absolute terms, holding the environment fixed, there is still substantial scope to improve the predictions of our model. An open question, therefore, is whether a more principled approach to variable selection in the choice modeling have led to further improvements. It's also not clear whether alternative non-structural approaches would have led to better predictions.

The school choice application we considered here is only one application of structural demand modeling in economics. The setting and counterfactual policy have the virtue of being relatively simple and therefore showing important possibilities for counterfactual predictions where substitu-

tion patterns across choices is central. We show that even when we focus on the simple task of predicting aggregate choice behavior, there can be significant unforeseen error due to changes in the environment that are outside of the model but influence its accuracy. We expect the unpredictability of the environment to play a major role in other applications as well, especially when the structural model becomes more complex. The sources of uncertainty in these models not only involves estimates of demand, but also uncertainty about the underlying environment. Our findings illustrate the risk of putting too much weight on the former and not on the latter in counterfactual analysis.

TODO

- Read McFadden book checking forecast; all McFadden literature
- Update refs
- Conditioning on epsilon (Peng)
- Joint test of coefficient stability
- standardize moment, oracle terminology; prediction vs. forecast; lexicographic =¿ Lexicographic
- Show applicant pool table?

6.1 Unobserved heterogeneity

Table 12

This focuses on the 395 students who ranked in K1 in 2013 but did not get assigned and are coming in K2 in 2014, plus the 128 students who ranked in K1 and got assigned and are coming in K2 in 2014 and not simply ranking their continuing option as their only choice. Purpose to show is that the gain in doing this more complicated thing is also limited.

A Computation Appendix

Unlike in the MNL model, the log likelihood function associated with the mixed MNL model is difficult to evaluate directly, involving many multi-dimensional integrals. Hence, we estimate it using Markov Chain Monte Carlo (MCMC) instead of maximum likelihood.

The basic framework to calibrate estimate MMNL models using MCMC have been established in previous works such as Train (2003), and is based on Gibbs sampling and the Metropolis-Hasting algorithm. However, the examples there do not have as many fixed coefficients as we have. (We have a fixed effect for every school and over 80 schools, and since only the relative difference between fixed effects matter for preferences, but not their absolute values, these fixed effects are likely to be highly correlated.) It is known that the simple Metropolis Hastings with random walk proposals does not perform well when estimating many dimensions (see Katafygiotis and Zuev (2008)), especially if the dimensions are correlated. We therefore modify the framework to use Metropolis-Within-Gibbs (MWG), which samples blocks of coordinates iteratively (rather all coordinates at once), and Hamiltonian Monte Carlo (HMC), which incorporates gradient information for directions to sample. We describe these methods in greater detail in Section A.2.

A.1 Specifying the Likelihood Function

The first step of applying MCMC techniques is specifying the full likelihood function of observing the data given the model parameters. An equivalent representation of the MMNL model from Section 3.2.2 is as follows. Let the vector of characteristics $x_{ij} = (x_{ijr}, x_{ijf})$, where x_{ijr} corresponds to the first L components, which represent the terms with random coefficients, and x_{ijf} the last $K - L$ components, which have fixed coefficients. Let coefficient vector $\beta = (\beta_r, \beta_f)$ similarly. The latent utilities are as follows.

$$u_{ij} = \delta_{s(j)} + \beta_f \cdot x_{ijf} + \gamma_i \cdot x_{ijr} + \epsilon_{ij}, \tag{2}$$

$$\gamma_i \sim \mathcal{N}(\beta_r, \Sigma), \tag{3}$$

$$\epsilon_{ij} \sim Gumbel(0, 1), \tag{4}$$

The set of parameters to be estimated is (δ, β, Σ). In order for the model to be well-specified, we normalize the last component of δ to be zero. Moreover, the covariance matrix Σ can be written in the block diagonal form

$$\Sigma = \begin{pmatrix} \Sigma_1 & & \\ & \Sigma_2 & \\ & & \Sigma_3 \end{pmatrix},$$

where Σ_1, Σ_2 and Σ_3 are 1×1, 1×1 and 3×3 symmetric positive definite matrices.

The data to fit these parameters are the observed choices of every students along with the observed characteristics vector x_{ij}. Suppose that student i makes m_i choices, and let the chosen programs from best to worst be $y_{i1}, y_{i2}, \cdots, y_{im_i}$.

The likelihood function can be expressed as follows. Given γ_i, the conditional likelihood is

$$\phi_i(\delta, \beta_f | \gamma_i) = \prod_{c=1}^{m_i} \frac{\exp(\delta_{s(y_{ic})} + \beta_f \cdot x_{iy_{ic}f} + \gamma_i \cdot x_{iy_{ic}r})}{\sum_{d=c}^{m_i} \exp(\delta_{s(y_{id})} + \beta_f \cdot x_{iy_{id}f} + \gamma_i \cdot x_{iy_{id}r})}. \tag{5}$$

This is the MNL likelihood function. The full likelihood function incorporating all the data is

$$\Phi(\delta, \beta_f, \beta_r, \Sigma) = \prod_{i=1}^{n} \int_{\mathbb{R}^5} \phi_i(\delta, \beta_f | \gamma_i) \exp(-\frac{1}{2}\Sigma^{-1} \|\gamma_i - \beta_r\|^2) d\gamma_i. \tag{6}$$

Here, n is the number of students; recall that the random coefficients γ_i each has five dimensions.)

Our estimates will be based on sampling the parameters based on this likelihood function Φ. Because Φ is complex, we do this by MCMC. As a detour, we will give an overview of MCMC and the specific techniques we use. Readers who are familiar with these techniques can jump to Section A.3.

A.2 Overview of the MCMC procedure

The idea behind Markov Chain Monte Carlo (MCMC) is to samples from a distribution by constructing a Markov chain whose unique stationary distribution is the desired distribution of interest. So if the chain is easy to simulate and if it is fast-mixing, meaning that it converges quickly to the stationary distribution, then we can sample by simply simulating the chain. After throwing out a so-called "burn-in" period at the beginning, we would have arrived at samples from the desired distribution.

The work horse of MCMC are Gibbs sampling and Metropolis-Hasting. Gibbs sampling is used when the desired distribution can be factored into several marginal distributions that are easier to sample. For example, to sample from a joint distribution on x, y and z, one might iteratively sample one variable at a time conditional on the other ones. Precisely speaking, we initialize x^0, y^0 and z^0 arbitrarily. For each $t \geq 1$, sample iteratively from the following conditional distributions:

$$\begin{array}{rcl} x^t & | & y^{t-1}, z^{t-1} \\ y^t & | & x^t, z^{t-1} \\ z^t & | & x^t, y^t \end{array}$$

After a sufficient number S of samples, and after throwing out the initial burn-in of B samples, $\{(x^t, y^t, z^t) : B < t \leq S\}$ would approximate samples from the original distribution, although successive samples are not independent. One can remove the serial correlation by either sampling independently from this set, or by keeping only samples in which t is a multiple of Δ, where Δ is a chosen positive integer.

Metropolis-Hasting is a technique to sample from an arbitrary distribution with given likelihood function $L(x)$. There are many variants, but the common idea is to use a proposal distribution that is easy to sample from and reject certain samples to get the likelihood ratios to be correct. The proposal distribution may depend on the current iterate x. Let transition probability density be

$T(y|x)$; this is the probability density of proposing y given that the current sample is x. In order to obtain the correct likelihoods, we can only accept a fraction of the samples proposed, and reject the others. The probability that we accept proposal y given the previous iterate being x is

$$A(y|x) = \min(1, \frac{L(y)T(x|y)}{L(x)T(y|x)}).$$

Note that if $T(y|x)$ is proportional to $L(y)$, then the acceptance probability is always 1 as the proposal distribution already matches the target. Otherwise, the above formula is tuned so that the following identity, called "detailed balance" in the literature, holds:

$$L(x)T(y|x)A(y|x) = L(y)T(x|y)A(x|y).$$

This equation guarantees that the desired density $p(x)$ is a stationary distribution of the Markov chain induced by the proposal and acceptance process. Furthermore, if the chain is ergodic, which is true for example if the proposal distribution has full support, then $p(x)$ is the only stationary distribution.

The sampling procedure is then to initialize x^0 arbitrarily, and for each $t \geq 1$

1. Draw y according to $T(y|x^{t-1})$.
2. Set $x^t = \begin{cases} y & \text{with prob. } A(y|x^{t-1}), \\ x^{t-1} & \text{otherwise.} \end{cases}$

By iterating this many times and discarding sufficiently many burn-in samples, we would have arrived at the desired distribution.

Because of the flexibility in the proposal distributions, there are many variants of the above techniques. The goal is to find a proposal distribution that strikes a good balance of being easy to sample from and approximating the target distribution locally. If it is not easy to sample from, then each step would take too long; if it is too far from the target distribution, then the acceptance probabilities would be very low and the chain may get stuck at a certain iterate for a very long time. In the following sections we present the three variants we use: Random Walk Metropolis (RWM), Metropolis-Within-Gibbs (MWG), and Hamiltonian Monte Carlo (HMC).

