Algorithmic Transparency and Accountability through Crowdsourcing: A Study of the NYC School Admission Lottery

Amélie Marian
amelie.marian@rutgers.edu
Computer Science Department
Rutgers, the State University of New Jersey
New Jersey, USA

ABSTRACT

Algorithms are used to aid decision-making for a wide range of public policy decisions. Yet, the details of the algorithmic processes and how to interact with their systems are often inadequately communicated to stakeholders, leaving them frustrated and distrusting of the outcomes of the decisions. Transparency and accountability are critical prerequisites for building trust in the results of decisions and guaranteeing fair and equitable outcomes. Unfortunately, organizations and agencies do not have strong incentives to explain and clarify their decision processes; however, stakeholders are not powerless and can strategically combine their efforts to push for more transparency.

In this paper, I discuss the results and lessons learned from such an effort: a parent-led crowdsourcing campaign to increase transparency in the New York City school admission process. NYC famously uses a deferred-acceptance matching algorithm to assign students to schools, but families are given very little, and often wrong, information on the mechanisms of the system in which they have to participate. Furthermore, the odds of matching to specific schools depend on a complex set of priority rules and tie-breaking random (lottery) numbers, whose impact on the outcome is not made clear to students and their families, resulting in many "wasted choices" on students' ranked lists and a high rate of unmatched students. Using the results of a crowdsourced survey of school application results, I was able to explain how random tie-breakers factored in the admission, adding clarity and transparency to the process. The results highlighted several issues and inefficiencies in the match and made the case for the need for more accountability and verification in the system.

CCS CONCEPTS

• **Social and professional topics** → *Government technology policy*;
• **Human-centered computing** → **User studies**.

KEYWORDS

school matching, crowdsourcing, accountability, transparency

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06…$15.00
https://doi.org/10.1145/3593013.3594009

ACM Reference Format:
Amélie Marian. 2023. Algorithmic Transparency and Accountability through Crowdsourcing: A Study of the NYC School Admission Lottery. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3593013.3594009

1 INTRODUCTION

Automated and algorithmic decision systems are used extensively in public policy. These systems are often opaque, not publicly audited, and their details are hidden through contracts with third-party vendors despite being funded by taxpayers and used for public good allocations and decisions. As more and more government agencies are relying on automated systems to allocate resources and make decisions that have significant impacts on people's lives, many advocates and scholars are raising concerns about the lack of transparency of decisions systems [12, 28, 30, 32]. Without transparent and explainable processes, it is not possible to verify whether the outcomes satisfy ethical and fairness constraints. Without accountability, there can be no trust in the decisions produced by these systems.

An example of the challenges of providing transparent and accountable decision systems can be shown from the experience of the Automated Decision Systems Task Force (ADS Task Force) which was established by a New York City Council law in 2018. The task force, the first of its kind, was tasked with recommending a process for reviewing the City's use of automated decision systems (algorithms). Because many City agencies and offices use algorithms to aid their decision-making, and because automated decision systems are becoming more prevalent in all fields, the City's goal was to examine ways to ensure these systems align with the goal of making New York City a fairer and more equitable place. Unfortunately, the ADS Task Force did not succeed. Members of the Task Force grew frustrated with the lack of transparency from the city agencies. After 18 months, they had yet to see an actual automated decision system[14, 20].

This culture of secrecy is endemic in many government agencies. In NYC, one automated decision system that impacts the lives of many families is the school admission process run by the city's Department of Education (DOE). Each year approximately 150,000 rising middle- and high-schoolers are being assigned to schools using a deferred acceptance (DA) matching algorithm [1]. The algorithm matches students to schools based on students' preferences, admission priorities, and the schools' admission-ranked lists. Transparency in such a scenario is critical, without a transparent and explainable process, families are losing trust in the system.

