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

Parag A. Pathak and Peng Shi[†]

October 2017

**Abstract**

Discrete choice models of demand are widely used for counterfactual policy simulations, yet their out-of-sample performance 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 considered several new choice plans that differ in where applicants can apply. At the request of the mayor and district, we estimated discrete choice demand models to forecast the effects of these alternatives. This work led to the adoption of a plan which significantly altered choice sets for thousands of applicants. Pathak and Shi (2014) update forecasts prior to the policy change and describe prediction targets involving access, travel, and unassigned students. Here, we assess how well these ex ante counterfactual predictions compare to the actual choices made under the new choice sets. For equilibrium outcomes, a simple *ad hoc* model performs as well as the more complicated structural choice models for one of the two grades we examine. However, the inconsistent performance of the structural models is largely due to prediction errors in the characteristics of applicants, which are auxiliary inputs. Once we condition on the characteristics of the actual applicants, the structural choice models outperform the ad hoc alternative in predicting both equilibrium outcomes and choice patterns. Moreover, refitting the models using the new choice data does not significantly improve their prediction accuracy, suggesting that the choice models are indeed "structural" and are robust across the reform. Our findings show that structural choice models can be effective in predicting counterfactual outcomes, as long there are accurate forecasts about auxiliary input variables.

[*] The first report of this project was released as NBER Working Paper 19859 in January 2014. We thank Boston Mayor Thomas Menino and Boston Public School Superintendent Carol Johnson for authorizing this study. Boston Public Schools staff, including Kamal Chavda, Tim Nicolette, Peter Sloan, Kim Rice, and Jack Yessayan, provided essential help. We are grateful to our discussant Liran Einav for comments inspiring Section 7, Dan McFadden and Ariel Pakes for encouragement, and participants at the McFadden 80th Birthday conference for input. We also thank Nikhil Agarwal, Isaiah Andrews, Josh Angrist, Steve Berry, Glenn Ellison, Patrick Kline, and Michael Whinston for feedback. Financial support is from 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

The aim of developing models capable of quantitatively forecasting the effects of policy changes has been an objective of economics since at least Hurwicz (1950) and Marschak (1953). In recent years, design-based research strategies that estimate particular parameters or causal effects have become increasingly popular. The design-based approach, however, does not immediately allow for ex ante policy evaluations of changes far outside of historical experience. Both Angrist and Pischke (2010) and Heckman (2010) attribute the growth of design-based research to skepticism about structural modeling for counterfactual analysis given its reliance on parametric and behavioral assumptions.

Though opinions vary on the value of structural models, there is one area of consensus: there are relatively few systematic evaluations of ex ante counterfactual predictions of structural models to the aftermath of the policy change. For instance, Angrist and Pischke (2010, "Industrial Disorganization") 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 many studies 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 sought to modify the plan to assign students to schools closer to their homes, in part to reduce transportation costs. BPS publicized a number of plans that redraw the boundaries of the city and shrink applicant choice sets. BPS and the broader community were interested in predicting the choices families would make under these alternatives and the ensuing final assignments.[3] The mayor and superintendent delayed the timeline for selecting a new plan and asked us to forecast the effects of these alternatives, stating (Menino, 2012b):

[1] Misra and Nair (2011) use a structural agency model to design and implement a compensation scheme, and report that the new scheme's outcomes match those from the model. For merger analysis, Peters (2006) examines the predictive value of structural simulation methods for airline mergers and finds they do not predict post-merger ticket prices well. Ashenfelter and Hosken (2008) argue that design-based estimates of mergers differ markedly from structural estimates. In response to Angrist and Pischke (2010), Nevo and Whinston (2010) describe a few counterfactual validations in the context of merger analysis and Einav and Levin (2010) support more retrospective analyses of past mergers, but express skepticism about cross-merger extrapolation.

[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 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).

> "We have the opportunity to generate an advanced analysis that will allow us to better predict how families would make choices in the real world [...] This is something we have never been able to do before."

Our policy report, Pathak and Shi (2013), uses historical participation and rankings to predict new choices under the different proposals. BPS administrators and the public referred to the report to compare alternatives and ultimately select a new plan.

In January 2014, families throughout Boston ranked schools under new choice sets. For a typical applicant, the new system adds three new school choices, removes sixteen choices, and keeps nine choices intact. Figure 1 summarizes the timeline of the reform. Pathak and Shi (2014) updates the earlier report using the most recent pre-reform data. In this paper, we use choices in the first post-reform year to evaluate the accuracy of predictions.

To describe our approach and questions, we first introduce some notation. Let $X$ encode the characteristics of student and schools, including the set of schools to which each student can apply. Let $Y$ encode the choice outcome of students, which is a rank-ordering over eligible school programs for each student. We observe $(X, Y)$ under the existing policy and can compute equilibrium outcomes of interest, such as the chance students from various neighborhoods are assigned to higher performing schools, as well as the distance students travel and the number unassigned. Let $M(X, Y)$ denote these equilibrium outcomes, which is a well-defined function of $X$ and $Y$, since the assignment mechanism can be exactly recreated before or after the reform. From an ex-ante perspective, the outcome $M(X, Y)$ is a random variable as both covariates $X$ and choices $Y$ are uncertain.

The forecasting problem is to predict what happens under the new policy. We use demand models to learn about the conditional distribution of choices given covariates, $Y|X$. Our paper focuses on two questions:

1. How well do structural models predict equilibrium outcomes important for the school choice context?

2. How well do structural demand models predict raw choice patterns?

Let $(X^*, Y^*)$ be the dataset observed under the new policy regime. The first question compares the actual equilibrium outcomes $M(X^*, Y^*)$ with the forecast $M(X, Y)$. The second question compares actual choices $Y^*$ with the predicted choices conditioned on the actual applicant characteristics, $Y|X^*$. Conditioning on the actual covariates $X^*$ isolates the performance of the demand model from auxiliary forecasts of characteristics.

An innovation of our research design, illustrated in Figure 2, is that we made predictions prior to the policy change. The benefit of making forecasts before the new policy is that we cannot modify forecasts after observing the post-reform data. In this respect, our exercise contrasts with other studies of structural models which 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 on unexpected outcomes and therefore provide a genuine out-of-sample assessment. Motivated by

Nevo and Whinston (2010)'s call to compare structural models to other possibilities, we also report forecasts not based on random utility maximization. This simple alternative provides a reference point from which we judge relative performance.

Our prediction targets come from Pathak and Shi (2013) and were central to the Boston policy debate. For students in each of Boston's 14 neighborhoods, we predict the chance of being assigned to a high-performing school, the average travel distance to school, and the number unassigned. To focus on the choice models, we predict individual choices and the distribution of choices in each neighborhood.

The discrete choice models we fit are the multinomial logit (MNL) model and the mixed MNL model. In a pioneering contribution, McFadden and co-authors used the MNL model to study the impact of BART, San Francisco's rapid transit system (McFadden, Reid, Talvitie, Johnson, and Associates, 1979). They collect data on the travel behavior of a sample of individuals in 1972, prior to the introduction of BART, and estimate MNL models to predict the behavior of the same individuals in 1975 after BART began. McFadden, Talvitie, and Associates (1977) provide a detailed account of the performance of these models. McFadden (2001) summarizes

> "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 other areas of economics involving choice (McFadden, 2001). There have also been many developments in choice modeling in the subsequent four decades. Yet our study is one of the only post-BART out-of-sample validations of discrete choice models of demand of which we are aware.

Our exercise holds 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 historical revealed preferences. The data not only includes a student's top choice, but his entire ranking of schools. Rank order list data contain rich information about substitution patterns among choices beyond what is contained in the top choice (see, e.g., Berry, Levinsohn, and Pakes (2004)). Moreover, our dataset includes a large number of observables, including student characteristics and exact geographic location. In addition, Boston's choice plan has been in existence for more than two decades, so there is a wealth of knowledge and shared experience about the system among participants. 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 reflects this feature.[4] Our exercise should also be particularly informative on substitution patterns since the policy changes applicant choice sets. As preference data from school assignment plans become more widely available, there is a rapidly

[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.

growing literature estimating school demand. Our results speak to the reliability of school demand modeling as a policy planning tool for school districts.[5] Finally, the relatively simple policy change allows us to compare predictions and easily decompose sources of error, which may not be possible in more complicated structural models.

On the other hand, the premise of our exercise, and other predictions based on discrete choice models of demand is that preferences are stable, can be estimated well, and can be used 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 (see Figure 3 shows how choices are presented). In a field experiment, Hastings and Weinstein (2008) provide show that choice behavior in Charlotte's school choice plan can be swayed by informational cues. In other contexts, interventions simplifying information can significantly alter choice behavior (Kling, Mullainathan, Shafir, Vermuelen, and Wrobel, 2012). If these features dominate decision-making, then they may interfere with the reliability of forecasts that assume stable preferences over time.

The rest of this paper is structured as follows. Section 2 provides details on the Boston student assignment plan and events leading up to the adoption of a new plan in 2014. Section 3 describes the data, forecast targets, and how we generate predictions. Section 4 discusses how we evaluate predictions and Section 5 reviews hypotheses motivated by back-testing our framework. Section 6 compares our predictions to the the actual outcomes in the first year of the new plan. It also decomposes sources of prediction error by examining changes to the set of participants and the underlying stability of the demand model. Section 7 reports on how prediction errors affect the ranking of plans other than the one Boston ultimately selected. Section 8 concludes and discusses directions for future work.

## 2 Background

### 2.1 School Choice in Boston

Boston Public Schools has one of the nation's most well-known school choice plans. From 1988 to 2013, the city was divided into the North, West, and East Zone for elementary school admissions, shown in Figure 3. There roughly 25 to 30 elementary schools in each zone. Students residing in a zone are allowed to rank any school in the zone as well as any school within a 1 mile walk zone of their residence and a handful of city-wide schools. 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. Until 2013, for half of the program seats, students residing in the walk zone obtain priority, but this

[5] The growing literature estimating school demand from similar datasets includes: Abdulkadiroğlu, Agarwal, and Pathak (2015), 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), Harris and Larsen (2015), Hastings, Kane, and Staiger (2009), He (2012), Hwang (2015), Kapor, Neilson, and Zimmerman (2017), Ruijs and Oosterbeek (2012), and Walters (2014).

priority does not extend to the other half. A single lottery number serves as the tie-breaker.[6]

Since 2005, after students submit their choices, they are processed through a version of Gale and Shapley (1962)'s student-proposing deferred acceptance (DA) algorithm (Abdulkadiroğlu, Pathak, Roth, and Sönmez, 2005; Pathak and Sönmez, 2008). This algorithm takes as input the submitted preference rankings of students, and the priorities of students to generate an assignment. DA 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 together with applicants that it admitted provisionally in the previous round, 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 there are no new applicants (some may remain unassigned). Under DA, it is a weakly dominant strategy for all participants to rank schools truthfully (Dubins and Freedman, 1981; Roth, 1982). Moreover, this algorithm produces a stable assignment (Gale and Shapley, 1962; Abdulkadiroğlu and Sönmez, 2003).

### 2.2 Policy Reform

The new policy, which began in 2014, affects the set of schools each applicant is allowed to rank. There were two major rationales for the reform. The first was the desire to assign students closer to home from families who wanted assignments at nearby schools and the district which wished to reduce busing costs.[7] Second, there were longstanding concerns about inequities in the three zone system.