A.2.1 Random Walk Metropolis (RWM)

This method is the easiest to sample from, as it uses a simple random walk to propose the next value: if the current iterate is x, it proposes $y = x + \epsilon$, where ϵ is multivariate normal distributed, $\epsilon \sim \texttt{Normal}(0, \rho I)$, where I is the identity matrix and ρ is a scale parameter. Other covariance matrices can also be used instead of the identity but it must be the same for every x. The scale parameter is tuned to match the overall variance of the desired distribution. Too small a ρ and successive samples and there will be too much serial correlation; too large a ρ and acceptance probability might be near zero so the chain may get stuck. We tune ρ by multiplying it up or

down so that the average acceptance ratio since last tuning is between 0.4 and 0.6, which is the ball park value suggested by the literature.[26] The number of steps we wait before tuning increases exponentially, so that after our burn- in sample until our last iteration there is no tuning.

This method performs well when the target distribution has not too many dimensions, and has approximately the same scale in each dimension. However, when there are many dimensions, it becomes exponentially harder to guess the right direction, and the method may take very long to converge; when there are dimensions that are at very different scales, then there may exist no ρ that is good for all dimensions.

A.2.2 Metropolis Within Gibbs (MWG)

This is a simple extension of RWM that allows various sub-blocks of coordinates to have different scales. It is simply to sample each sub-block iteratively, conditional on the others, much like running several RWM within a Gibbs sampling framework. This also reduces the number of dimensions sampled at each step. The draw back is that more samples are needed.

Precisely speaking, instead of sampling all dimensions of vector x simultaneously, write it in terms of sub-vectors $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix}$. Each sub-vector may represent several coordinates. Initialize x^0 arbitrarily and for $t \geq 1$, sample

$$\begin{array}{lcl} x_1^t & | & x_2^{t-1}, \cdots x_k^{t-1} \\ x_2^t & | & x_1^t, x_3^{t-1}, \cdots x_k^{t-1} \\ & \cdots & \\ x_k^t & | & x_1^t, \cdots x_{k-1}^t \end{array}$$

Each of the above is sampled using RWM, perhaps with different scale parameters for different sub-vectors. In each Gibbs iteration, for each of the variables, we only take one step of Metropolis-Hasting, which involves one proposal and possible acceptance. Because of detailed balance, embedding Metropolis-Hasting into Gibbs sampling in this way also works.

A.2.3 Hamiltonian Monte Carlo (HMC)

This method uses the gradient of the log likelihood function to inform the proposals, which can significantly improve the acceptance probabilities in high dimensions. The drawback is that each iteration is slower as several gradient calls is needed. The method is motivated by Hamiltonian dynamics in physics. It models the current iterate x as a location vector, and treats the negative log likelihood function as an energy potential. In each step, it samples a random momentum vector and simulates the trajectory of the object by discretizing time and alternatively updating the momentum using the potential function and updating the position using the momentum. To

[26] See Roberts, Gelman, and Gilks (1997).

make detailed balance work out, the first and last steps of simulation are half-steps. Precisely speaking, let the gradient of the log likelihood function be $G(x) = \nabla(\log(L(x)))$. Let ϵ and Δ be tuning parameters, representing the discretization in time and the number of steps to simulate respectively. The proposal is based on the pseudocode in this algorithm (this is taken from Neal (2011)):

Algorithm 1 Pseudocode for one step of HMC

Function HMC_STEP(x):
Draw momentum $p_0 \sim \texttt{Normal}(0, I)$.
Initialize $y = x, p = p_0$.
Update $p = p - \epsilon G(y)/2$.
for $\Delta - 1$ iterations **do**
  Update $y = y + \epsilon p$
  Update $p = p - \epsilon G(y)$.
end for
Update $y = y + \epsilon p$.
Update $p = p - \epsilon G(y)/2$.
return $\begin{cases} y & \text{with prob. } A(y|x) = \min(1, \frac{L(y)}{L(x)} \exp(\frac{\|p_0\|^2 - \|p\|^2}{2}) \\ x & \text{otherwise} \end{cases}$

Note that the chance of proposing y given x is simply the chance of drawing momentum p_0. Moreover, by the reversibility of the intermediate steps of discrete simulation, if we started at y and drew a momentum of $-p$ (where p is the final momentum vector in HMC_STEP), then the proposal would be x. This implies that

$$\frac{T(y|x)}{T(x|y)} = \frac{\exp(-\frac{1}{2}\|p_0\|^2)}{\exp(-\frac{1}{2}\|-p\|^2)},$$

which implies that

$$\frac{T(y|x)A(y|x)}{T(x|y)A(x|y)} = \frac{\exp(-\frac{1}{2}\|p_0\|^2)}{\exp(-\frac{1}{2}\|-p\|^2)} \frac{L(y)}{L(x)} \exp(\frac{\|p_0\|^2 - \|p\|^2}{2}) = \frac{L(y)}{L(x)}.$$

So detailed balance holds and the following is a valid Metropolis-Hasting sampler: Initialize x^0 arbitrarily. For $t \geq 1$, set $x^t = \text{HMC_STEP}(x^{t-1})$.

One can show that as the time discretization $\epsilon \to 0$, for any fixed total simulation time $\epsilon\Delta$, the acceptance probability goes to 1. Hence, we would like ϵ to be small enough so the chain does not get stuck and $\epsilon\Delta$ large enough so that successive samples are not too serially correlated. In practice, we fix $\Delta = 20$ and tune ρ so that the empirical acceptance rate since last tuning is between 0.5 and 0.8. As before, we increase the interval between tuning times exponentially so that no tuning happens in the sample we keep (after burn-in and before the last iteration). Another detail is that to prevent cases in which $\epsilon\Delta$ is exactly what makes the proposal y go back to original point x, instead of using the same ϵ, we draw $\tilde{\epsilon} \sim \texttt{Uniform}(0.85\epsilon, 1.15\epsilon)$ before each call to HMC_STEP, and use $\tilde{\epsilon}$ as the step size throughout that call. Because this distribution is a-priori fixed, detailed balance is

also preserved. All these are according to the best practices for applying HMC as outlined in Neal (2011).

A.3 The MCMC Sampler

Our MCMC procedure is based on the one in Train (2003) but breaking up the estimation of the fixed coefficients into two steps, one step using Hamiltonian Monte Carlo (HMC) and the other Metropolis Within Gibbs (MWG). We use HMC to estimate the school fixed effects and MWG to estimate the other fixed coefficients. These techniques allow us to accommodate the large number of school fixed effects and the unequal scales across the other fixed coefficients.

To sample from the full likelihood function $\Phi(\delta, \beta_f, \beta_r, \Sigma)$ (Equation 6), we initialize $\delta^0, \beta_f^0, \beta_r^0, \Sigma_1^0, \Sigma_2^0, \Sigma_3^0$ arbitrarily. For each $t \geq 1$, we do a few layers of Gibbs sampling. In some of the layers we embed a form of Metropolis-Hasting; but in each Gibbs iteration we only take one step of Metropolis-Hasting, much as it is in MWG. Furthermore, let T be a parameter indicating how long we wait before tuning. We initialize T to be 1 and increase this parameter steadily, so that tuning becomes exponentially less frequent. For $t \geq 1$, each MCMC step is as follows:

1. Draw $\gamma_i^t | \delta^{t-1}, \beta_f^{t-1}, \beta_r^{t-1}, \Sigma^{t-1}$. This is done using one iteration of RWM with likelihood function
$$L(x) = \phi_i(\delta^{t-1}, \beta_f^{t-1}, x) \exp(-\frac{1}{2}(\Sigma^{t-1})^{-1} \|x - \beta_r^{t-1}\|^2)$$
and starting value γ_i^{t-1}. (See Equation 5 for definition of ϕ_i.) We initialize $\rho = 0.05$ and initially to tune for each i every $\texttt{Uniform}(1000T, 1500T)$ steps.

2. Draw $\beta_r^t | \gamma_i^t, \Sigma^{t-1}$. This is sampling from $\texttt{Normal}(\frac{1}{n}\sum_{i=1}^n \gamma_i^t, \frac{1}{m}\Sigma^{t-1})$.

3. Draw $\Sigma^t | \gamma_i^t, \beta_r^t$. This can be done as follows: For $l \in \{1, 2, 3\}$, let $\mathbf{C}_l^t$ be the covariance matrix of the lth block of γ_i^t assuming mean as in the lth block of β_r^t. (Recall that the random coefficients are organized into 3 blocks, with ell match being the first block, walk zone being the second, and distance, mcas, and % white/asian being the third.) Let k_l be the number of variables in the lth block and let n be the number of students. Draw Σ_l^t according to the Inverse Wishart Distribution with degree of freedom $\nu = k_l + n$ and scale matrix $\Psi = k_l I_{l \times l} + n\mathbf{C}_l^t$.

4. Draw $\delta^t | \gamma_i^t, \beta_f^{t-1}$. This is done using one step of HMC with likelihood function
$$L(x) = \prod_{i=1}^n \phi_i(x, \beta_f^{t-1} | \gamma_i^t),$$
and constraining the last component to be zero. We initialize $\epsilon = 0.015$, and $\Delta = 20$. We tune every $1000T$ steps.