In this paper, I relate my experience working with groups of parents to increase transparency and force accountability in the NYC school admission process through the use of crowdsourced admission data. I first describe how NYC assigns students to schools, and recent changes in policy (Section 2). I discuss how this work shed clarity on the "lottery," random numbers generated to prioritize students in the match (Section 3). The crowdsourced admission results highlighted some unintended effects and perverse incentives of the process and helped push the NYC DOE to be more accountable to families (Section 4). I reflect on the effects of the work on the information that is now shared with families, and on the efforts that are still needed to improve outcomes for underserved students (Section 5), discuss how this effort relates to existing work (Section 6), and conclude (Section 7).

2 BACKGROUND

2.1 The NYC School Matching Algorithm

NYC middle- and high-school admissions use a matching algorithm [1] similar to the stable marriage algorithm designed by Gale and Shapley [16] and used for the medical school resident match. The NYC school matching process consists of three parts: the schools rank students based on some desired criteria (academic, geographic, demographic), the students list schools in order of preference, and the matching algorithm, handled by a clearinghouse and designed to optimize some notion of utility, produces the rankings. The NYC school matching algorithm is designed to optimize students' choices while guaranteeing stable matches. An important property of the matching is its strategy-proofness (on the students' side) so that the best strategy for families is always to list their choices in their true order of preference.

In NYC, the matching algorithm factors in various policy choices and admission priorities, such as set-asides for low-income students, geographic priorities, seats reserved for students who qualify for special education services, or priorities for continuing students or their siblings. The school rankings of students have considered these priorities as well as some measures of the academic or artistic preparation of students.

This "screening" of students has been a source of strong criticism due to the belief that it leads to a lack of diversity in the city schools [19]. Typically, schools were ranking students using a weighted-sum function of their grades, state test scores, and absences. However, the actual ranking functions used by the schools were often coarse and would lead to unintended outcomes [15]; some schools for instance ended up, unknowingly, basing most of their rankings on the number of days a student was absent or late, disproportionally penalizing students for illnesses. In addition, school choice systems are often complicated to navigate for families [27]. As a result, minority families often do not apply to selective schools, leading to disproportionate representation in popular schools. In addition, school principals do not understand the implications of their school admission criteria on class composition, making the system clearer and more transparent would increase equity [10].

2.2 Admission and policy changes

Several changes to middle- and high-school admissions were implemented starting in the 2021 admission season. Because of the COVID-19 pandemic, 2020 state tests were canceled and many schools did not assign grades. This led the NYC Department of Education (DOE) to move to a lottery-based approach for admissions to all middle schools and some selective high schools that were previously using academic screens and auditions [7]. In addition, several equity-motivated modifications to the process were included, such as the removal of some geographic admission priorities, and the addition of set-asides for low-income students in many popular schools.

Algorithms are fickle; a seemingly small change to a process can trigger unanticipated changes in the outcome of the algorithm. The changes in admission policies meant that the odds of admissions at several popular schools were significantly different from what they had been in the past. Unfortunately, the changes, and their impacts on the chances of admissions at different schools, were not communicated well by the DOE. This meant that families, who often were not aware of the changes, or did not understand their impacts, were basing their ranked choice decisions on expected match outcomes from the past. This resulted in a large increase in the number of applicants not getting any of their 12 choices. In 2021, 7% of high school applicants (close to 5,000 students) were unmatched citywide, with up to 18% of students in a Manhattan district, a 125% increase compared to previous years (information received via an NYS Freedom of Information Law request), unmatched student data for the 2022 admission cycle was similar [11].

3 INCREASING TRANSPARENCY BY PROVIDING FAMILIES WITH THEIR RANDOM NUMBERS

Lotteries are not new in NYC school admissions, the DOE has used them for years for pre-K and Kindergarten admissions, and to break ties when there are more qualified applicants than seats at a school. However, the increased use of lotteries and the importance of the random number in admission odds following the changes in admissions in 2021 has raised questions from parents about the lottery process, including how the numbers are drawn, the odds of admission to preferred schools, and whether they could have access to their student's lottery number.