The reform was informed by the Pathak and Shi (2013) report, which used choice modeling and simulations to predict the effects of the 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.[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 more details on the role of the report. It's worth noting that there were two prior failed attempts to reform the choice sets of students in 2003 and 2009. Decision-makers did not have access to comparable forecasts during these prior attempts.

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

[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 similar objectives; see, e.g., the discussion about Seattle in Pathak and Sönmez (2013).

[8] BPS's initial plans divided 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)).

Based on the Pathak and Shi (2013) report and other discussions, the Boston school committee adopted the Home-Based plan (see Seelye (2013) and Shi (2013)). This plan constructs customized choice sets based on applicants' exact residential address. It uses a BPS categorization of schools into quality tiers, which are computed using a schools' prior Massachusetts Comprehensive Assessment System (MCAS) test test score growth and levels. Tiers were finalized as of January 2013 for 2014 admissions. Under the new plan, every applicant can 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 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 3 illustrates how participants see choice information. The online application platform lists information on transportation, tier category, and why the choice can be ranked. Previous years' school brochures did not include comparable information.

Aside from the changes to choice menus, the new plan also eliminates walk zone priority (Dur, Kominers, Pathak, and Sönmez, 2016). The school priorities are: continuing students, followed by siblings, followed by other students. As before, a single lottery number serves as tie-breaker. There are no other changes to the DA implementation.

The new plan involves large changes in applicant choice sets. This fact can be seen in Table 1, which shows that for an average grade K1 student, the reform adds three new options, removes sixteen options, and keeps nine options intact. While the choice set changes are substantial, there is still overlap among likely top choices. Only 7% of post-reform grade K1 applicants cannot apply to their top choice under the old plan. Conversely, Panel B shows that 16% of pre-reform grade K1 applicants cannot apply to their top choice in the new plan. The new plan resulted in similar changes in the choice set for grade K2 applicants.

# 3 Prediction Approach

## 3.1 Data Sources

Our data comes from BPS round 1 choice and enrollment files covering years 2010 to 2014. We focus on round 1 assignment, which takes place in January and February, because over 80% of students are assigned then. Forecasts are based on data from 2010 through 2013. We use the 2014 data,

[9] There are a few exceptions to this formula. 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 tunnel travel, East Boston students are eligible for any East Boston school. 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 English language learners or special needs have additional choices. Level 1, 2, and 3 ELL students are allowed to apply to any compatible ELL program within their ELL zone, a six-zone overlay of Boston. Substantially-separate special education students do not apply in round 1.

from the first post-reform year, to evaluate these forecasts.

The choice data contains preference rankings and demographic information for every round 1 participant. The fields include student ID number; English language learner (ELL) 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); lottery number; first 10 choices and priorities at each; school program to which the student was assigned and the priority used for that assignment. Using the assigned school and program codes, we infer the capacity available for round 1 assignment for each school program. We place students to one of 14 neighborhoods using the geocode.[10]

The enrollment data is a December snapshot and contains additional student demographics. The fields are enrolled school and program, grade, geocode, address, gender, race, and languages spoken at home. The file covers the vast majority of the students in the choice data, and can be linked by student ID number. When there is a conflict between the demographic information in the choice and enrollment files, we use the choice file. We also match geocodes to 2010 census block groups, which contain median household income.

For the schools, we have characteristics for each of year (2010-2014). The school file has the building code, address, school type, % of students of each race, % of ELL students, % of students who have special education 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 distance to school, we use walking distance estimates from Google Maps API.[11]

## 3.2 Forecast Targets

Each equilibrium outcome we target corresponds to a single number for each grade and each neighborhood. They are defined as follows:

- **Access to Quality:** A student's chance of being assigned a top tier school if he had ranked it. In particular, we compute the average chance a student from a neighborhood is assigned to any Tier 1 or 2 schools within his choice menu, assuming that he ranks all such schools above other schools and given other students' submitted rankings.

- **Distance:** The average distance between the student's residence and school assignment.

- **Unassigned:** The number of students who are unassigned at the end of round 1.

Using the notation introduced above, these outcomes depend on applicant characteristics $X$ and choices $Y$, as well as the matching mechanism. We therefore refer to these equilibrium outcomes as $E[M(X,Y)]$. As mentioned above, we chose these outcomes since the Boston debate focused on equity of access to quality schools and assignment close to home. Travel distance plays an important

[10] For internal reporting, BPS classifies students into 16 neighborhoods. We combine three neighborhoods with few students, Central Boston, Back Bay, and Fenway/Kenmore, into one neighborhood that we call "Downtown."

[11] For students with missing address information, we treat the outcome centroid of the student's geocode as the address.

role because BPS is required to cover busing costs for all Boston pupils. Unassigned students loom large for facilities planning, staffing, and other budgeting issues.

The second set of targets involve student choices. We predict both individual student choices and choice patterns across a group of students.

For **individual choices**, we predict the top $k$ options ranked highest for each student, where $k$ varies from 1 to 3. For each pair of options in a student's choice set, we also predict whether the student would prefer one option over the other.

For **distribution of choices**, we predict the percentage of top $k$ choices for each school by grade and neighborhood. Furthermore, we predict the aggregate distribution of the top two choices of students, for students who rank at least two choices.

Using the notation introduced above, these choice outcomes depend only on applicant characteristics $X$ and choices $Y$, but not the outcome of the mechanism.

### 3.3 Generating Predictions

We use data from before the reform to fit choice models to forecast outcomes for the first year of the reform and compare these to outcomes induced by the actual choice data.[12] To protect the integrity of our out-of-sample comparison, we specify choice models and forecasts prior to the reform by posting a pre-analysis plan before the data following the reform became available in Pathak and Shi (2014).[13]

Our counterfactual predictions come from three approaches, two of which are based on random utility models. The centerpiece of each approach is the choice model, which maps the characteristics of an individual student as well as the set of schools in his menu to a ranking. The three choice models we examine are:

- **Multinomial Logit (MNL):** This widely-used and easy-to-estimate model is motivated by random utility maximization. It is also the basis of the Pathak and Shi (2013) report.
- **Mixed MNL (MMNL):** This model is a popular alternative to MNL since it can capture substitution patterns that violate the Independence of Irrelevant Alternatives property of MNL models. Mixture models are a significant development in discrete choice models of demand in the years following McFadden (1974) (see e.g., Berry, Levinsohn, and Pakes (1995) and Nevo (2001)).[14]

[12] 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. We abstract away from forecasting capacities this they are at the discretion of the school board and outside the scope of our structural model.

[13] Pathak and Shi (2014) describe the specification of the mixed MNL model, but did not report estimates before posting the report. Estimating the mixed MNL model was too computationally-intensive to complete in time. Pathak and Shi (2015) update the report with the mixed MNL forecasts.

[14] McFadden (2001) states that the MNL methods used to account for substitution between modes of transportation in the BART study are inferior to current methods.

- **Lexicographic:** This model serves as our benchmark for models not motivated by random utility maximization. The model is motivated by psychology and marketing literature. It assumes that applicants rank programs based on an intuitive heuristic.

We describe the choice models in more detail below. For choice outcomes, we take the actual set of students who applied in the first year of the reform and their characteristics, and use the choice model to predict the relative ranking of options within the choice menu. We use the actual set of students to isolate choice prediction from population forecasting. For the MNL-based choice models, we then draw the parameters as jointly normal random variables, using the estimated means and covariance matrix. We next simulate a complete ranking over for each student's choice menu drawing from distribution of idiosyncratic tastes. For individual choices, we predict the *modal* outcome after many simulations.[15] For predicting choice patterns, we use the empirical choice distribution.

For forecasting equilibrium outcomes, there are additional simulation layers. Rather than using the actual applicants as with the choice forecasts, we simulate the pool of applicants and their characteristics. At the time of the typical counterfactual forecast, an analyst does not know future participants. With the simulated applicant pool, we then use each choice model to generate a complete ranking of options within each student's menu, similar to method in the choice forecasts.

We then truncate the generated preference rankings to the first ten choices. Truncation is necessary because the choice data we receive from BPS only has the first ten choices, although there is no restriction on the number of choices in the mechanism. More importantly, this assumption allows us to sidestep modeling students' outside options, for which we have little data.[16] In Pathak and Shi (2013), we performed sensitivity analysis on list length and found ten to be reasonable. In Section 6.2, we further examine this assumption. Another parameter that affects the equilibrium outcome is the number of seats in each school program. For the purpose of the prediction exercise in this paper, we generate predictions based on the assumption that the school board uses the same capacities as in the previous year. In practice, Boston runs DA several times with minor tweaks to capacity but does not report the outcome until the final round 1 run. To abstract away from this back-and-forth iteration, we use these capacities when we compute actual equilibrium outcomes using the actual choice submissions in the first post-reform year.

Finally, we generate i.i.d. lottery numbers for each student, and compute the assignment using DA. In computing access to quality, we compute the probability that the student receives a lottery number that is good enough to be assigned one of the Tier 1 or 2 schools in his menu.[17]

[15] We focus on the mode because the best deterministic prediction of a biased coin that yields heads 60% of the times is that it always yields heads.

[16] Moreover, students often enroll in options they did not rank but could have ranked, undermining the usual assumption that an unranked option is inferior to the student's outside option. 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.

[17] This probability is estimated in a tractable way as follows. If there is at least one Tier 1 or 2 school with a program with excess capacity for which the student is eligible for, then the student's access to quality is 100%. If all such programs are full, then we compute a lottery cutoff for the student, which is the worst lottery number needed for that student to displace out at least one currently assigned student from one of these programs, and we report the chance that the student gets a lottery number at least as high. This approach is exact in a continuum model like

### 3.3.1 Multinomial Logit (MNL) Choice Model

For each student $i$ and each program $j$, we define the MNL model by letting $u_{ij}$ be the indirect utility 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 ELL and the program is an ELL program. The $k^{\text{th}}$ component of this vector is denoted $x_{ij}^k$. The indirect utility of program $j$ is

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

where $s(j)$ denotes the school containing the program[18], $\delta_{s(j)}$ is a school effect, $\beta$ is a $K$-dimensional vector of coefficients, and $\epsilon_{ij}$ represents an unobserved idiosyncratic taste. We assume that $\epsilon_{ij}$ is distributed according to a type-I extreme value distribution, Gumbel(0,1). Since utility has no scale, we normalize the scale parameter to one. The school effect captures unobserved school characteristics such as safety, reputation, facilities, environment, and teacher quality. The estimated parameters are $(\delta, \beta)$.

The rationale for and list of characteristics in $x_{ij}$ is in our pre-analysis plan Pathak and Shi (2014). The final list includes the following:

- distance: walking distance from the student's residence 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 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.

The list of characteristics also includes student-school interaction terms. The interacted student characteristics are the median household income of the student's census block group and race.[19] The interacted school characteristics are distance and the following:

Azevedo and Leshno (2016).

[18] Each school may have multiple programs such as regular education or a specialized program for English language learners. Since students may later transfer between programs within a school, and since Pathak and Shi (2013) did not find significant program fixed effects, we include a school effect rather than a program effect.

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

- mcas: the proportion of the school's students who score Advanced or Proficient in the previous year's MCAS standardized test for math, averaging the proportions for grades 3, 4, and 5.[20]
- % white/asian: the proportion of the school's students who are white or asian.