5. Draw $\beta_f^t|\gamma_i^t, \delta^t$. This is done using one iteration of MWG with likelihood function

$$L(x) = \prod_{i=1}^{n} \phi_i(\delta^t, x|\gamma_i^t).$$

We break the fixed coefficients β_f into 6 subvectors: 1) "continuing;" 2) "sibling;" 3) "ell language match;" 4) "distance*black/hispanic" and "distance*income est."; 5) "mcas*black" and "mcas*income est."; 6) "% white/asian*black/hispanic" and "% white/asian*income est." We initialize the scales ρ for each subvector to be .5, .5, .1, .1, .5, and .5 respectively. We tune every $\texttt{Uniform}(100T, 150T)$ steps.

We run these steps 1,000,000 times, increasing the tuning interval parameter T by a factor of 1.2 every 5000 iterations. We throw out the first 500,000 iterations as burn-in. Note that in the interval we keep, no tuning happens. This ensures the correctness of the Markov chain in this period.

For a robustness check, we re-ran this procedure 6 times, sometimes with different initial values, and we found near identical results each time.

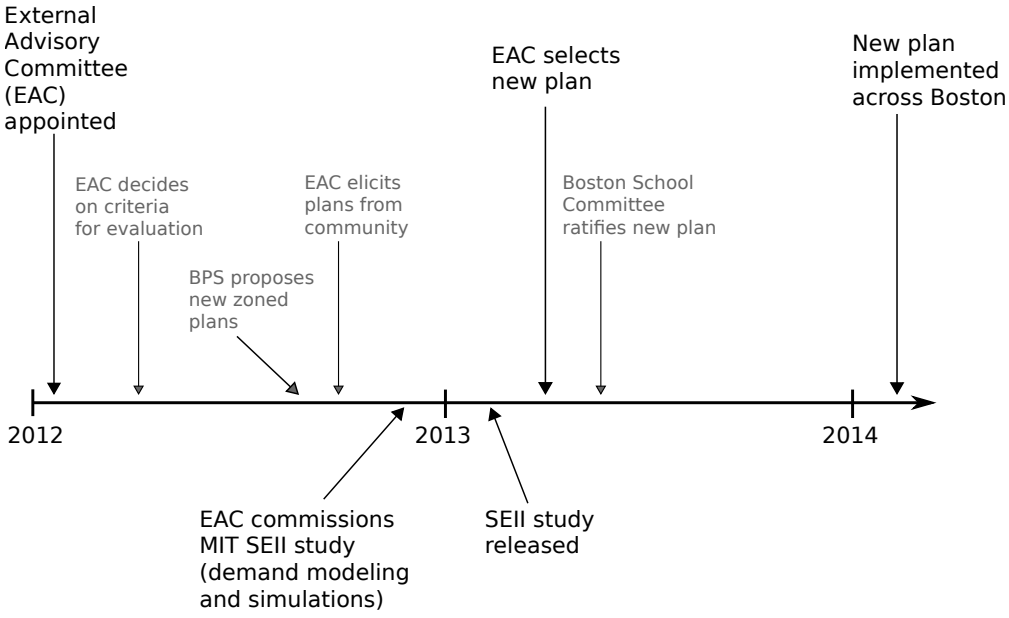


FIGURE 1: Timeline of Policy Reform

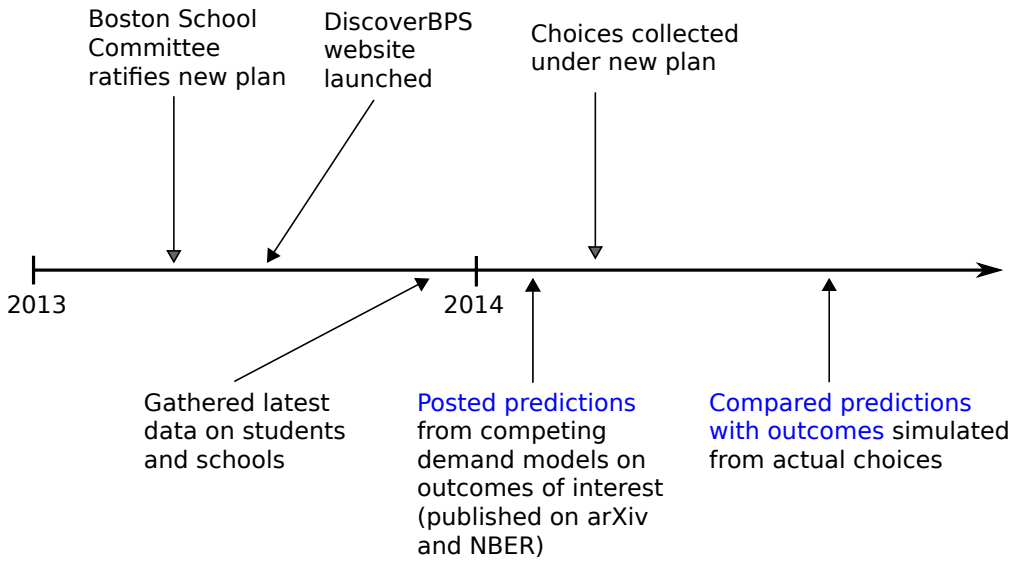
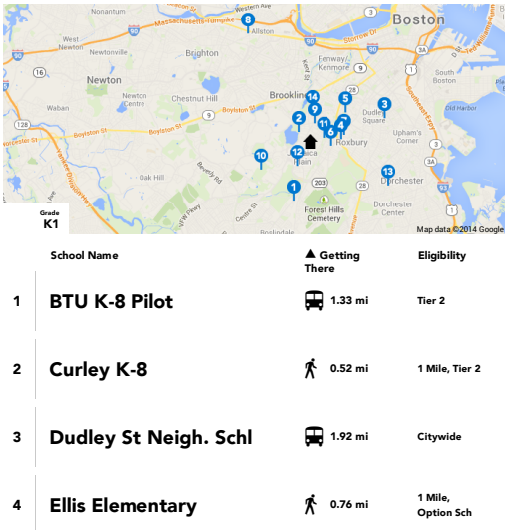


Figure 2: Overview of Research Design



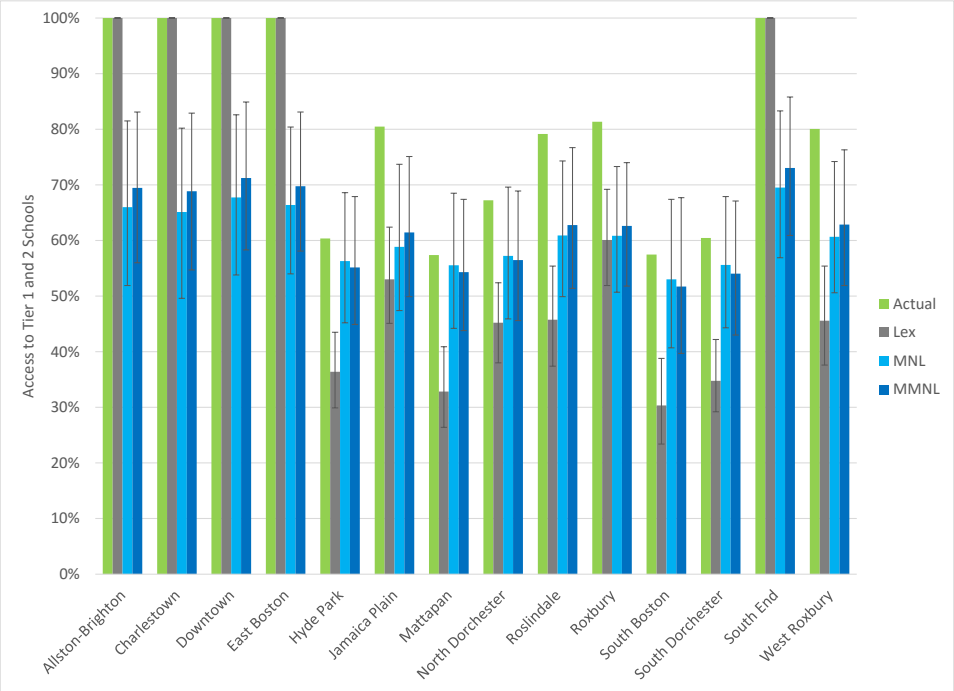
(A) Before reform (in 2013)