3.1 One Number per Student

Each student is assigned a single random lottery number that is used to determine their admission priority for schools that admit their students through total or partial lotteries, and to break ties for schools that admit their students through priority groupings, or batches, if there are more students in a batch desiring to gain admission than there are available seats at the school for that batch.

The decision to use the same lottery number for all the schools, rather than having a separate lottery number for each school, is one that often puzzles and infuriates parents who believe the system is unfair to students with an unlucky draw and that students would have a better chance if they could draw one number per school. In fact, the literature shows that using the same number for all schools does not penalize students [4, 26], instead, it slightly increases their

chances of matching to their top choice [2]. In a single number system, students with a good lottery number are more likely to be assigned to their preferred school, but if schools hold separate lotteries, to get their top school, students have to receive a (lower odds) good lottery number at their preferred school. The chances of not getting an offer are roughly the same.

3.2 Gaining Access to the Generated Lottery Number

Incredibly, at first, the DOE declined to provide families with any information on their lottery numbers, or much details about the process. The DOE first told parents that the numbers were "truly random," but could not be shared with families because they consisted of long strings of numbers and letters that they —the families— could not understand! In addition, the DOE claimed that the numbers would not give information to families as to their relative chances in their preferred schools, as these chances depend on the choices of other applicants. Students were receiving school assignments, largely based on the results of an opaque lottery process using a randomly-generated number that they, or their families, had no way to access for verification or transparency.

A parent group had to launch a campaign asking parents to request their student's lottery numbers[22] under New York State's Freedom of Information Law (FOIL)[1] —parents are legally entitled to see any information that is used to decide their child's admission to public schools— for the DOE to finally relent and agree to provide the lottery numbers, at first upon email request *after the results of the Spring 2021 match*. As the demand from parents increased in the subsequent admission season, the lottery number was made available to *all middle- and high-school students* on the DOE admission portal *before applications were due*, a win for transparency and as we will see below, an important piece of information to help families navigate the process.

3.3 Understanding the Lottery Number

When the DOE finally provided families with their random number, it did so with scant contextual information. Most received, without additional information, a lottery number that looked like this:

`6ba829b3-fa99-4752-a931-2119fb0c1fea`

While this is indeed a "long string of numbers and letters" as described by the DOE, it is not so complicated that parents could not understand it. Anyone with a computer science background will easily recognize the string as a 32-characters hexadecimal number. Most families are not familiar with hexadecimal numbers, but they are quite easy to explain: hexadecimals are base 16 numbers (they use digits 0–9, then a-f) that are often used in programming because computers encode everything in binary (bits are base 2: 0 or 1), and 16 is a power of 2. A hexadecimal character can be represented in 4 bits.

We can only guess why the lottery numbers are encoded as long hexadecimals; the DOE has declined to provide more information on the choice. However, a longer glance at the numbers shows that they are in a format called UUID (Universally Unique Identifiers), Version 4 (the version can be identified by looking at the first character of the third block of characters — the 13th character). UUID V4 are used to generate random unique identifiers, it makes sense for the programmers of the NYC DOE lottery to have used an existing — and well-tested — random number generator library function that creates uniformly distributed UUID V4 numbers.

Once the format of the lottery numbers was clear, understanding how they were processed was the next step. Once again the DOE at first declined to provide such information, but using the crowdsourced data from the 2021 admission cycle (see Section 4), it was possible to conclude that the UUID numbers were sorted lexicographically: the numbers are compared left to right, in increasing order, from 0 to f (0–9 then a-f). This means that the first character is enough to give a rough idea of how good a number is: a lottery number that starts with 0 is in the first $1/16^{th}$ (6.25%), one that starts with f in the last $1/16^{th}$. To differentiate further, we can look at the first two characters: a lottery number that starts with '00' is in the first 256th (0.4%), '01' in the 2nd 256th, and so on. The first two characters are sufficient to identify where your number is expected to stand in comparison to other numbers, with a 0.4% precision. Note that these are the theoretical percentiles, i.e., the expected percentile based on a uniform distribution of random numbers.