Hausman and Ruud (1987) extend MNL models to situations with ranking data and we estimate the parameters $(\delta, \beta)$ by maximum likelihood. To quantify uncertainty in the estimation, we estimate a covariance matrix by taking the inverse of the Hessian of the log likelihood function at the maximum. Table A1 reports estimated parameters and standard errors.

### 3.3.2 Mixed MNL (MMNL) Choice Model

This mixed MNL model adds random coefficients to the MNL model. Suppose we place random coefficients on the first $L$ components of $x_{ij}$. The model specifies the indirect utility as

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

where $\delta$ and $\beta$ are fixed effects and coefficients in the MNL model and $\gamma_i$ 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 $\beta$, and the covariance matrix $\Sigma$ satisfies certain restrictions which we specify below. The idiosyncratic term, $\epsilon_{ij}$, is distributed Gumbel(0,1) as in the MNL model. The estimated parameters are $(\delta, \beta, \Sigma)$. The set of characteristics $x_{ij}$ are the same the MNL model, and also include mcas and % white/asian (which are explained in Section 3.3.1).

We allow random coefficients for the following characteristics, which we organize into "blocks." We assume independence across blocks, but allow arbitrary covariance within each block. The blocks are:

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

The covariance matrix $\Sigma$ therefore satisfies the restriction

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

[20] 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. MCAS performance levels need not be a measure of school effectiveness. Abdulkadiroğlu, Pathak, Schellenberg, and Walters (2017) show that in New York City, applicant preferences are uncorrelated with effectiveness once we control for peer quality.

where $\Sigma_1$, $\Sigma_2$, and $\Sigma_3$ are $1 \times 1$, $1 \times 1$ and $3 \times 3$ symmetric positive definite matrices. This formulation allows students to have heterogeneous preferences for ELL programs (if applicable), for schools in the walk zone, and for distance, school performance, and school demographics.

Because the model no longer has a closed form log-likelihood function and the log-likelihood functions are not necessarily globally concave, we fit the model by Markov Chain Monte Carlo (MCMC) methods. One difficulty with our specification is that there are 75 school effects. As far as we are aware, the state-of-the-art MCMC techniques for including fixed effects in mixed logit models, described in Train (2003), involve adding a layer of Gibbs sampling and simulating the conditional distribution of the fixed effects using the Random Walk Metropolis-Hasting algorithm. However, simulating a 75-dimensional distribution is prohibitively slow using Random Walk Metropolis. We therefore use Hamiltonian Monte Carlo (HMC) , which incorporates the gradient of the log likelihood function, to quickly update the 75-dimensional fixed effect (Neal, 2011). 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, we repeat the sampling six times with independent draws with random starting values, and found the results to be nearly identical. Additional details are in Appendix A.

### 3.3.3 Lexicographic Choice Model

When evaluating structural choice models, Nevo and Whinston (2010) emphasize the importance of comparing to an alternative. We therefore consider a model motivated by intuitive 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 pertains 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 parameter estimation. However, it is still motivated by past choice behavior and by 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. 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 tier using distance. For ELL students, BPS staff thought that families prefer ELL programs since they offer targeted programming, especially ELL programs in their home language.

The Lexicographic model is motivated by 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, the agent chooses the better option according to this attribute; if there is little difference, then the agent goes to the next attribute. This encapsulates the Lexicographic model above if we define the academic quality of schools based on tier. 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 is therefore a special case of elimination by aspects with a deterministic ordering of aspect given by the hierarchy.[21]

We picked this alternative to compare with our discrete choice model given its empirical support. 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 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.

### 3.3.4 Predicting Who Applies

For grades K1 and K2, all students who wish to be assigned a Boston Public school must participate in the choice process. The set of applicants is therefore an important determinant of our equilibrium targets. A large influx of new applicants in a given neighborhood would increase the number unassigned and reduce average access to top tier schools. If we had data on all potential applicants and their non-BPS options, we might include the decision to participate as part of the structural model. Since we don't have this data, we still need to reflect this uncertainty and to capture any trends in the neighborhood participation patterns.

To predict who applies, we use demographic trend projections. A similar approach was used in the McFadden, Reid, Talvitie, Johnson, and Associates (1979) BART study, who write in Chapter

[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.

IV.3

> "It is in the nature of auxiliary forecasting that one does not have available complete structural or causal models; hence, forecasting must use data analysis and trend projection techniques, combined with available external forecasts."

McFadden et al. use census demographic data and projections to construct a representative sample of San Francisco households. Since we have the universe of participants in previous years, we directly observe the joint distribution of household characteristics and use it to predict the applicant pool.

We construct our applicant pool as follows. For continuing students, we exploit the fact that they are already in the enrollment data of the previous year, and focus on the predicted the probability that a given student who is currently enrolled will choose to continue on to the next grade. Once we have a predicted probability, we include each currently enrolled student as a continuing student with this probability, independently from everything else, and assume that the student will continue in the same program in the next year. We model the probability of continuing to be normally distributed and common across students for each grade-neighborhood combination. The common probability for the neighborhood allows for a common shock on the number of continuing students. For each grade and each neighborhood, the mean and standard deviation of the normal random variable are estimated based on previous years' data. To detect time trends, we regress the number of students per neighborhood by year using four years of data from 2010-2013. For grade-neighborhood combinations for which the slope of the regression is not significant at a 95% confidence level, we discard any time trend and use the sample mean and sample standard deviation from the previous four years. For the grade-neighborhood combinations in which the regression slope has 95% significance, we use the predicted mean and standard error of the regression.[22]

For new students, we use previous year's applicant demographics as proxies. We first forecast the total number of new applicants from each grade and neighborhood, and then sample with replacement from the set of new applicants of the previous year from this grade and neighborhood. We model the number of new applicants as the product of two independent normals, one representing a BPS-wide shock and one a neighborhood-specific shock. The common shock captures macro effects such as BPS publicity or economic factors driving private school enrollment. The neighborhood-specific 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 trend in the same way. Pathak and Shi (2014) provide additional details about the performance of this approach. All predictions of equilibrium outcomes are based on 1,000 independent simulated samples of the applicant pool.

[22] Grade K1 Charlestown and K2 Downtown are the only two grade-neighborhood combinations with a steady upward trend in the number of applicants.

# 4 Evaluating Predictions

## 4.1 Equilibrium Outcomes

Our metrics for evaluating equilibrium outcomes (access to quality, unassigned, or distance) take several forms of uncertainty into account. Let $\omega_h$ be a random variable that corresponds to the simulated outcome for neighborhood $h$ generated by a choice model and let $\omega = (\omega_1, ..., \omega_H)$. This variable is random due to the randomness in the generation of the applicant pool, uncertainty in estimated choice model parameters, the choice model's taste shock, and the lottery numbers. The prediction for neighborhood $h$ is $\bar{\omega}_h \equiv \mathbb{E}[\omega_h]$. This quantity can be estimated by sampling $\omega_h$ 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 $\bar{\omega}_h$ for each $h$.

Let $\omega^*$ denote the actual outcome vector, computed using the actual population, actual choices, actual lottery numbers, and the school capacities described above. The root mean squared error (RMSE), which is our main measure of prediction error for equilibrium outcomes, is defined as

$$\mathbf{RMSE} \equiv |\bar{\omega} - \omega^*|_2 \equiv \sqrt{\sum_{h=1}^{H} (\bar{\omega}_h - \omega_h^*)^2}.$$

The RMSE is an overall measure of the amount of prediction error across neighborhoods. To measure uncertainty in the prediction, we define

$$\mathbf{Expected\ RMSE} \equiv \mathbb{E}[|\bar{\omega} - \omega|_2].$$

The expected RMSE measures how much prediction error we should expect when the choice model is correct.

In addition, we estimate a 95% confidence interval by computing 2.5$^{\text{th}}$ and 97.5$^{\text{th}}$ percentile of each predicted outcome for each grade-neighborhood combination. Let this interval be denoted as $\Omega_h$. The proportion of neighborhoods for which the actual prediction is within the confidence interval,

$$\mathbf{\%\ in\ 95\%\ C.I.} = |h : \omega_h^* \in \Omega_h|/H,$$

is our last measure of prediction accuracy. If the model is correct, then we expect this to be close to 95% on average.

## 4.2 Choice Forecasts

To measure prediction accuracy for individual choices, we report the percentage of mistakes for top choices and for pairwise comparisons from the rank order list. To define the metrics precisely, we first define some notation. For a given choice model, let $y_i$ be the vector of simulated preference rankings of student $i$. $y_{i1}$ is the index of the top choice, $y_{i2}$ is the index of the second choice, and so on. Define the set of top $k$ choices as $Y_{ik} \equiv \{y_{i1}, ..., y_{ik}\}$. Similarly, let $y_i^*$ denote the student's actual choice ranking, $r_i$ denote the actual number of choices ranked, and $Y^*_{ik}$ denote the actual set of top $k$ choices. Denote $I_k = \{i : r_i \geq k\}$ as the set of students that ranked at least $k$ choices.

Given a choice model, define $\hat{Y}_{ik}$ to be the best prediction of the top $k$ choices for student $i$. We predict by simulating the top $k$ choices $Y_{ik} = \{y_{i1}, \cdots, y_{ik}\}$ many times and taking the $k$ most common choices across simulations. Our first measure computes mistakes among the top $k$ choices:

$$\textbf{\% Mistakes in Top k Choices} = \frac{1}{|I_k|} \sum_{i \in I_k} \frac{|\hat{Y}_{ik} \backslash Y^*_{ik}|}{k}.$$

That is, we tabulate the average proportion of predicted top $k$ choices in the set $\hat{Y}_{ik}$ that are not in the actual set of top $k$ choices, $Y^*_{ik}$, counting only students who ranked at least $k$ choices. When $k = 1$, for example, this measures the fraction of top choices that are incorrectly forecast.

A second measure of prediction error considers pairwise comparisons. Given the actual ranking $y^*_i$, define the set of pairwise comparisons implied by this ranking to be a collection of ordered pairs:

$$C^*_i = \{(j, l) : \text{program } j \text{ is ranked before } l \text{ in } y^*_i\}.$$

A pair of programs $(j, l)$ is in this set if both programs $j$ and $l$ are ranked and $j$ is preferred, or if $j$ is ranked and $l$ is unranked. Given a choice model, define $\hat{z}_i(j, l)$ to be the indicator variable if we predict that student $i$ prefers option $j$ over option $l$. This indicator is equal to 1 if the probability that $j$ is preferred over $l$ is over 50%, and is 0 otherwise. Define the percentage of mistakes in pairwise comparisons to be the proportion of comparisons that the choice model predicts incorrectly:

$$\textbf{\% Mistakes in Pairwise Comparisons} = \frac{1}{|I_1|} \sum_{i \in I_1} \frac{1}{|C^*_i|} \sum_{(j,l) \in C^*_i} (1 - \hat{z}_i(j, l)).$$

For the distribution of choices, we use statistical distance to aggregate comparisons. We compute this metric for the distribution of top choices for each neighborhood and the joint distribution of the top two choices. For neighborhood $h$, define market share $s_{hjk}$ to be the average proportion of top $k$ choices from this neighborhood that are for school $j$, counting only students who ranked at least $k$ choices. Let $I_{hk}$ denote the set of students from this neighborhood who ranked at least $k$ choices. The predicted top $k$ market share of school $j$ in neighborhood $h$ is

$$s_{hjk} = \frac{1}{|I_{hk}|} \sum_{i \in I_{hk}} \mathbb{E}[|\text{programs in } Y_{ik} \text{ at school } j|]/k.$$

Similarly, define the actual market share $s^*_{hjk}$ using the actual set of top $k$ choices $Y^*_{ik}$ instead of the predicted set $Y_{ik}$ for each student $i$. The statistical distance between two vectors is the minimum mass needed to transform one vector to the other. It is also sometimes called the total variation.