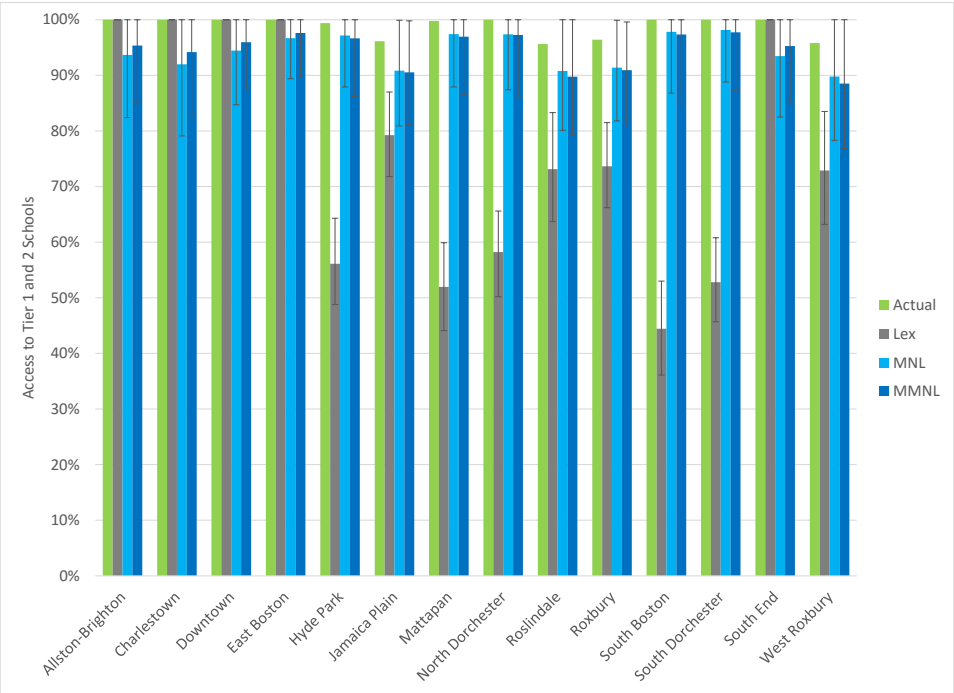
(B) After reform (in 2014)

FIGURE 3: Illustration of the Change in Choice Sets

Panel (a) shows the geographic zones under the Three Zone plan in 2013. The choice set include all schools in a student's zone with a few exceptions for city-wide schools and for students residing on zone boundaries. Panel (b) shows a computer-generated list of choices under the Home-Based Plan in 2014 for a given address. The set of choices is generated for each student based on her address and the school's tier.



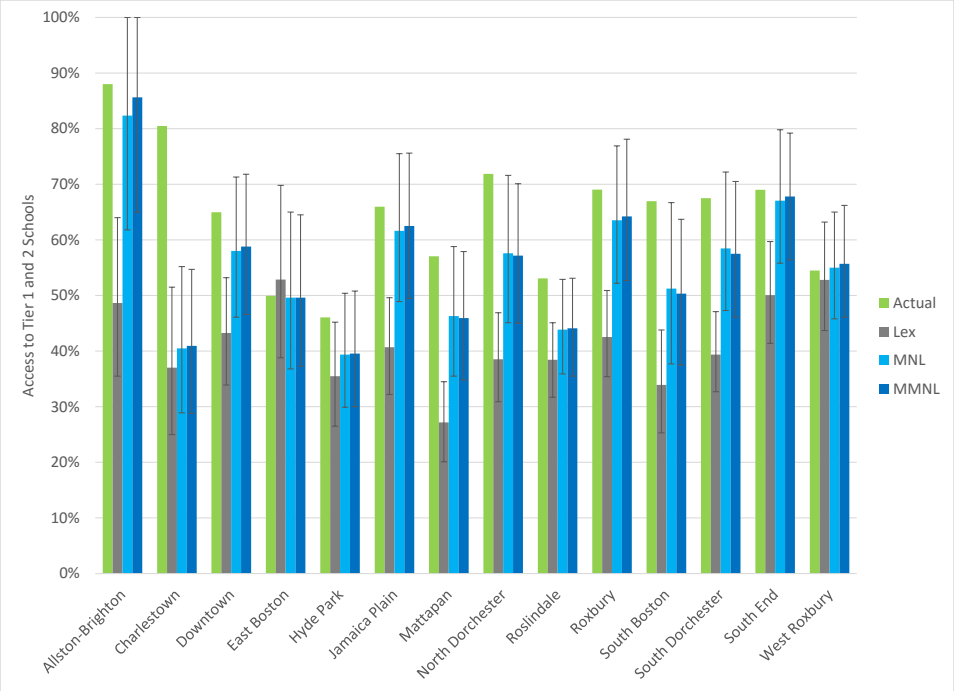
(A) Grade K1



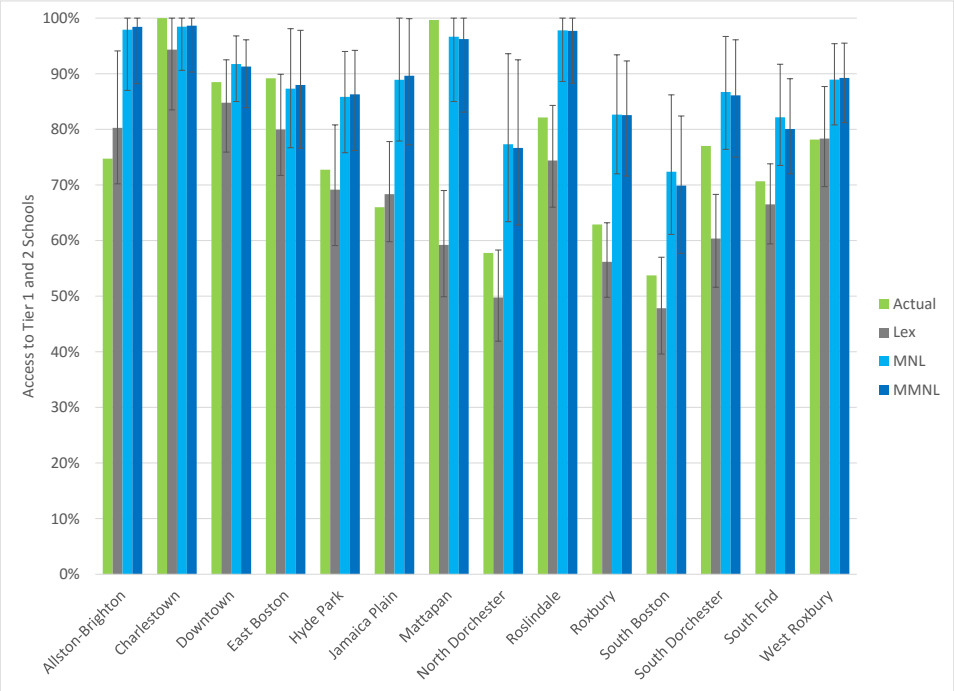
(B) Grade K2

FIGURE 4: Backtesting Access to Quality

Access to quality is the chance a student has of being assigned a tier 1 or tier 2 school if it is ranked. For each grade, this figure shows the predicted access to quality averaged across each neighborhood for each demand model. The choice models are estimated using choices two years prior to the reform (2012). Access to quality is computed using choices one year prior to the reform (2013). Whisker bars represent 95% confidence intervals.