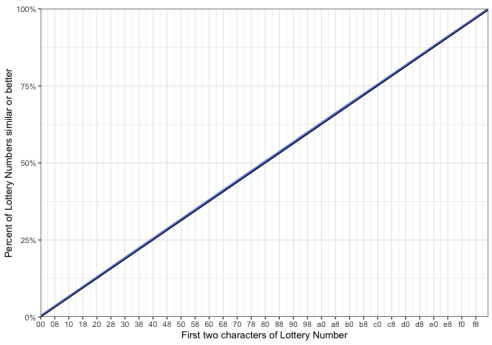
The DOE has declined to provide applicants with percentile information about their lottery numbers, stating:

> *"Until the application has closed and we know the full pool of applicants, the relative rank of this lottery number cannot be calculated."*

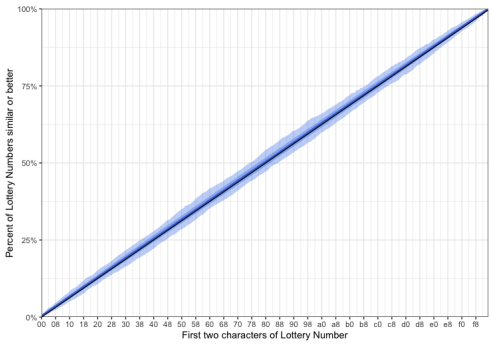
This is technically correct. The random numbers are drawn from a uniform distribution and random samples may differ from the expected values. However, sample percentiles are known to be asymptotically normally distributed around theoretical percentiles, with the variance depending on the sample size [29]. For large samples (such as lottery numbers for all applicants to high- or middle-schools), this means the sample percentiles will be very close to that of the theoretical distribution given above, as shown in Figure 1a, which represents the distribution of the first two characters for a simulation over 1,000 samples of 60,000 lottery numbers (similar to a citywide cohort of high- or middle-school applicants). To illustrate how the empirical distribution of lottery numbers may differ from the theoretical one, I report on the simulation of 1,000 samples for 4 different sample sizes: 60,000, 2,500, 600, and 100, and plotted the theoretical percentiles (black line), median sample percentile (dark blue line), 10–90 percentile range of sample percentiles (where 80% of sample percentiles will fall, in medium blue), and the full range of observed sample percentiles (in light blue) in Figure 1. The x-axis represents the first two characters of the lottery numbers, and the y-axis is the observed percent of lottery numbers that are similar or better in the simulated sample.

The sample sizes were chosen to represent various comparison groups: 60,000 for a cohort of citywide students (there are typically between 60,000 and 80,000 students in each grade cohort), 2,500 represents a large sample of applicants (e.g., all applicants from a district, or to a high-demand high school — some high schools have more than 5,000 applicants), 600 and 100 represent medium and small samples of applicants (e.g., average- or low- demand high- or

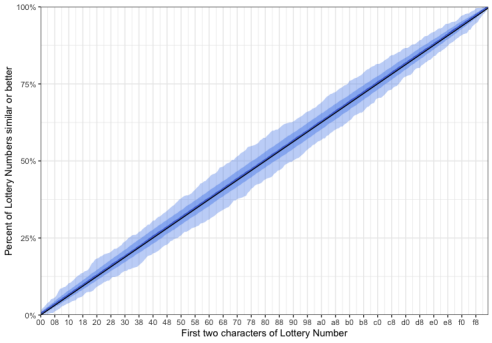
[1] New York State's Freedom of Information Law (FOIL) is similar to the Federal Freedom of Information Act (FOIA).