The average statistical distance across $H$ neighborhoods is defined as:

$$\textbf{Statistical Distance in Top } k \textbf{ Market Share} = \frac{1}{2H}\sum_{h=1}^{H}\sum_{j}|s_{hjk} - s^*_{hjk}|.$$

The final metric we examine is for the joint distribution of two highest ranked schools. For pair of options $(j, l)$, define $p_{jl}$ as the proportion of students who ranked at least two choices who ranked school $j$ first and school $l$ second. Similarly, we define $p^*_{jl}$ for the corresponding actual choice rankings. The following measures the minimum mass needed to transform one distribution to the other:

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

## 5 Back-testing and Hypotheses Formulation

Even though the policy reform involves a simple change to choice sets, our forecast approach rests on several assumptions. Is there any hope that our predictions will be reasonable? To set expectations, we report on a back-testing exercise which applies our prediction methodology on data from prior years: 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 the predicted and actual access to quality by neighborhood in 2013. Each bar corresponds to a choice model. For each prediction, the figure also plots the 95% confidence interval. These estimates allow us to compute the overall root mean squared error (RMSE) across neighborhoods, our summary measure of prediction error, and the proportion of neighborhoods for which the actual access to quality falls within the confidence interval.

For each grade and for each moment of interest, the MNL and MMNL models exhibit nearly-identical RMSE. This fact is shown in Table 2. Moreover, for every combination except access to quality in grade K1, the MNL-based models exhibit smaller RMSE than the Lexicographic model. However, the absolute performance of the MNL-based models involves several inaccuracies: the RMSE is larger than the expected RMSE and the predicted outcome is within the predicted 95% confidence interval (C.I.) only about 70% of the time, averaging across outcomes. For the Lexicographic model, the performance is worse: the actual outcome is inside the 95% C.I. less than 40% of the time.

The MNL and MMNL models also outperform the Lexicographic model for individual choice predictions. Table 3 reports on MNL performance for each grade and for the top 1, top 2, and top 3 choices. As a benchmark, Table 3 also tabulates the accuracy of random guessing. For grade K1, the table shows that random guessing predicts the top choice wrong 97% of the time. The Lexicographic model predicts the top choice incorrectly 63% of the time, which means that it predicts the top choice (out of more than 30 options) correctly 27% of the time. For pairwise comparisons, random guessing predicts wrongly 50% of the time by definition. The Lexicographic model predicts incorrectly 30% of the time, 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, the Lexicographic model is not much better than random guessing, and for the joint distribution of top 2 choices, it can be even worse than guessing. These facts are shown in Panel B of Table 3. The Lexicographic model does not allow for students to prefer a more distant school in the same tier to a closer school, if the continuing, sibling, and English language learner status of the student are the same at both schools. For these metrics, the performance of the MNL-based models is nearly identical, and both outperform random guessing and the Lexicographic model.

As a result of this analysis, we formulate the following hypotheses before choices are submitted in the new plan:

- 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 correct.

- For choice forecasts, the comparison across choice models would be the same as the equilibrium forecasts. Moreover, the Lexicographic model would reasonably predict individual choices, but would perform poorly for the distribution of choices.

# 6 Comparing Forecasts and Prediction Errors

## 6.1 Equilibrium and Choice Forecasts

Figure 5 shows the actual access to quality in the first post-reform year (2014), as well as the predicted access to quality according to each choice model based on pre-reform data for each neighborhood. For each prediction, the figure also contains the 95% confidence interval.

The MNL-based models outperform the Lexicographic model only for grade K1, but not for grade K2. Table 4 shows that 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-test. However, for grade K2, the Lexicographic model exhibits similar RMSE as the MNL model, with slightly better prediction accuracy for two out of the three targets: access to quality and number unassigned. Moreover, for these two targets, 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.[23]

[23] This phenomenon is not due to greater uncertainty in the Lexicographic model prediction. Column 5 of Table 4 shows the expected RMSE of Lexicographic is similar to the MNL-based models.

We cannot reject the other hypotheses about equilibrium outcomes in Table 4. In all cases, the MNL and the MMNL models exhibit near-identical results, regardless of whether we consider the RMSE or the fraction of predictions within the 95% confidence interval. The expected RMSE is also similar between the two models. In addition, regardless of the model or the metric, the actual RMSE is higher than the expected RMSE, which shows that none of the models is accurate in an absolute sense.

For choice forecasts, the results of the prediction exercise are similar to those reported in the back-test, and are consistent with our hypotheses. Table 5 shows that the MNL and MMNL models exhibit near-identical performance, regardless of the grade and the metric. 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 the distribution of choices than individual choices. This pattern was present in the back-test.

## 6.2 Decomposing Prediction Errors

When we use the actual applicants and their characteristics but not their choices, the MNL-based choice models systematically outperform the Lexicographic model. Table 6 shows the predictions about relative performance of the choice models are comparable to the back-test when we use the actual set of applicants and their characteristics, instead of predicting these pre-reform. The MNL and MMNL models have near-identical performance, 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 predict applicants. For the number unassigned, all of the choice models have similar prediction accuracy, and the RMSEs are much smaller than the corresponding RMSEs that predict applicants. Nevertheless, the actual RMSE is larger than the expected RMSE in all cases, consistent with expectations set 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. Table 7 reports on how prediction accuracy changes with 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 Old Demand Model: Using the actual set of applicants, but choices from demand model fit with old applicants, shown in Table 6.
- New Applicants with Refit Demand Model: Using the actual set of applicants and choice models estimated from the actual choices.[24]
- New Applicants with Refit Demand Model and Ranking Length: Using the actual set of applicants and choice models estimated from the actual choices as well the actual number of choices ranked by each applicant.

[24] Table A1 and Table A2 contain the coefficient estimates

- Sampling Actual Choices and Using Applicant Forecast: 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 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 Old Demand Model" is similar compared to "New Applicants with Refit Demand Model."
2. The assumption about rank-order list length is 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 unassigned. In comparison, predictions of the applicant pool are first-order, as the RMSE improves significantly for every metric when we compare the original forecasts to the version with the actual applicants.
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 sampling actual choices using the applicant forecast. new applicants.

The inconsistent performance of the MNL-based models in the original forecasts is driven by errors in the applicant forecast rather than the choice models. 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 or post-reform data. These findings suggest that discrete choice models can be effective in predicting counterfactual outcomes, as long as there are accurate forecasts about auxiliary input variables.

The stability of the MNL-based choice models across the reform is shown in Table 8, which compares model performance when we estimate using pre-reform data and post-reform data. The prediction error decreases when we estimate the model from post-reform choices, but the reduction is relatively small. This fact provides 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 are a close proxy to a choice model from choices made under the new policy.

The new presentation of the choice menu has a small effect on the distribution of preferences. Figure 6 plots the percentage of actual top choices that are Tier 1 compared to predicted choices from the choice models. Tier has an effect in grade K2, where top choices shift toward Tier 1 schools by a few percentages compared to the MNL-based predictions. The Lexicographic model overstates the importance of tier since it predicts a larger shift toward Tier 1 schools.

While preferences measured by the choice model appear stable across policies, the set of applicants are not. Table A3 shows that there are three major errors: (1) the number of continuing

K2 students is much larger than predicted, (2) the number of new grade K1 and K2 students is significantly less than predicted, and (3) the proportion of grade K2 ELL students are less than predicted.

Were these errors foreseeable? It is difficult to comment on this with any level of rigor since almost anything can seem foreseeable after the fact. Nevertheless, we give our best guesses below. We think that the first source of error was possibly foreseeable, as it is caused by misunderstanding of how BPS assigns continuing students. We assumed that currently enrolled students who wish to continue are assigned the same program code for the next grade, but in reality BPS sometimes changes the program code when students change grades and our forecast did not capture these changes adequately. The second error is unexpected as the number of applicants had been rising in previous years. The low number of applicants is either due to a break in the previous trend in the number of kindergarten-aged children in Boston, or due to a greater substitution to school options outside of BPS, including charter and private schools or public schools in neighboring districts. Our data do not allow us to distinguish these two alternatives. The third discrepancy is driven by a simultaneous change in the test that BPS uses to determine eligibility to ELL programs, which decreased the proportion of eligible students. This third change was done by the BPS Office of English Learners, which has little overlap with the office in charge of school assignment, and was therefore hard for us to foresee.

# 7 Selecting Another Policy

While our analysis has focused on the absolute accuracy of the choice models, it's also worth considering whether BPS would have chosen a different choice plan given the prediction errors. Even if the prediction errors are large in an absolute sense, they may not affect the relative ranking of alternative plans and BPS's policy decision.

The alternative choice plans we consider are the 2012-2013 school assignment reform proposals described in Pathak and Shi (2013). Most proposals partition the city into alternative zones, ranging from six to twenty-three. Two proposals are variants of the Home-Based plan. The decision-making that led Boston to adopt the Home-Based plan involved a compromise across several dimensions. But the effects on access and proximity were central, and the school board was also concerned about insufficient school capacity. We therefore evaluate the relative performance of other plans with respect to these three equilibrium targets.

Table 9 reports on access to quality for grade K1 for the Allston-Brighton neighborhood. Each entry of Panel A reports access to quality for eight plans for different choice models and applicant samples. Column 1, for example, shows that access to quality is highest under the 10 Zone plan, according to the MNL choice model estimated with post-reform choices with post-reform applicants. In contrast, access to quality is 72.0% under the status quo. Since we cannot directly compare plans that were not implemented, column 1 serves as our reference point. The ranking of plans is unchanged when we use the MNL model fit pre-reform, but with the post-reform applicants. Access to quality is highest under the 10 Zone plan and lowest under the status quo. Panel B shows that of the possible comparisons (e.g., Status Quo vs. Home Based A, Status Quo vs. Home Based B, Home Based A vs. Home Based B, etc.), there are no reversals of pairwise comparisons.[25]

A more direct assessment of the impact of prediction errors on the ranking of alternative plans is shown in column 3 of Table 9. Here, we report forecasts of access to quality based on pre-reform choices and applicants. This more closely mirrors the Pathak and Shi (2013) report. The forecast provides a more optimistic scenario for both versions of the Home-Based plan compared to the reference in column 1. Specifically, the Home-Based plans have the highest access to quality after the 10 Zone plan, but in column 1 they have the lowest access to quality after the status quo. In fact, there are reversals across 8 of 22 possible non-trivial pairwise comparisons of plans. This suggests that if Boston chose a plan based only on access to quality in the Allston-Brighton, the MNL model forecast could have led to a different choice. However, this pattern also is present with the Lexicographic model, where access is higher under the Home-Based plans than other alternatives. There are more reversals of pairwise comparisons under Lexicographic than MNL.