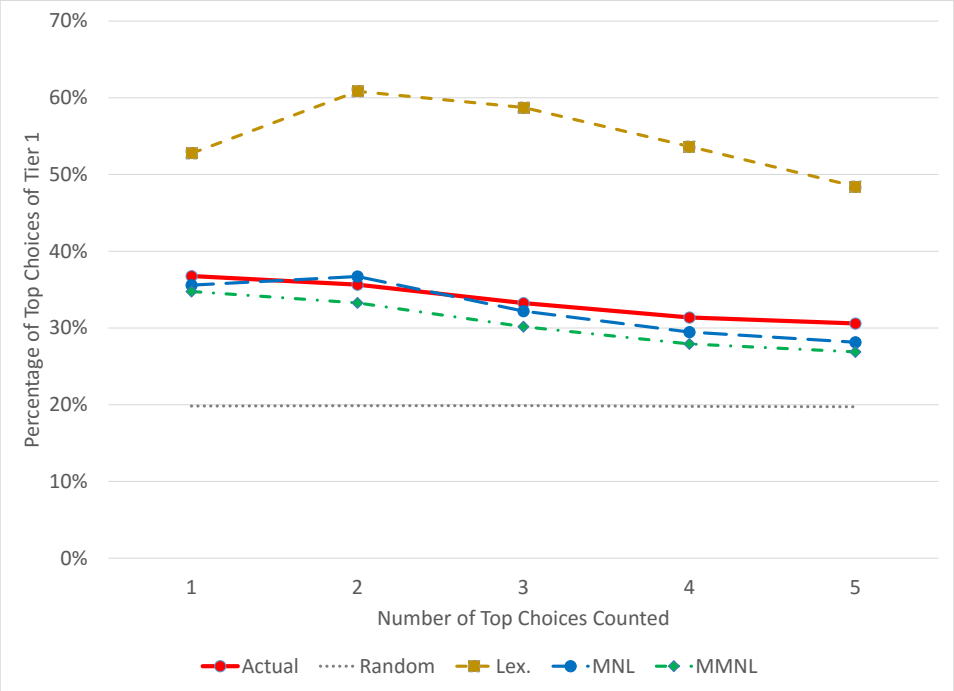
(A) Grade K1



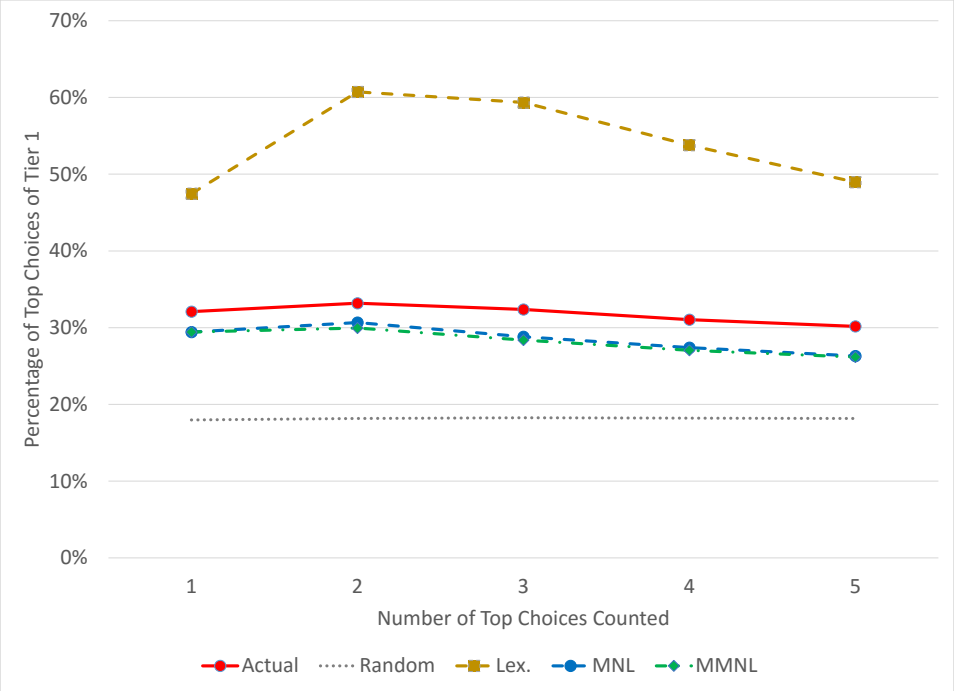
(B) Grade K2

FIGURE 5: Predicted vs. Actual Access to Quality

Access to quality is the chance a student has of being assigned a tier 1 or tier 2 school if it is ranked. For each grade, this figure shows the predicted access to quality averaged across each neighborhood for each demand model. The demand models are estimated using choices from the year prior to the reform (2013). Actual access to quality is computed using choices from the first year of the new assignment plan (2014). Whisker bars represent 95% confidence intervals.



(A) Grade K1



(B) Grade K2

FIGURE 6: Predicted vs. Actual Fraction of Top Choices Ranking Tier 1 Schools

This figure shows the percentage of top choices that are tier 1 schools for different demand models estimated using data from the old assignment plan (in 2013) compared to the percentage from actual choices in the new assignment plan (in 2014).

Table 1. Comparison of Choice Sets

	Grade K1	Grade K2
	(1)	(2)
	A: Applicants in New Assignment Plan	
Schools in New and Old Choice Set	9.4	12.6
Schools Added to New Choice Set	2.5	2.9
Schools Removed from Old Choice Set	16.1	18.4
Applicants for whom Top k Choices in New Choice Set are in Old Choice Set		
Top 1	93%	95%
Top 3	91%	92%
Top 5	90%	91%
	B: Applicants in Old Assignment Plan	
Applicants for whom Top k Choices in Old Choice Set are in New Choice Set		
Top 1	84%	79%
Top 3	76%	73%
Top 5	68%	68%

Notes: This table compares choice sets between the old three-zone plan (in 2013) and the new home-based assignment plan (in 2014). Schools in New and Old Choice Set is the average number of choices an applicant in the new plan has in the new choice set that are available in the old choice set. Schools Added to New Choice Set is the average number of choices an applicant in the new plan has in the new choice set, but not in the old choice set. Schools Removed from Old Choice Set is the average number of choices an applicant in the new plan would have in the old choice set, but not in the new choice set. Applicants for whom Top k Choices in New Choice Set are in Old Choice Set reports whether highly ranked choices in the new choice set are available in the old choice set. For each student in the new plan, we compute the percentage of their top k choices that are still available under the old choice set for k=1, 3, and 5. Applicants for whom Top k Choices in Old Choice Set are in New Choice Set reports whether highly ranked choices in the old choice set are available options in the new choice set. For each student in the old plan, we compute the percentage of top k choices that are still available options under the new plan, for k=1, 3, and 5.

Table 2. Backtesting Equilibrium Predictions Using Data from Two Years Prior

	Grade K1			Grade K2		
	RMSE	Exp. RMSE	% in 95% C.I.	RMSE	Exp. RMSE	% in 95% C.I.
	(1)	(2)	(3)	(4)	(5)	(6)
			A: Access to Quality			
Lexicographic	22%	(3%)	36%	31%	(3%)	36%
MNL	23%	(6%)	36%	5%	(4%)	100%
MMNL	20%	(6%)	36%	5%	(4%)	100%
			B: Distance (miles)			
Lexicographic	0.72	(0.25)	36%	0.46	(0.11)	29%
MNL	0.44	(0.18)	64%	0.26	(0.10)	64%
MMNL	0.40	(0.18)	64%	0.26	(0.10)	79%
			C: Unassigned			
Lexicographic	51	(17)	14%	26	(13)	57%
MNL	19	(17)	93%	19	(10)	71%
MMNL	18	(17)	93%	18	(11)	79%

Notes: This table reports backtesting results for equilibrium outcomes for the last year of the old assignment plan (in 2013) using choice models estimated from data two years prior (in 2012). In both years, we use the choice set from the old assignment plan (in 2013). Access to Quality in percentages is defined as the chance each student has of being assigned a tier 1 or tier 2 school, distance in miles is the Google-Maps walk distance between each student and their assigned school (conditional on being assigned), and Unassigned is the number of students unassigned in each neighborhood. Lexicographic, multinomial logit (MNL), and mixed MNL (MMNL) are the three choice models. For each grade, each outcome of interest, each choice model, and each of the 14 neighborhoods, we compute the prediction error, defined as the squared difference between the predicted outcome for this neighborhood (based on the demand model) with the actual outcome (based on the actual choices). Columns 1 and 4 report the root mean squared error (RMSE), defined as the square root of the average of the squared differences across the 14 neighborhoods. Columns 2 and 5 reports how large a RMSE we should expect from a random sample if the model were correct. Exp. RMSE is computed by simulating each outcome 1,000 times given a choice model, accounting for uncertainty from sampling students, coefficient estimates, and lottery numbers. Each sample of the simulation is a vector of 14 dimensions, one for each neighborhood. We compute the RMSE of each sample with respect to the sample mean (the Euclidean distance between the two 14-dimensional vectors) and report the average of these RMSEs across the 1,000 samples. Columns 3 and 6 present another metric of how unexpected the RMSE is if the model were completely correct. % in 95% C.I. is the percentage of neighborhoods for which the actual outcome lies within the predicted 95% confidence interval of the outcome. The confidence interval is estimated from the 2.5 and 97.5 percentiles of 1,000 simulations of each choice model.

Table 3. Backtesting Choice Predictions Using t-2 to Predict t-1

	Grade K1				Grade K2			
	Random	Lexicographic	MNL	MMNL	Random	Lexicographic	MNL	MMNL
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	A: Individual Choices (% Mistakes)							
Top Choice	97%	63%	58%	58%	97%	37%	35%	34%
Top 2 Choices	94%	70%	59%	58%	94%	67%	57%	57%
Top 3 Choices	90%	69%	54%	54%	90%	67%	54%	54%
All Pairwise Comparisons	50%	30%	18%	18%	50%	16%	10%	10%
	B: Distribution of Choices (Statistical Distance)							
Market Shares by Neighborhood								
Top Choice	56%	46%	22%	21%	52%	26%	16%	15%
Top 2 Choices	52%	48%	19%	18%	52%	48%	20%	19%
Top 3 Choices	49%	49%	16%	16%	48%	51%	17%	16%
Joint Distribution of Top 2 Choices	64%	76%	45%	41%	67%	79%	51%	47%

Notes: This table reports backtesting results for choices for the last year of the old assignment plan (in 2013) using choice models estimated from data two years prior (in 2012). In both years, we use the choice set from the old assignment plan (in 2013). Panel A reports on individual choices of students and panel B reports on the distribution of choices of students, averaged across 14 Boston neighborhoods. Each column corresponds to a choice model: Random in columns 1 and 5 denotes uniformly random choices, Lexicographic in columns 2 and 6 denotes the lexicographic model, MNL in columns 3 and 7 denotes the multinomial logit model, and MMNL in columns 4 and 8 denotes the mixed MNL model. % Mistakes in Panel A uses each demand model's best guess of the student's choice and reports the fraction of incorrect guesses. Top Choice is for the first choice. Top 2 Choices is for the unordered set of first and second choice, and we report the percentage of elements in this set that are wrongly predicted and average over students who ranked at least two options. Top 3 Choices reports the analog for the unordered set of first, second and third choice, averaging over students who ranked at least three choices. Pairwise is the set of pairwise comparisons of options implied by the student's actual ranking compared to the best guess of each comparison from each choice model. Statistical distance in Panel B is the total variation distance between the predicted distribution of choices and the actual distribution for neighborhood-level market shares. The first three rows report on the joint distribution of the neighborhood-level market share of top choices following Panel B, averaged across neighborhoods. Joint distribution of top 2 choices aggregates all students and compares the predicted joint probability distribution of the first and second choice of students who ranked at least two choices and the actual distribution.