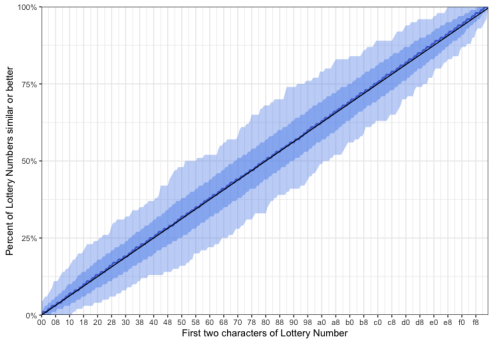
(a) Distribution of the first two characters of a sample of 60,000 Lottery Numbers



(b) Distribution of the first two characters of a sample of 2,500 Lottery Numbers



(c) Distribution of the first two characters of a sample of 600 Lottery Numbers



(d) Distribution of the first two characters of a sample of 100 Lottery Numbers

Figure 1: First Two Characters of Lottery Numbers to Percentile Conversion — Empirical Data with Different Sample Sizes, 1,000 simulations

middle-school). As discussed above, the sample percentiles for the 60,000 sample (Figure 1a) track very closely with the theoretical percentiles. With a sample size that is not as large, 2,500 numbers (Figure 1b) we start seeing some variations, but the actual percentile will be within 1% of the theoretical percentage 80% of the time (medium blue ribbon). Smaller sample sizes will have more variance (Figure 1c). In those cases, the percentiles given above are more approximate, but can still give you a good idea of where a lottery number stands compared to others. For instance, for a school with 600 applicants, a lottery number starting with '3a' is expected to be in the 23rd percentile. In over 1,000 simulations, that lottery number fell between the 21st and 25th percentile 80% of the time, but in some extremely skewed cases could present in the 18th or 28th percentile. Finally, the distribution of small sample percentiles is obviously more skewed as more variations are possible. This is due to large variability in small samples, and can be shown in Figure 1d. This means that a lottery number will have more variability in how it compares to that of other applicants in a school with a small number of applicants. However, this is unlikely to have much impact as schools with low numbers of applicants typically are not selective.

Figure 1d also illustrates how families can easily come to distrust the process. Parents often believe that the system is rigged because they know a group of applicants with unusually bad (or good) lottery numbers, as small samples are known to be likely to have outlier results according to the Law of Small Numbers [31].

The truth is that families don't care much about the actual numbers or even the exact percentile where it falls; rather they want an idea of their student's chances and guarantees that the numbers were generated fairly. The 32-character numbers look unnecessarily complex; most of the characters are just noise and have no impact on the student's admission outcome, only the first characters will ever be used to compare students in the same cohort. The use of overly long and opaque numbers are raising more questions than they answer: parents on internet boards are convinced that the DOE is tipping the scales by favoring students from some schools, or demographics, over others; that the numbers are encoding all type of information used in the match. They use anecdotal data to confirm their fears. The lack of transparency is the main cause of mistrust.

For a decision process to be transparent and accountable, it needs to be simply explainable. If the DOE had clearly stated how the numbers were generated (maybe sharing which library function is used) and explained how the numbers are processed from the beginning, families would have more trust in the system.

3.4 Leveraging Knowledge of Lottery Numbers

It seems counter-intuitive to advocate for lottery numbers to be released before the results of the school match are given to families. A lottery seems fair at face value, couldn't families just wait for the outcome, and possibly get their lottery number then? Wouldn't knowing the lottery number in advance create wrong incentives and undesirable strategic behavior? Can families just list the schools in which they are interested in their true order of preference and see how things pan out?

The NYC school admission matching process is not a pure lottery: families are ranking their choices and admission priorities are in place. While the original stable matching algorithm was designed to be strategy-proof [16], the NYC School matching algorithm itself is not completely strategy-proof [1, 2]. Because the number of schools a student can rank is limited to 12, there is a possibility that a suboptimal selection of 12 schools could result in a student not being matched to any of their choices (in which case the DOE assigns them to a school with available seats after the match). While the best strategy is always to order the 12 schools in a student's ranked list in their true order of preference, there is some strategizing when it comes to deciding which schools to include in the list of 12 choices.