Access to quality in a given neighborhood is not the only factor used to select among plans. We therefore report on how the ranking across plans changes under different choice models, aggregating across the three outcomes and 14 neighborhoods in Table 10. About 5% of the pairwise comparisons across plan dimensions change when the MNL model is fit from pre-reform data compared to post-reform data. In other words, the MNL choice model generates a similar ranking of plans for each metric and neighborhood before and after the policy reform. However, column 3 shows that there are larger reversals across the ranking of plans with the choice model fit pre-reform on pre-reform applicants. The relative rankings reverse on average 16% of the times, shown in the last row of the table. The extent of reversals of plan rankings is more than three times higher the number of reversals when we had the same applicant pool as the point of reference in column 2. That is, the relative ranking of alternative policies changes significantly due to errors in forecasting applicants. In other words, Boston may have chosen another plan had there been a better forecast of the auxiliary variables.

Does the susceptibility of the MNL forecast to errors in who applies undermine its value for decision-making? The answer to this question depends on the performance of the alternative. Column 3 shows that errors in forecasting applicants do not erase the benefits of the MNL model compared to Lexicographic. Despite the errors in the forecast of the auxiliary variables, there are significantly fewer prediction reversals with the MNL model fitted from past data than the Lexicographic model, under which nearly one-third of pairwise comparisons across plans are reversed. The performance of the Lexicographic here is not much better from a random prediction, which would reverse one-half of comparisons. In summary, even though counterfactual comparisons are sensitive to prediction errors in auxiliary covariates, there is still value in using a structural choice model instead of our alternative.

[25] For this tabulation, we only consider comparisons where the difference in access is at least 1%, to avoid tallying trivial differences across plans.

# 8 Conclusions

This paper report on an out-of-sample validation of structural models of school demand. Forecasts from these models influenced a policy change that affected thousands of Boston families. We made predictions prior to the policy change, so it is not possible to modify predictions after observing realized outcomes. Since we observe choices participants made in the new policy, we also conduct a decomposition of sources of prediction error.

We find that, once we control for changes in the environment outside of the structural model, the choice models are reasonably accurate compared to expectations set by back-testing. Both the MNL and mixed MNL choice model significantly outperform the Lexicographic model, when using the actual applicants. Moreover, the performance of the MNL-based models is similar when refit with post-reform data, suggesting that the distribution of preferences measured by the choice model are stable even with a large change in choice sets and how choices are framed. We also find that the MNL model's performance is similar to the mixed MNL model, a fact foreshadowed in the back-tests. The micro-level data we have on individual characteristics likely reduces the potential benefit of the more flexible and computationally-intensive specification.

The scenario where an analyst has access to the actual participants under the new policy allows us to focus attention on choice model performance. But it is a hypothetical scenario that does not correspond to any real-world forecasting problem. Without the actual participants, the magnitude of the error from the applicant forecast is so large for grade K2 that it undermines the performance benefit of the MNL model. In fact, without using the actual applicants, the prediction error from the Lexicographic specification is smaller for several forecast targets compared to the MNL model in grade K2. Our decomposition shows that the superior performance of Lexicographic is driven by the fact that errors in the applicant and choice forecasts counteract each other. The error in the MNL forecast is large enough to change the ranking of several other alternative policies, and may have led the city to pick a different plan. However, the negative effects of errors in the applicant forecast does not erase the benefit of structural modeling: despite the presence of the errors in auxiliary inputs, the correctly specified model fitted from past data still reproduces the majority of counterfactual comparisons across plans, and does so much more consistently than the ad hoc alternative.

In absolute terms, there is still substantial scope to improve the demand model predictions. An open question is whether a more principled approach to variable selection in the choice models have led to further improvements. It's also possible that alternative non-choice based approaches would have improved performance.

Structural demand models have widespread application in economics beyond demand for schools. Our setting and policy change show possibilities for scenarios where substitution among choices is central. While standard choice models may succeed in predicting choice behavior, there can still be significant unforeseen error for outcomes that depend on choices due to changes in the environment that are outside of the model. Difficulty predicting these auxiliary inputs likely plays a large role in other applications.

# A Estimating the Mixed MNL Choice Model

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

Train (2003) reviews the basic framework to calibrate estimate MMNL models using MCMC. It is based on Gibbs sampling and the Metropolis-Hasting algorithm. However, our setting has more fixed coefficients since we have a fixed effect for every school. 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 3.3.2 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.

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

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

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

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

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

where $\Sigma_1$, $\Sigma_2$ and $\Sigma_3$ are $1 \times 1$, $1 \times 1$ and $3 \times 3$ 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 $\gamma_i$, 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 $\gamma_i$ each has five dimensions.)

Our estimates will be based on sampling the parameters based on this likelihood function $\Phi$. Because $\Phi$ 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 sample from a distribution by constructing a Markov chain whose unique stationary distribution is the desired distribution of interest. 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 arrive at samples from the desired distribution.

The workhorse 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. 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 $\Delta$, where $\Delta$ 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 arrive 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 $\epsilon$ is multivariate normal distributed, $\epsilon \sim \texttt{Normal}(0, \rho I)$, where $I$ is the identity matrix and $\rho$ 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 $\rho$ and successive samples and there will be too much serial correlation; too large a $\rho$ and acceptance probability might be near zero so the chain may get stuck. We tune $\rho$ 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 $\rho$ that is good for all dimensions.

### A.2.2 Metropolis Within Gibbs (MWG)

Metropolis Within Gibbs 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. It also reduces the number of dimensions sampled at each step. The drawback 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 make detailed balance work out, the first and last steps of simulation are half-steps. Precisely

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

speaking, let the gradient of the log likelihood function be $G(x) = \nabla(\log(L(x)))$. Let $\epsilon$ and $\Delta$ 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 $\epsilon$ 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 $\rho$ 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 $\epsilon$, 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, we preserve detailed balance. Neal (2011) describes these as best practices for applying HMC.

## 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 $\gamma_i^{t-1}$. (See Equation 5 for definition of $\phi_i$.) 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 $l$th block of $\gamma_i^t$ assuming mean as in the $l$th block of $\beta_r^t$. (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 $l$th block and let $n$ be the number of students. Draw $\Sigma_l^t$ 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 $\beta_f$ 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 $\rho$ 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.

# B Computing Equilibrium Forecasts

All post-reform equilibrium forecasts are computed by averaging the results of 1000 iterations of the following sequence of steps.

1. Sample applicant pool $X$ according to the assumptions described in Section 3.3.4. More details are given in Section 4.2 of the Part I report, Pathak and Shi (2015).
2. Sample choice model parameters.
   - For the Lexicographic model, we skip this step since the model does not have parameters.
   - For the MNL model, we sample
     $$(\delta, \beta) \sim N(\mu, \Sigma),$$
     where $\mu$ is the maximum likelihood estimate of the fixed effect $\delta$ and coefficients $\beta$, and $\Sigma$ is the inverse of the Hessian of the log-liklihood function evaluated at $\mu$.
   - For the MMNL model, we sample $(\delta, \beta, \Sigma)$ from the posterior distribution from MCMC, and independently sample for each student $i$ the individual coefficients $\gamma_i \sim N(\beta_r, \Sigma)$.
3. For each student, compute a relative ranking of all options within his choice menu that he is eligible for, truncating to the top 10 choices. (This corresponds to the $Y|X$, using the notation introduced in Section 1. ) For the MNL and MMNL models, this involves independently sampling idiosyncratic taste shocks $\epsilon_{ij} \sim \text{Gumbel}(0, 1)$ for every student $i$ and eligible option $j$. We also sample a lottery number $l_i$ for each student $i$, $l_i \sim Uniform(0, 1)$.
4. Compute the assignment using the deferred acceptance algorithm described in Section 2 using the following inputs.
   - The simulated choice rankings from the previous step.
   - The program capacities imputed from the round 1 assignment from the previous year (which is 2013 for the calculation of post-reform forecasts.)

- The following priority structure. Define the priority of student $i$ for program $j$ to be (the higher the better)

$$\pi_{ij} = Boost_{ij} + l_i, \tag{7}$$

$$Boost_{ij} = 8Continuing_{ij} + 4PresentSchool_{ij} + 2Sibling_{ij} + SameSide_{ij}, \tag{8}$$

  where the variables on the right hand side of (8) are binary indicator variables for whether the student is a continuing student for program $j$, a continuing student for another program in the same school as program $j$, has a sibling in the school of program $j$, or is on the same side of the East Boston bridge as the school housing program $j$.

5. Compute the equilibrium outcome of interest for each of the fourteen neighborhoods.

   - Access to quality: Let the set of students assigned to school $j$ be denoted $I_j$, and define

   $$z_j = \begin{cases} \min_{i \in I_j} \pi_{ij} & \text{if school } j \text{ is full,} \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

     This is an estimate of the minimum priority needed to get into school $j$, given the generated preferences and priorities of other students. The estimate is based on the large market approximation of Azevedo and Leshno (2016). Define the access of student $i$ to school $j$ to be the probability that his lottery number is high enough for his priority to be higher than the cutoff of $z_j$,

   $$Access_{ij} = \max(\min(Boost_{ij} + 1 - z_j, 1), 0), \tag{10}$$

     and the student's access to quality as the maximum $Access_{ij}$ over all program $j$ in his menu from a Tier 1 or 2 school. The final result is the average of the access to quality estimates for every student $i$ living within the neighborhod.

   - Distance: compute the average walking distance an assigned student from the neighborhood to his assigned school. The walking distance is from Google Maps API, based on the student's home address and the school's address. For students for whom we do not have the home address, we use the centroid of the geocode where the student lives as a proxy.

   - Unassigned: compute the number of students from the neighborhood who is not assigned.

In each of the 1000 iterations, we compute for each neighborhood a scalar estimate for each of the three equilibrium outcomes of interest. The final forecast is the average of these 1000 values. The estimated 95% confidence intervals are from the empirical 2.5 and 97.5 percentiles of these 1000 values.

In computing the actual outcome, only Steps 4 and 5 are needed. Instead of the simulated values from steps 1-3, we use the actual applicant pool $X^*$, the actual choices $Y^*$, and the actual lottery number $l_i$ for each student $i$. As a result, only one iteration is needed.

The pre-reform forecasts (from the back-testing exercise) are computed similarly, except that Step 4 above is altered to account for the different priority structure. Instead of the same-side priorities above, the pre-reform assignment plan contains walk-zone priorities, which only apply to 50% of the seats. The exact implementation is as follows. Each program $j$ is split into two bins of equal size, $j_1$ and $j_2$. Bin $j_1$ is called the walk-zone bin and $j_2$ is the open bin. If program capacity is odd, then the walk-zone bin has one additional seat. The preferences of students are augmented to be over the bins, so that for the same program, every student prefers the walk-zone bin over the open bin, but the relative preference between programs is as before. Priorities are now computed for every student $i$ and every bin. For a walk-zone bin $j_1$ of program $j$, the priority boost is

$$Boost_{ij_1} = 8Continuing_{ij} + 4PresentSchool_{ij} + 2Sibling_{ij} + WalkZone_{ij},$$

where $WalkZone_{ij}$ is a binary indicator variable for whether student $i$ lives in the walk-zone of the school housing program $j$. For an open bin $j_2$, the boost is as above except without the $WalkZone_{ij}$ term. Given these preferences over bins and priorities of students to bins, we compute an assignment of students to bins using the deferred acceptance algorithm. For access to quality, we define the access of each student to each bin using the analog of Equation (10) for bins, and define a student's access to quality by finding the maximum access to an eligible quality bin, which is defined to be a bin of a program from a Tier 1 or 2 school in the student's menu.