Table 4. Accuracy of Equilibrium Predictions Compared to Actual Outcomes

	Grade K1			Grade K2		
	RMSE	Exp. RMSE	% in 95% C.I.	RMSE	Exp. RMSE	% in 95% C.I.
	(1)	(2)	(3)	(4)	(5)	(6)
			A: Access to Quality			
Lexicographic	26%	(5%)	14%	13%	(4%)	86%
MNL	13%	(6%)	71%	15%	(5%)	36%
MMNL	13%	(6%)	79%	14%	(5%)	36%
			B: Distance			
Lexicographic	0.34	(0.14)	50%	0.14	(0.09)	71%
MNL	0.19	(0.12)	57%	0.13	(0.07)	71%
MMNL	0.19	(0.12)	57%	0.14	(0.07)	71%
			C: Unassigned			
Lexicographic	30	(16)	57%	34	(9)	43%
MNL	22	(16)	86%	41	(7)	14%
MMNL	21	(17)	86%	40	(8)	14%

Notes. This table reports the accuracy of predictions under three choice models for equilibrium outcomes using data from 2013 (the last year of the old assignment plan) compared to data from 2014 (the first year of the new assignment plan). For each grade, each outcome of interest, each choice model, and each of the 14 neighborhoods, we compute the prediction error as the squared difference between the predicted outcome for this neighborhood (based on the demand model) with the actual outcome (based on the actual choices). Table 2 notes contain definitions of the prediction targets.

Table 5. Accuracy of Choice Predictions Compared to Actual Choices

	Grade K1				Grade K2			
	Random	Lexicographic	MNL	MMNL	Random	Lexicographic	MNL	MMNL
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	A: Individual Choices (% Mistakes)							
Top Choice	93%	59%	54%	53%	93%	33%	32%	33%
Top 2 Choices	85%	62%	51%	51%	87%	60%	54%	55%
Top 3 Choices	78%	58%	47%	47%	80%	56%	50%	51%
All Pairwise Comparisons	50%	28%	23%	23%	50%	14%	12%	13%
	B: Distribution of Choices (Statistical Distance)							
Market Shares by Neighborhood								
Top Choice	47%	41%	20%	21%	45%	22%	15%	15%
Top 2 Choices	41%	43%	16%	16%	43%	48%	19%	20%
Top 3 Choices	37%	41%	13%	14%	38%	44%	15%	17%
Joint Distribution of Top 2 Choices	62%	72%	41%	41%	67%	75%	49%	48%

Notes: This table reports on the accuracy of choice predictions using data from 2013 (the last year of the old assignment plan) compared to data from 2014 (the first year of the new assignment plan). Table 3 notes define prediction targets.

Table 6. Accuracy of Equilibrium Predictions Using New Applicants with Estimated Choices

	Grade K1			Grade K2		
	RMSE	Exp. RMSE	% in 95% C.I.	RMSE	Exp. RMSE	% in 95% C.I.
	(1)	(2)	(3)	(4)	(5)	(6)
			A: Access to Quality			
Lexicographic	22%	(2%)	7%	22%	(2%)	0%
MNL	5%	(3%)	64%	12%	(3%)	79%
MMNL	6%	(3%)	64%	12%	(3%)	79%
			B: Distance			
Lexicographic	0.21	(0.07)	43%	0.15	(0.03)	14%
MNL	0.15	(0.08)	57%	0.08	(0.04)	57%
MMNL	0.15	(0.08)	50%	0.09	(0.05)	71%
			C: Unassigned			
Lexicographic	8	(4)	50%	15	(3)	21%
MNL	7	(4)	79%	14	(4)	36%
MMNL	7	(4)	64%	15	(4)	43%

Notes: This table reports the accuracy of predictions under three choice models for equilibrium outcomes using data from 2013 (the last year of the old assignment plan) compared to data from 2014 (the first year of the new assignment plan), with the actual set of applicants. Unlike Table 4, which randomly samples the applicant pool using past data, the calculation here uses the actual set of number of applicants and their characteristics. Choices are generated from demand model estimates fit from old data. Table 2 notes contain definitions of the prediction targets.

Table 7. Prediction Improvements Using Data from New Assignment Plan

	Grade K1		Grade K2	
	MNL	MMNL	MNL	MMNL
	(1)	(2)	(3)	(4)
		A: Access to Quality		
Original Prediction	13%	13%	15%	14%
New Applicants with				
Choices Predicted from Demand Model Fit with Old Applicants	5%	6%	12%	12%
Choices Predicted from Demand Model Refit with New Applicants	7%	7%	13%	13%
Choices Predicted from Demand Model Refit with New Applicants + Ranking Length	3%	3%	7%	7%
New Choices and Sampled New Applicants	8%		10%	
		B: Distance		
Original Prediction	0.19	0.19	0.13	0.14
New Applicants with				
Choices Predicted from Demand Model Fit with Old Applicants	0.15	0.15	0.08	0.09
Choices Predicted from Demand Model Refit with New Applicants	0.16	0.15	0.07	0.07
Choices Predicted from Demand Model Refit with New Applicants + Ranking Length	0.18	0.18	0.09	0.10
New Choices and Sampled New Applicants	0.08		0.07	
		C: Unassigned		
Original Prediction	22	21	41	40
New Applicants with				
Choices Predicted from Demand Model Fit with Old Applicants	7	7	14	15
Choices Predicted from Demand Model Refit with New Applicants	7	7	14	14
Choices Predicted from Demand Model Refit with New Applicants + Ranking Length	6	6	10	10
New Choices and Sampled New Applicants	22		15	

Notes: This table compares the accuracy of predictions from Table 4 using additional information from the new assignment plan. Each cell entry is the RMSE of the prediction error. Table 2 notes contain definitions of the prediction targets. Original Prediction is reproduced from columns 1 and 4 of Table 4. New Applicants with Choices Predicted from Demand Model Fit with Old Applicants uses new applicants in 2014 and their characteristics, predicted choices from the demand model fit in 2013, following columns 1 and 5 of Table 5. New Applicants with Choices Predicted from Demand Model Refit with New Applicants uses new applicants in 2014 and predicted choices from demand model refit in 2-14. New Applicants with Choices Predicted from Demand Model Refit with New Applicants + Ranking Length uses new applicants in 2014, predicted choices from demand model refit in 2014, and and the actual number of choices ranked by each new applicant in 2014. New Choices and Samples New Applicants uses actual choices (not demand model predicted choices). To do this, we consider continuing and non-continuing students separately. Continuing students are already registered in BPS in a lower grade. Non-continuing students are new to the system. We predict the set of continuing students using the same methodology as in the original prediction and assume each chooses their previous choice. For non-continuing students, we use the same methodology as in the original prediction and sample actual applicants in 2014 with replacement.

Table 8. Accuracy of Choice Predictions from Refit Demand Models

	Demand Model Fit Using Data	Grade K1		Grade K2	
		MNL	MMNL	MNL	MMNL
		(1)	(2)	(3)	(4)
		A: Individual Choices (% Mistakes)			
Top Choice	Old	54%	53%	32%	33%
	New	49%	49%	31%	30%
Top 2 Choices	Old	51%	51%	54%	55%
	New	50%	49%	51%	51%
Top 3 Choices	Old	47%	47%	50%	51%
	New	45%	45%	48%	48%
All Pairwise Comparisons	Old	23%	23%	12%	13%
	New	21%	21%	11%	11%
		B: Distribution of Choices (Statistical Distance)			
Market Shares by Neighborhood					
Top Choice	Old	20%	21%	15%	15%
	New	18%	17%	12%	11%
Top 2 Choices	Old	16%	16%	19%	20%
	New	14%	13%	15%	14%
Top 3 Choices	Old	13%	14%	15%	17%
	New	11%	10%	11%	11%
Joint Distribution of Top 2 Choices	Old	41%	41%	49%	48%
	New	39%	37%	45%	43%

Notes: This table compares the accuracy of choice predictions from choice models fitted using 2013 data (the last year of the old assignment plan) and choice models fitted using 2014 data (the first year of the new assignment plan). Accuracy is evaluated compared to the actual choices of students in 2014. Table format follows Table 7, except we include an additional row for each outcome specifying the source year for the data used to fit the demand model. We consider only the multinomial-logit (MNL) model (columns 1 and 3), and the mixed MNL (MMNL) model (columns 2 and 4). Table 3 notes contain definitions of the prediction targets.

References

Abdulkadiroğlu, A., N. Agarwal, and P. Pathak (2013): "The Welfare Effects of Coordinated School Assignment: Evidence from the NYC High School Match," Working paper, MIT.

Abdulkadiroğlu, A., P. A. Pathak, A. E. Roth, and T. Sönmez (2005): "The Boston Public School Match," American Economic Review, Papers and Proceedings, 95, 368–371.

Abdulkadiroğlu, A., P. A. Pathak, J. Schellenberg, and C. Walters (2017): "Do Parents Value School Effectiveness?," Working paper, MIT.