# C Evaluating Choice Forecasts

Using the notation of Section 4.2, the quantities that need to be computed to evaluate choice forecasts for a given choice model are as follows.

1. Best prediction $\hat{Y}_{ik}$ for the set of top $k$ choices of student $i$, where $k \in \{1, 2, 3\}$.
2. For each $k \in \{1, 2, 3\}$, market share $s_{hj}$ of top $k$ choices from this neighborbood that is for a program in school $j$.
3. For each tuple of schools $(j, l)$, the proportion $p_{jl}$ of students who ranked at least two choices, who ranked school $j$ first and $l$ second.
4. For each tuple of programs $(j, l)$, best prediction $\hat{z}_i(j, l)$ for whether student $i$ prefers program $j$ over program $l$.

Items 1-3 can be computed using many samples of the permutation of top 3 choices, $(y_{i1}, y_{i2}, y_{i3})$, for each student $i$. For $\hat{Y}_{ik}$ this is because due to the way the percentage of mistakes in Top $k$ choices is defined, we have by linearity of expectations that the optimal deterministic prediction $\hat{Y}_{ik}$ if we believe the choice model to be correct is simply the top $k$ most commonly occurrent options in the set $\{y_{i1}, \cdots, y_{ik})$. For $s_{hl}$ and $p_{jl}$, having many samples of the permutation of top 3 choices suffices since the empirical market shares and empirical proportions are unbiased estimates of the true values.

- For the Lexicographic model, one sample of $(y_{i1}, y_{i2}, y_{i3})$ for each student $i$ suffices since the model is deterministic.
- For the MNL model, we sample 5000 independent draws of model parameters $(\delta, \beta) \sim N(\mu, \Sigma)$, where $\mu$ is the maximum likelihood estimate and $\Sigma$ is the inverse of the Hessian of the log-likelihood function at $\mu$. For each draw of $(\delta, \beta)$, and for each student $i$ and program $j$, we produce 200 independent draws of $\epsilon_{ij} \sim \text{Gumbel}(0, 1)$, and use these to simulate rankings. Hence, for each student, we have 1,000,000 samples of $(y_{i1}, y_{i2}, y_{i3})$ which are almost independent of one another.[27]
- For the MMNL model, we use the same recipe as above: we produce 5000 independent samples of the model parameters $(\delta, \beta, \Sigma)$ from the MCMC posterior and for each of these samples and each student $i$, we produce an independent draw of individual coefficients $\gamma_i \sim N(\beta_r, \Sigma)$. For each of the 5000 combinations of $(\delta, \beta, \gamma)$, we produce 200 draws of $\epsilon_{ij}$ for each student $i$ and program $j$ as before and compute 1,000,000 almost independent samples of $(y_{i1}, y_{i2}, y_{i3})$.

Item 4 can be computed easily for the Lexicographic model. For the MNL based methods, the desired quantity $\hat{z}_i(j, l)$ has the following form:

$$\hat{z}_i(j, l) = \begin{cases} 1 & \text{if } \mathbb{P}(u_{ij} \geq u_{il}) \geq 0.5, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Define $\bar{u}_{ij} = u_{ij} - \epsilon_{ij}$. This is student $i$'s utility for program $j$ without counting his idiosyncratic taste shock $\epsilon_{ij}$. Define $\bar{u}_{il}$ similarly. Observe that for the MNL model,

$$\mathbb{P}(u_{ij} \geq u_{il}) = \mathbb{E}\left[\frac{\exp(\bar{u}_{ij})}{\exp(\bar{u}_{ij}) + \exp(\bar{u}_{il})} \mid \beta, \delta\right] \tag{12}$$

Hence, we can estimate the above quantity using the 5000 independent samples of model parameters $\beta$ and $\delta$ from the previous calculations for items 1-3. For the MMNL model, the same technique can be applied except that Equation (12) also requires conditioning on $\gamma_i$, and we use the 5000 independent samples of $(\delta, \beta, \gamma)$ from before.

Another benchmark we use in evaluating choice forecasts is random guessing, in which case the choice ranking $y_i$ is assumed to be a uniformly random permutation of options within student $i$'s menu. For the metrics on individual choice, we do not need to explicitly sample but can write explicitly formula for computing the relevant quantities. Let $|S_i|$ be the number of options in student

[27] They are not completely independent because 5000 draws of $(\delta, \beta)$ are shared across students and across each 200 draws of $\epsilon_{ij}$. The completely independent alternative would be to produce one million independent draws of $(\delta, \beta)$ for each student, which is computationally expensive and we think will not change the results.

$i$'s menu.

$$\% \text{ Mistakes in Top } k \text{ Choices} = 1 - k/|S_i|,$$
$$\% \text{ Mistakes in Pairwise Comparisons} = 0.5.$$

For the metrics on distribution of choices, we can compute the top $k$ market shares simply by distribution the market share of each student uniformly among his available options, and averaging over students of each neighborhood. For the joint distribution of top two chocies, we assume that every ordered pair of distinct options is equally likely and apply the linearity of expectations and average across students.

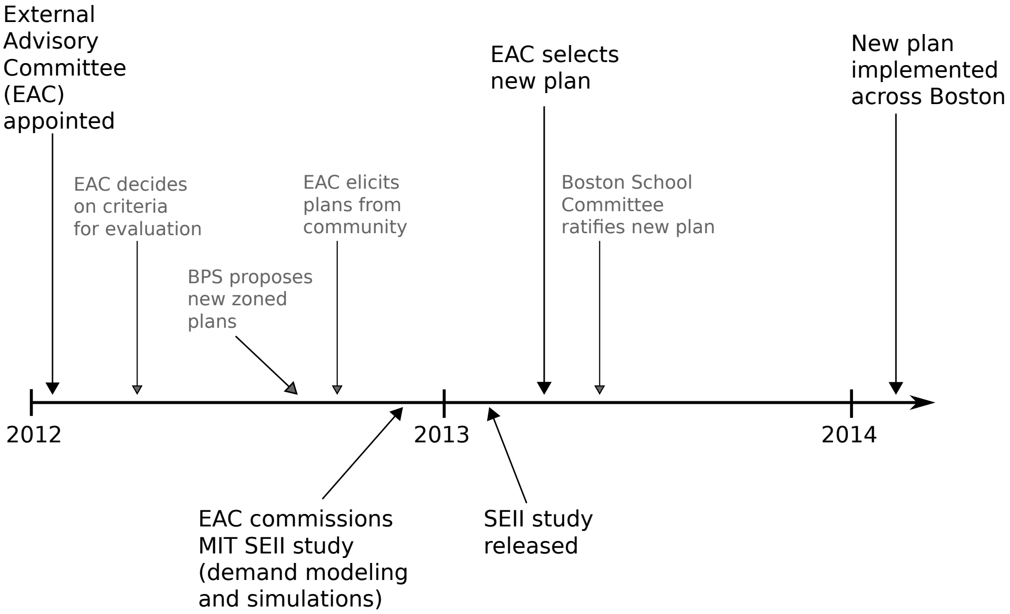

External Advisory Committee (EAC) appointed
EAC decides on criteria for evaluation
BPS proposes new zoned plans
EAC elicits plans from community
EAC selects new plan
Boston School Committee ratifies new plan
New plan implemented across Boston
2012
2013
2014
EAC commissions MIT SEII study (demand modeling and simulations)
SEII study released

Figure 1: Timeline of Policy Reform

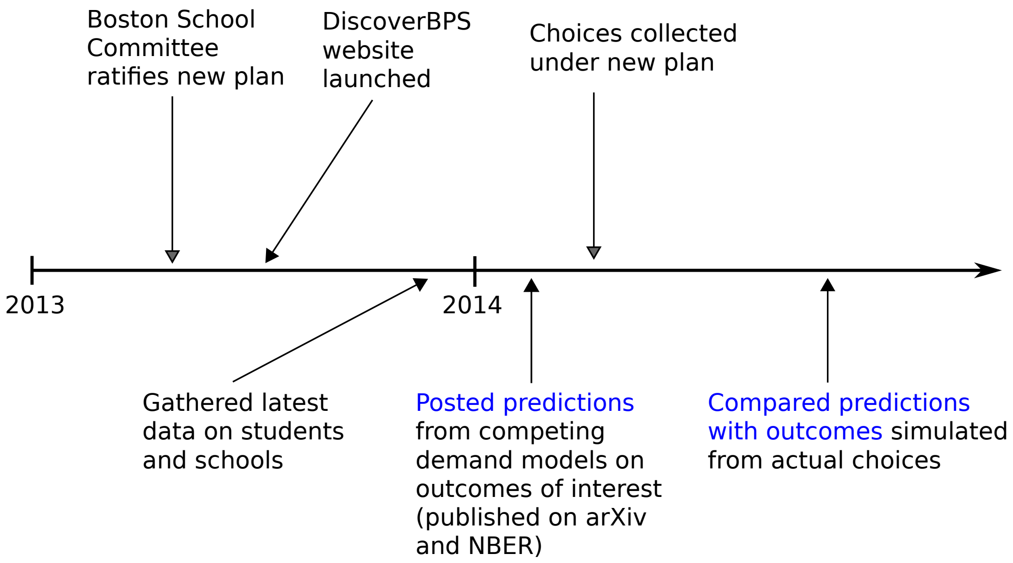
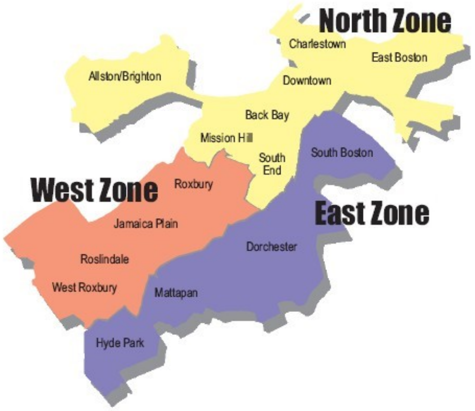
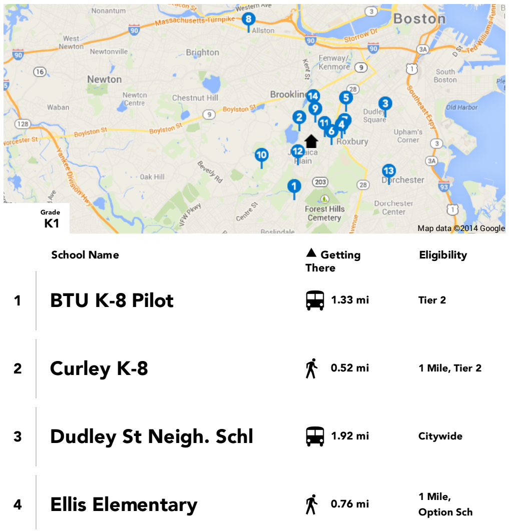

Boston School Committee ratifies new plan
DiscoverBPS website launched
Choices collected under new plan
2013
2014
Gathered latest data on students and schools
Posted predictions from competing demand models on outcomes of interest (published on arXiv and NBER)
Compared predictions with outcomes simulated from actual choices

Figure 2: Overview of Research Design



(A) Before reform (in 2013)