Abdulkadiroğlu, A., and T. Sönmez (2003): "School Choice: A Mechanism Design Approach," American Economic Review, 93, 729–747.

Agarwal, N., and P. Somaini (2014): "Demand Analysis Using Strategic Reports: An Application to a School Choice Mechanism," NBER Working Paper 20775.

Angrist, J., and J.-S. Pischke (2010): "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics," Journal of Economic Perspectives, 24(2), 3–30.

Ashenfelter, O., and D. Hosken (2008): "The Effects of Mergers on Consumers Prices: Evidence from Five Selected Case Studies," NBER Working Paper 13589.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," Econometrica, 63(4), 841–890.

——— (2004): "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market," Journal of Political Economy, 112(1), 68–105.

Burge, K. (2012): "Study Finds Inequalities in Schools' Zone Plans," Boston Globe, October 1.

Burgess, S., E. Greaves, A. Vignoles, and D. Wilson (2015): "What Parents Want: School Preferences and School Choice," Economic Journal, 125(587), 1262–1289.

Calsamiglia, C., C. Fu, and M. Guell (2017): "Structural Estimation of a Model of School Choices: the Boston Mechanism vs. Its Alternatives," Working paper, CEMFI.

Drolet, and Luce (2004): "The Rationalizing Effects of Cognitive Load on Response to Emotional Tradeoff Difficulty," Journal of Consumer Research, 31(1), 63–77.

Dubins, L. E., and D. A. Freedman (1981): "Machiavelli and the Gale-Shapley algorithm," American Mathematical Monthly, 88, 485–494.

Dur, U., S. D. Kominers, P. A. Pathak, and T. Sönmez (2016): "Reserve Design: Unintended Consequences and the Demise of Walk Zones in Boston," forthcoming, *Journal of Political Economy*.

Einav, L., and J. Levin (2010): "Empirical Industrial Organization: A Progress Report," Journal of Economic Perspectives, 24(2), 145–162.

Fishburn, P. (1974): "Lexicographic Orders, Utilities and Decision Rules: A Survey," Management Science, 20(11), 1442–1471.

Gale, D., and L. S. Shapley (1962): "College Admissions and the Stability of Marriage," American Mathematical Monthly, 69, 9–15.

Glazerman, S., and D. Dotter (2016): "Market Signals: Evidence on the Determinants and Consequences of School Choice from a Citywide Lottery," Mathematica Policy Research, June.

Goldstein, D. (2012): "Bostonians Committed to School Diversity Haven't Given Up on Busing," *The Atlantic*, October 10.

Handy, D. (2012): "Debate on Overhauling Boston Schools' Assignment System Continues," 90.9 *WBUR*, November 13.

Harris, D., and M. Larsen (2015): "What Schools Do Families Want (and Why?)," Technical Report, New Orleans, LA: New Orleans Education Research Alliance.

Hastings, J., T. J. Kane, and D. O. Staiger (2009): "Heterogeneous Preferences and the Efficacy of Public School Choice," Working paper, Brown University.

Hastings, J., and J. M. Weinstein (2008): "Information, School Choice and Academic Achievement: Evidence from Two Experiments," Quarterly Journal of Economics, 123(4), 1373–1414.

Hausman, J. A., and P. A. Ruud (1987): "Specifying and Testing Econometric Models for Rank-ordered Data," Journal of Econometrics, 34(1), 83–104.

He, Y. (2012): "Gaming the Boston Mechanism in Beijing," Working paper, Rice University.

Heckman, J. (2010): "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy," Journal of Economic Literature, 2, 356–398.

Hurwicz, L. (1950): "Prediction and Least Squares," in Statistical Inference in Dynamic Economic Models, ed. by T. C. Koopmans. John Wiley & Sons.

Hwang, S. (2015): "A Robust Redesign of High School Match," University of Britsh Columbia.

Johnson, E. J., R. J. Meyer, and S. Ghose (1989): "When choice models fail: Compensatory models in negatively correlated environments," Journal of Marketing Research, 26(Aug), 255–290.

Kapor, A., C. Neilson, and S. Zimmerman (2017): "Hetereogneous Beliefs and School Choice Assignment Mechanisms," Working Paper, Princeton University.

Katafygiotis, L., and K. Zuev (2008): "Geometric insight into the challenges of solving high-dimensional reliability problems," Probabilistic Engineering Mechanics, 23(2–3), 208 – 218, 5th International Conference on Computational Stochastic Mechanics.

Kling, J. R., S. Mullainathan, E. Shafir, L. C. Vermuelen, and M. V. Wrobel (2012): "Comparison Friction: Experimental Evidence from Medicare Drug Plans," Quarterly Journal of Economics, 127, 199–235.

Kohli, R., and K. Jedidi (2007): "Representation and Inference of Lexicographic Preference Models and their Variants," Marketing Science, 26(3), 380–399.

Manzini, P., and M. Mariotti (2012): "Choice by lexicographic semiorders," Theoretical Economics, 7, 1–23.

Marschak, J. (1953): "Economic Measurements for Policy and Prediction," in *Studies in Econometric Methods*, eds., Hood and Koopmans, New York Wiley, p. 1-26.

McFadden, D. (1974): "The Measurement of Urban Travel Demand," Journal of Public Economics, 3, 303–328.

——— (2001): "Economic Choices," American Economic Review, 91(3), 351–378.

McFadden, D., A. Talvitie, and Associates (1977): "Validation of Disaggregate Travel Demand Models: Some Tests," *Urban Travel Demand Forecasting Project*, Final Report, Vol. V. Institute of Transportation Studies, University of California Berkeley.

McFadden, D., and K. Train (2000): "Mixed MNL Models for Discrete Response," Journal of Applied Econometrics, 15, 447–470.

Menino, T. (2012a): "Press Release," http://www.cityofboston.gov/news/default.aspx?id=5873 November 29.

——— (2012b): "State of the City Address," January 17 Available at http://www.cityofboston.gov/.

Neal, R. (2011): "MCMC using Hamiltonian dynamics," in Handbook of Markov Chain Monte Carlo, ed. by S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, chap. 5. CRC Press.

Nevo, A. (2001): "Measuring Market Power in the Ready-to-Eat Cereal Industry," Econometrica, 69(2), 307–342.

Nevo, A., and M. Whinston (2010): "Taking the Dogma out of Econometrics: Structural Modeling and Credible Inference," Journal of Economic Perspectives, 24(2), 69–82.

Pathak, P., and P. Shi (2013): "Simulating Alternative School Choice Options in Boston," Working Paper, MIT.

Pathak, P. A., and P. Shi (2014): "Demand Modeling, Forecasting, and Counterfactuals, Part I," NBER Working Paper 19589.

——— (2015): "Demand Modeling, Forecasting, and Counterfactuals, Part I," Available at `http://arxiv.org/abs/1401.7359`.

Pathak, P. A., and T. Sönmez (2008): "Leveling the Playing Field: Sincere and Sophisticated Players in the Boston Mechanism," American Economic Review, 98(4), 1636–1652.

——— (2013): "School Admissions Reform in Chicago and England: Comparing Mechanisms by their Vulnerability to Manipulation," American Economic Review, 103(1), 80–106.

Payne, J. W., J. R. Bettman, and E. J. Johnson (1988): "Adaptive strategy selection in decision making," Journal of Experimental Psychology: Learning, Memory, and Cognition, 14, 534–552.

Peters, C. (2006): "Evaluating the Performance of Merger Simulation: Evidence from the US Airline Industry," Journal of Law and Economics, 49(2), 627–649.

Roberts, G. O., A. Gelman, and W. R. Gilks (1997): "Weak convergence and optimal scaling of random walk Metropolis algorithms," The Annals of Applied Probability, 7(1), 110–120.

Roth, A. E. (1982): "The Economics of Matching: Stability and Incentives," Mathematics of Operations Research, 7, 617–628.

Ruijs, N., and H. Oosterbeek (2012): "School choice in Amsterdam. Which schools do parents prefer when school choice is free?," Working paper, Amsterdam.

Seelye, K. Q. (2012): "4 Decades after Clashes, Boston Again Debates School Busing," New York Times, October 4.

Shi, P. (2013): "Closest Types: A Simple Non-Zone-Based Framework for School Choice," Working paper, MIT.

——— (2015): "Guiding school-choice reform through novel applications of operations research," Interfaces, 45(2), 117–132.

Slovic, P. (1975): "Choice between equally valued alternatives," Journal of Experimental Psychology: Human Perception Performance, 1, 280–287.

Thorngate, W. (1980): "Efficient decision heuristics," Behavioral Science, 25(May), 219–225.

Todd, P., and K. Wolpin (2006): "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility," American Economic Review, 96(5), 1384–1417.

Train, K. (2003): Discrete Choice Methods with Simulation. Cambridge University Press, Cambridge, UK.

Tversky, A. (1969): "Intransitivity of Preferences," Psychological Review, 76(1).

——— (1972): "Elimination by aspects: A theory of choice," Psychological Review, 79(4).