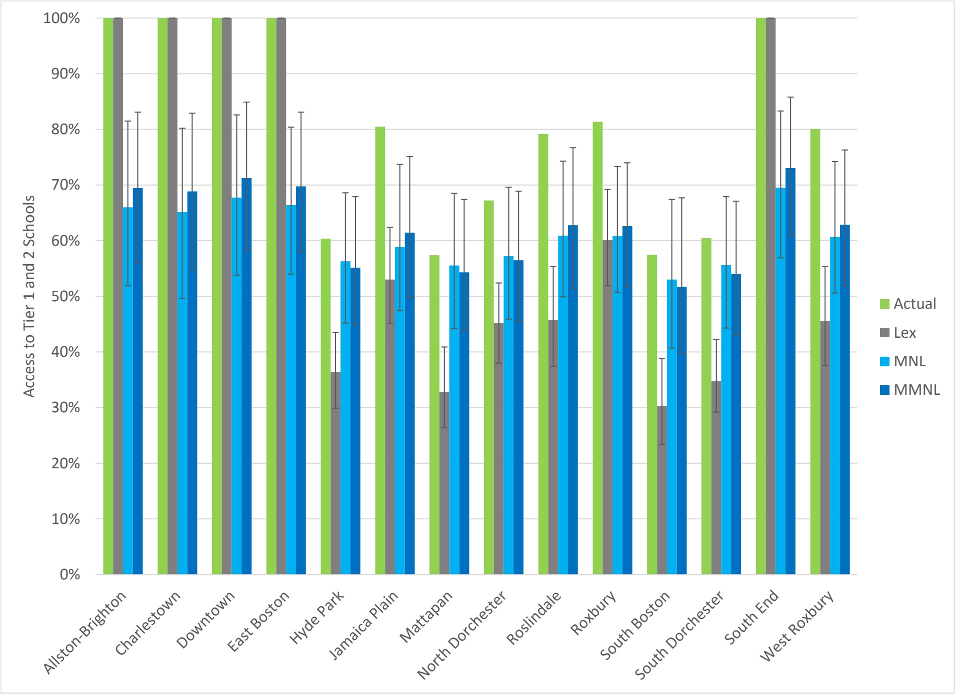


| | School Name | ▲ Getting There | Eligibility |
|---|---|---|---|
| 1 | BTU K-8 Pilot | 1.33 mi | Tier 2 |
| 2 | Curley K-8 | 0.52 mi | 1 Mile, Tier 2 |
| 3 | Dudley St Neigh. Schl | 1.92 mi | Citywide |
| 4 | Ellis Elementary | 0.76 mi | 1 Mile,<br>Option Sch |

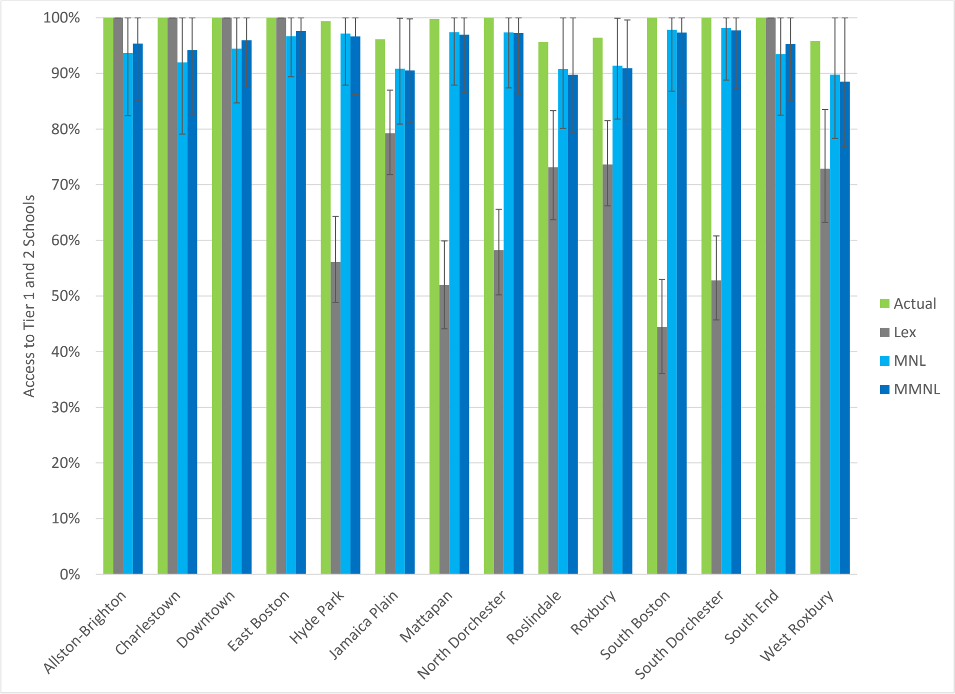
(B) After reform (in 2014)

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.



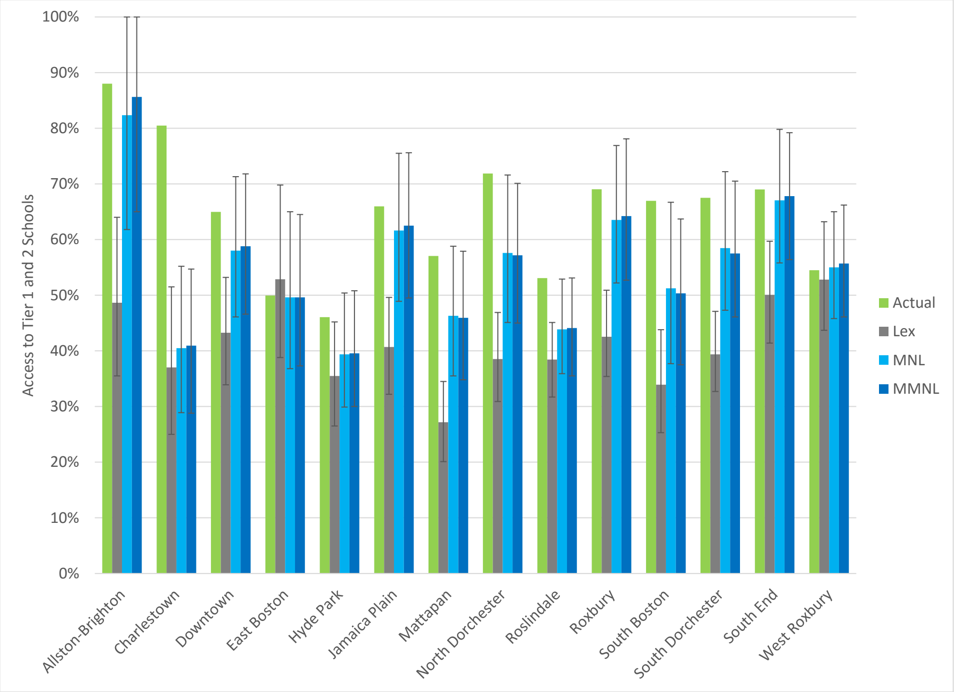
(A) Grade K1



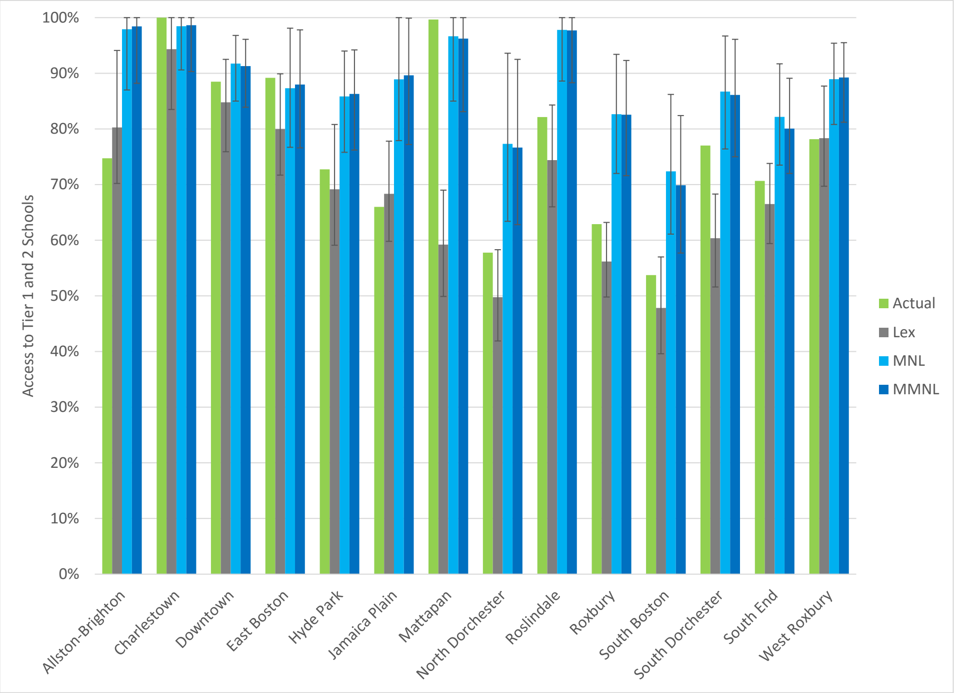
(B) Grade K2

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.



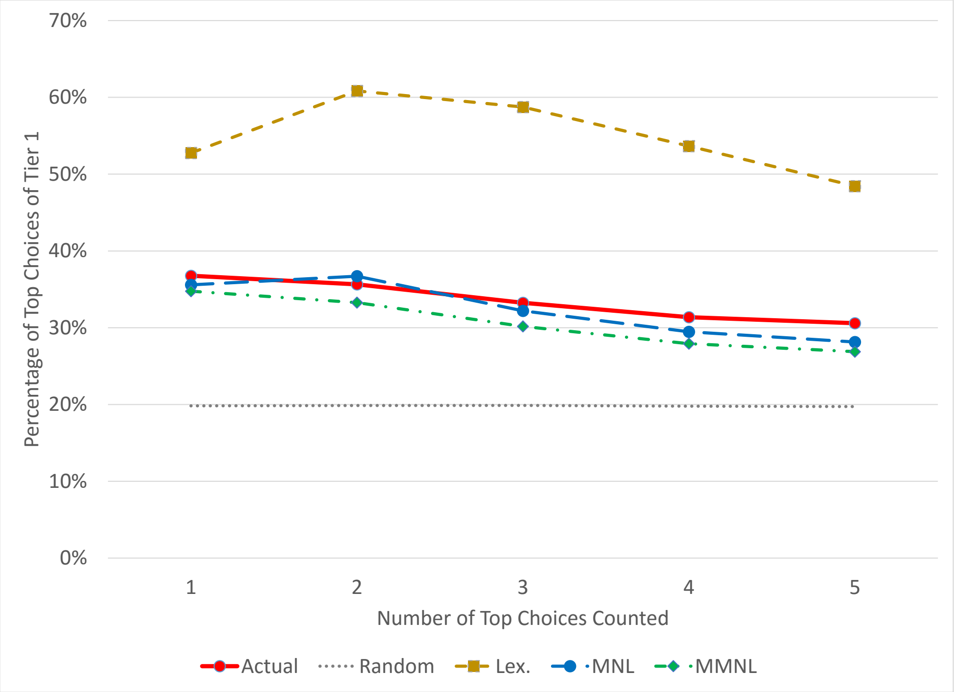
(A) Grade K1



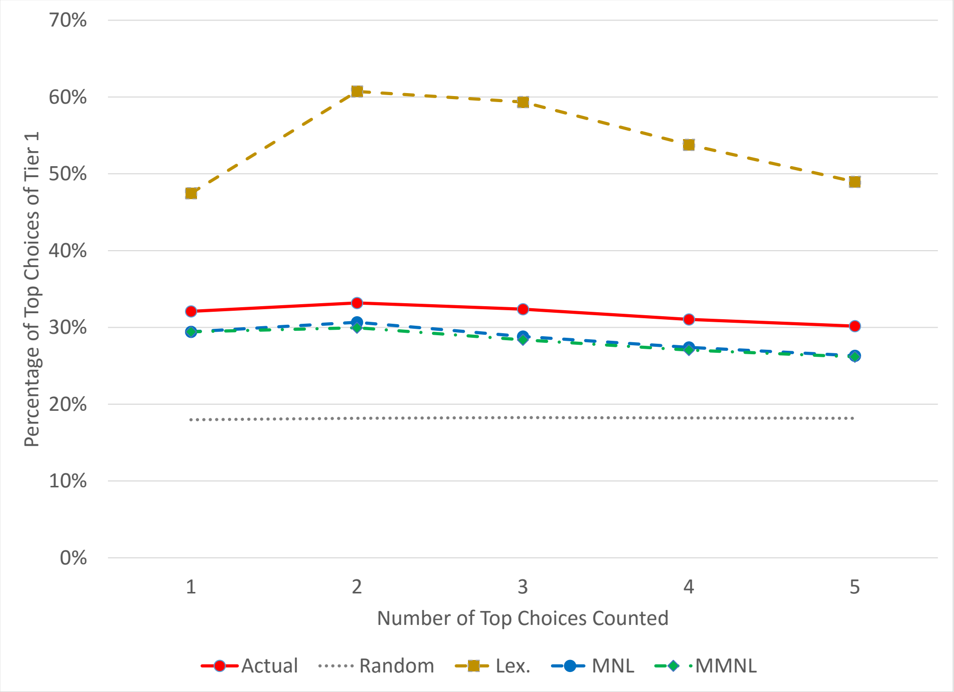
(B) Grade K2

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<br>(1) | Grade K2<br>(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 to Predict One Year 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 Data from Two Years Prior to Predict Choices from One Year Prior**

| | 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 Actual 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 Post-Reform Data**

| | 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 | | | | |
| Old Demand Model | 5% | 6% | 12% | 12% |
| Refit Demand Model | 7% | 7% | 13% | 13% |
| Refit Demand Model + Ranking Length | 3% | 3% | 7% | 7% |
| Sampling Actual Choices Using Applicant Forecast | 8% | | 10% | |
| | B: Distance | | | |
| Original Prediction | 0.19 | 0.19 | 0.13 | 0.14 |
| New Applicants with | | | | |
| Old Demand Model | 0.15 | 0.15 | 0.08 | 0.09 |
| Refit Demand Model | 0.16 | 0.15 | 0.07 | 0.07 |
| Refit Demand Model + Ranking Length | 0.18 | 0.18 | 0.09 | 0.10 |
| Sampling Actual Choices Using Applicant Forecast | 0.08 | | 0.07 | |
| | C: Unassigned | | | |
| Original Prediction | 22 | 21 | 41 | 40 |
| New Applicants with | | | | |
| Old Demand Model | 7 | 7 | 14 | 15 |
| Refit Demand Model | 7 | 7 | 14 | 14 |
| Refit Demand Model + Ranking Length | 6 | 6 | 10 | 10 |
| Sampling Actual Choices Using Applicant Forecast | 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 Old Demand Model 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 Refit Demand Model uses new applicants in 2014 and predicted choices from demand model refit in 2014. New Applicants with Refit Demand Model + 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. Sampling Actual Choices Using Applicant Forecast (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 choices in 2014 with replacement.

**Table 8. Accuracy of Choice Predictions from Refit Demand Models**

| | Demand Model | Grade K1 | | Grade K2 | |
|---|---|---|---|---|---|
| | Fit Using Data | 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.

**Table 9. Reversals of Counterfactual Predictions for Access to Quality in Grade K1 for the Neighborhood Allston-Brighton under Various Simulation Assumptions**

| Choice Model (Year of Fitting) | MNL (2014) | MNL (2013) | | Lexicographic |
|---|---|---|---|---|
| Applicant Pool | 2014 | 2014 | 2013 | 2013 |
| | (1) | (2) | (3) | (4) |
| A. Counterfactual Predictions | | | | |
| Status Quo (3 Zone) | 72.0% | 77.7% | 84.3% | 100% |
| Home Based A | 77.5% | 82.9% | 96.3% | 55.0% |
| Home Based B | 79.1% | 84.5% | 97.8% | 57.0% |
| 6 Zone | 87.3% | 91.3% | 94.5% | 54.3% |
| 9 Zone | 86.5% | 90.7% | 94.2% | 54.4% |
| 10 Zone | 98.4% | 99.8% | 100.0% | 64.9% |
| 11 Zone | 86.4% | 90.5% | 94.2% | 54.2% |
| 23 Zone | 86.7% | 91.3% | 94.6% | 57.6% |
| B. Reversal of Pairwise Comparisons | | | | |
| # of Non-Trivial Comparisons | (Point of reference) | 22 | 22 | 25 |
| # of Reversals of Non-Trivial Comparisons | | 0 | 8 | 14 |
| Percentage of Reversals | | 0% | 36% | 56% |

Notes. In panel A, we report the point predictions for access to quality in grade K1 for the neighborhood Allston-Brighton under various proposed plans. Each row corresponds to a plan proposed during the 2012-2013 Boston student assignment reform, with each plan representing a different set of choice menus and priorities. The first row is the pre-reform status quo, and the second is the plan chosen after the reform. The third row is a variant of the plan in the second row, except with more choices. The remaining plans represent alternative partitioning of Boston into assignment zones. Each column specifies the choice model and the applicant pool used in the simulations. Column 1 uses the multinomial-logit (MNL) choice model fitted from post-reform choices using the post-reform applicant pool in 2014. Column 2 uses the MNL model fitted from pre-reform choices from 2013, but still simulated using the post-reform applicant pool. Column 3 is similar to Column 2 except that it uses the pre-reform applicant pool in 2013. Column 4 uses the lexicographic choice model and the pre-reform applicant pool.

In panel B, we measure how much columns 2 through 4 in panel A are different from panel 1 in terms of the relative rankings of access to quality across plans. Since column 1 is the point of reference, it is left blank in panel B. Consider first the comparison between columns 1 and 2 of panel A, which are reported in column 2 of panel B. Since there are 8 plans, there are 28 comparisons. Each comparison corresponds to a pair of rows from panel A, and we call the comparison "trivial" if the access to quality predictions in these two rows are within an addititve difference of 1.0% of one another in both columns 1 and 2. For example, the comparison between the 11 and 23 Zone plans is trivial, but the comparison between the Status Quo and the Home Based A plan is not. Row 1 in column 2 of panel B reports the number of non-trivial comparisons between columns 1 and 2 of panel A. Row 2 of panel B reports the number of such comparisons that is reversed, which means that the columns differ in which plan results in a higher access to quality. For example, for columns 1 and 3 of panel A, the comparison between the Home Based A plan and the 23 zone plan is reversed, but between the Home Based A plan and the Status Quo is not. Row 3 of panel C reports the ratio between the previous two rows expressed as a percentage. As a benchmark, if the columns agree exactly on the relative rankings across plans, then the percentage of reversals should be 0. On the other hand, random guessing results in an expected 50% of reversals. Thus, we see that column 2 of panel A is a perfect proxy for column 1 in terms of relative ranking across plans, while column 3 results in 36% of reversals, which is still better than random guessing. However, column 4 results performs worse than random guessing in this case.

**Table 10. Percentage of Reversals of Counterfactual Predictions Under Alternative Simulation Assumptions (Averaged Across Neighborhoods)**

| | Choice Model (Year of Fitting) | MNL (2014) | MNL (2013) | | Lexicographic |
|---|---|---|---|---|---|
| | Applicant Pool | 2014 | 2014 | 2013 | 2013 |
| | | (1) | (2) | (3) | (4) |
| Access to Quality | | | | | |
| | K1 | | 13% | 17% | 32% |
| | K2 | | 4% | 23% | 35% |
| Distance | | | | | |
| | K1 | (Point of reference) | 2% | 7% | 18% |
| | K2 | | 1% | 11% | 24% |
| Unassigned | | | | | |
| | K1 | | 5% | 20% | 44% |
| | K2 | | 5% | 21% | 38% |
| Overall | | | 5% | 16% | 32% |

Notes: This table reports the aggregate result of the analysis in Table 9 on the percentage of reversals of pairwise comparisons of counterfactual predictions, when averaged across the fourteen neighborhoods and performed for each of the three equilibrium moments of interest. See the notes for Table 9 for description of the columns as well as for the eight proposed plans compared. See the notes for Table 3 for descriptions of the three moments of interest. The numbers in the first row correspond to conducting an analogous analysis as that in the last row of panel B of Table 9 for all fourteen neighborhoods instead of just Allston-Brighton, and reporting the average across neighborhoods. The second row is similar, except for grade K2 instead of K1. The next four rows are for different moments of interest, but the analysis is analogous, except for the following difference: recall from the notes of Table 9 the definition of "non-trivial" comparisons of a given pair of plans, and that the threshold for a non-trivial comparison is set to an additive difference of 1.0% for access to quality. For distance, this threshold is set to 0.01 miles. For unassigned, this is set to 0.5 students/neighborhood.

The last row reports the unweighted average of the first six rows, and corresponds to an aggregate measure of how much relative rankings of counterfactual predictions are different across simulation assumptions. As can be seen, conditional on using on using the post-reform applicant pool and the MNL model, whether or not one fits the MNL model from post-reform or pre-reform choices only reverses pairwise comparisons of counterfactual predictions 5% of the time. However, if one had used pre-reform applicant pool with the same MNL model, then the predictions would be reversed 16% of the time. If one used the lexicographic choice model instead of the MNL and the pre-reform applicant pool, then the predictions would be reversed 32% of time. As a benchmark, random guessing would result in about 50% of the predictions reversed.

# References

Abdulkadiroğlu, A., N. Agarwal, and P. Pathak (2015): "The Welfare Effects of Coordinated School Assignment: Evidence from the NYC High School Match," NBER Working Paper, 21046.

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?," NBER Working Paper, 23912.

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.

Azevedo, E. M., and J. D. Leshno (2016): "A Supply and Demand Framework for Two-sided Matching Markets," *Journal of Political Economy*, 124(5), 1235–1268.

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., F. Reid, A. Talvitie, M. Johnson, and Associates (1979): "Overview and Summary: Urban Travel Demand Forecasting Project," *Urban Travel Demand Forecasting Project*, Final Report, Vol. I. Institute of Transportation Studies, University of California Berkeley.

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.

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/.

Misra, S., and H. Nair (2011): "A Structural Model of Sales-Force Compensation Dynamics: Estimation and Field Implementation," *Quantitative Marketing and Economics*, 9(3), 211–225.

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.

——— (2013): "No Division Required in This School Problem," *New York Times*, March 12.

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. (2014): "The Demand for Effective Charter Schools," NBER Working Paper 20640.

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 student | 0.606*** | 0.672*** | 0.802*** | 0.610*** | 0.967*** | 0.989*** |
| | (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 | Std. Error | Actual |
|---|---|---|---|---|
| | | (1) | (2) | (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 block group income refers to the median household income of the census block group in which the student resides, based on the 2010 census.