Tversky, A., S. Sattah, and P. Slovic (1988): "Contingent weighting in judgment and choice," Psychological Review, 95, 371–384.

Vaznis, J., and T. Andersen (2012): "Plans Upend Boston School Assignments," Boston Globe, September 25.

Walters, C. R. (2013): "A Structural Model of Charter School Choice and Academic Achievement," Unpublished working paper, UC-Berkeley.

Wise, D. A. (1985): "Behavioral Model versus Experimentation: The Effects of Housing Subsidies on Rent," In *Methods of Operations Research 50*, ed. Peter Brucker and R. Pauly, 441-489, Koningsten: Verlag Anton Hain.

Yee, M., E. Dahan, J. Hauser, and J. Orlin (2007): "Greedoid-Based Noncompensatory Inference," Marketing Science, 26(4), 532–549.

Table A1. MNL and MMNL Coefficient Estimates

	MNL			MMNL		
	2012	2013	2014	2012	2013	2014
	(1)	(2)	(3)	(4)	(5)	(6)
distance	-0.365***	-0.403***	-0.557***	-0.638***	-0.674***	-0.793***
	(0.014)	(0.015)	(0.028)	(0.037)	(0.039)	(0.045)
continuing	4.027***	4.354***	4.369***	4.777***	4.966***	5.201***
	(0.052)	(0.054)	(0.064)	(0.069)	(0.068)	(0.085)
sibling	2.104***	2.102***	2.619***	2.478***	2.451***	3.089***
	(0.037)	(0.038)	(0.048)	(0.045)	(0.045)	(0.060)
walk zone	0.500***	0.399***	0.133***	0.339***	0.185***	0.053***
	(0.019)	(0.020)	(0.023)	(0.028)	(0.028)	(0.029)
ell program x ell student	1.548***	1.211***	0.543***	1.892***	1.311***	0.614***
	(0.035)	(0.040)	(0.045)	(0.058)	(0.059)	(0.061)
ell program language match x ell	0.606***	0.672***	0.802***	0.610***	0.967***	0.989***
student	(0.043)	(0.049)	(0.062)	(0.052)	(0.060)	(0.078)
distance x black/hispanic	0.115***	0.114***	0.216***	0.188***	0.183***	0.268***
	(0.010)	(0.011)	(0.019)	(0.024)	(0.024)	(0.031)
distance x block group income	-0.262***	-0.296***	-0.274***	-0.295***	-0.343***	-0.337***
	(0.021)	(0.023)	(0.039)	(0.049)	(0.052)	(0.062)
mcas x black	-0.874***	-1.062***	-0.901***	-1.100***	-1.371***	-1.283***
	(0.105)	(0.111)	(0.089)	(0.153)	(0.144)	(0.130)
mcas x block group income	0.424*	-0.906***	1.762***	1.065***	0.925***	2.388***
	(0.221)	(0.252)	(0.216)	(0.299)	(0.313)	(0.278)
% white/asian x black/hispanic	-2.581***	-2.666***	-2.984***	-3.732***	-3.861***	-4.000***
	(0.097)	(0.094)	(0.114)	(0.162)	(0.148)	(0.170)
% white/asian x block group income	1.982***	1.778***	1.052***	2.633***	2.217***	1.355***
	(0.211)	(0.219)	(0.249)	(0.322)	(0.311)	(0.351)

Notes: This table reports the estimated coefficients of the multinomial logit (MNL) and mixed MNL (MMNL) choice models. The year in each column corresponds to the source year for choice data. All models include a fixed effect for each school. Distance is the Google-maps walking distance from the school to the student's home. Continuing is a binary indicator variable for whether the student is continuing to the school from a previous grade. Sibling is an indicator for whether the student has an older sibling at the school. Walk zone is an indicator for whether the student is in the walk zone of the school, which is approximately a one-mile radius around the school. Ell program is an indicator for whether the program is for English language learners (ELL). Ell student is an indicator whether the student is classified by the district as an English learner and thus eligible to ELL programs. Ell program language match is an indicator for whether the program is an ELL program that targets students who speak a certain language and this language matches the student's home language. Black/hispanic and black are indicators for the student's racial classification. Mcas is the proportion of students at the school who scored "Advanced" or "Proficient" in the previous year's standardized test for math, averaging the proportions for grades 3,4, and 5. Income (est) is the medium household income of the census block group containing the centroid of the student's geocode of residence measured in hundreds of thousands of dollars. % white/asian is the proportion of the enrolled population at the school who are White or Asian. Standard errors are in parenthesis. Standard errors for MNL are computed using the Hessian matrix of the maximum likelihood at the point estimate of the coefficients. Standard errors for MMNL are computed using the sample standard deviation of the MCMC samples.
*significant at 10%; **significant at 5%; ***significant at 1%.

Table A2. Covariance Estimates for MMNL Model

	2012	2013	2014
	(1)	(2)	(3)
	A: Standard Deviations		
σ(ell program x ell student)	1.638***	1.358***	0.959***
	(0.058)	(0.063)	(0.068)
σ(walk zone)	0.981***	0.878***	0.703***
	(0.030)	(0.030)	(0.035)
σ(distance)	0.392***	0.409***	0.499***
	(0.011)	(0.011)	(0.016)
σ(mcas)	2.275***	2.121***	1.837***
	(0.093)	(0.101)	(0.083)
σ(%white/asian)	2.672***	2.512***	2.300***
	(0.093)	(0.106)	(0.106)
	B: Correlation Coefficients		
ρ(distance,mcas)	-0.232***	-0.285***	-0.134***
	(0.041)	(0.043)	(0.049)
ρ(distance, %white/asian)	-0.089**	-0.055	0.021
	(0.039)	(0.040)	(0.051)
ρ(mcas, %white/asian)	0.035	-0.110*	0.236***
	(0.056)	(0.061)	(0.068)

Notes: This table reports covariance matrix estimates for the random coefficients in the mixed multinomial logit (MMNL) model. The year in each column corresponds to the source year for choice data. The variables "ell program," "ell student," "walk zone," "distance," "mcas," and "% white/asian" are defined in Table A1 notes. Panel A reports the square root of the variance of each random coefficient. Panel B reports the Pearson correlation coefficient of the three pairs of random coefficients for which we allow correlation. Standard errors of the estimates are in parenthesis, computed using the sample standard deviation of the MCMC samples.

*significant at 10%; **significant at 5%; ***significant at 1%.

Table A3. Prediction Error in Applicant Count and Demographics

		Predicted (1)	Std. Error (2)	Actual (3)
		A. Count of Applicants		
Grade K1	Continuing	92	7	158
	New	2652	177	2313
Grade K2	Continuing	1482	30	2051
	New	2196	153	1875
		B. Applicant Demographics		
ELL (Grade K1)	Yes	44.4%	1.0%	46.7%
	No	55.6%	0.7%	53.3%
ELL (Grade K2)	Yes	30.8%	0.7%	14.6%
	No	69.2%	0.7%	85.4%
Race	Hispanic	35.1%	0.6%	36.5%
	Black	28.8%	0.5%	28.0%
	White	22.6%	0.5%	22.9%
	Asian	8.4%	0.3%	7.9%
	Other	5.1%	0.3%	4.7%
Median Block Group Income	0-25K	16.9%	0.4%	17.5%
	25-50K	49.7%	0.6%	50.0%
	50-75K	20.7%	0.5%	20.2%
	75K+	12.7%	0.4%	12.3%
Neighborhood	Allston-Brighton	4.5%	0.3%	4.8%
	Charlestown	3.5%	0.2%	3.2%
	Downtown	3.7%	0.3%	3.4%
	East Boston	12.7%	0.7%	12.3%
	Hyde Park	6.3%	0.2%	6.4%
	Jamaica Plain	6.7%	0.4%	7.2%
	Mattapan	6.8%	0.3%	6.8%
	North Dorchester	5.3%	0.5%	5.6%
	Roslindale	8.5%	0.4%	8.1%
	Roxbury	13.6%	0.4%	14.1%
	South Boston	3.2%	0.2%	3.0%
	South Dorchester	13.2%	0.5%	13.6%
	South End	4.7%	0.2%	4.4%
	West Roxbury	7.3%	0.4%	7.2%

Notes: This table compares the predicted and actual new applicants across demographic categories. Column 1 reports the prediction for each category of students and column 2 reports the standard deviation. These are computed from the 1,000 actual samples of applicant pools used for computing the aggregate forecasts. Column 3 reports the actual number of students of each type. Column 1 reports the predicted percentage and column 2 the standard deviation of the prediction. The predictions are based on 2013 data (the last year of the old assignment plan). The numbers shown are the sample mean and standard deviations of the percentage of applicants of each category in the 1,000 simulation samples used for Table 4. Column 3 reports the actual percentages in the 2014 data (the first year of the new assignment plan). Panel A compares the predicted number of applicants to the actual number. Continuing students register in BPS in the previous grade at the time of application. The remaining students are new applicants. Panel B reports applicant characteristics. ELL denotes whether the student is classified by BPS as an English language learner. Race information is mising students for students who applied but did not enroll in any school. Income and neighborhood information are based on centroid of student geocode. Median income is from 2010 census block group.