As the changes in admissions that started in 2021 have made lottery numbers the primary factor for admissions to middle-schools and most high-schools, knowing the lottery number in advance of the application becomes important for several reasons:

- **Researching schools is time- and resource-consuming.** Parents and students have to attend numerous open houses, learn about the course offerings of the different schools, understand the schools' admission priorities, figure out if the school requires additional admission material (auditions, essays), for at least 12 schools, likely more, to build their list of 12 choices. On the school side, hosting multiple open houses requires a significant time investment from school staff. Some popular schools have been known to host several tours for only a handful of seats available for non-continuing students. Knowing their lottery numbers allows families to focus on likely matches and not waste time researching and comparing highly-selective schools if their random number is unlikely to give them access to these schools.
- **Being unmatched is to be avoided as much as possible.** As discussed in Section 2.2, a large number of high-school applicants were left without a match in 2021 and 2022 [11]. Such outcomes are devastating for families: students who do not have another high-school placement (e.g., Specialized –or magnet– HS) are assigned by the DOE to a school that rarely matches their expectations. In fact, the DOE's strategy to assign unmatched students to schools has created many issues: students being assigned to career and technical schools for professions in which they had no interest, hearing students with no ties to the deaf community being assigned to an ASL school,...
- **Managing expectations.** Lastly, and perhaps more importantly, we are talking about children and teenagers on whom the process can be extremely stressful. Receiving the news that you did not match to a school that you had your heart set on while your friends did, or worst that you did not get any match, can be hard, especially for middle-school applicants who are only 10 or 11 years old; knowing the lottery numbers in advance can help parents steer their children's hopes towards schools to which they have a reasonable chance of getting matched.

4 PRESSING FOR ACCOUNTABILITY VIA CROWDSOURCING

Knowing whether they have a "good" or "bad" lottery number is of limited use to families without added information about which lottery numbers are good enough to gain an offer to their desired schools. In fact, one of the reasons for the DOE to originally refuse to give lottery numbers was that, on their own, the lottery numbers were not very informative because the chances of gaining admission at a school depend on the lottery numbers of the other students applying to the school.

As we saw in Section 3.3, this is not quite correct: statistically, the distribution of the lottery numbers of all applicants to any school will follow a distribution close to that of the citywide pool (Figure 1). The chances of gaining admission to a lottery-based school depend on *the number of available seats, the number of applicants, and how they ranked the school.* The interplay between students' choice rankings and selectivity of schools is something that has not been made clear by the DOE: a school that is ranked 1^{st} by a set of applicants will have different odds of admission from one that is ranked 12^{th} by the same set of applicants, due to the mechanism of the matching algorithm. How far each school will go down their list is not as simple as dividing the number of seats by the number of applicants, and depends on how applicants ranked the school. Unfortunately, the only information that is made available by the DOE is the number of applicants per seat for each school, regardless of the position in which the applicants rank their schools. Furthermore, the number of applicants for each priority group is not clearly communicated.[2]

4.1 Historical School Cutoffs

Past information can help families estimate their chances and plan their ranked list of choices. Knowing whether a school you are interested in admitted everyone who listed them in the past, or only admitted the first 5% of their applicant list is valuable information that can help families plan. Of course, these school cutoffs are not a guarantee as applicants' choice patterns and admission rules and priorities may change year to year, but they give valuable information for families to manage their expectations. In addition, transparency about school selectivity and cutoffs is necessary for accountability and trust.

Once again, the DOE has declined to make historical cutoff information available. More concerning, a FOIL request filed in September 2021 for all schools cutoffs was declined because:

> *"[...]a compilation of such data does not currently exist, and responding to your request would involve more*

[2]As discussed in Section 5, as a result of the crowdsourced work presented in this paper, and of continued pressure from elected parents and parent advocacy groups, the DOE has changed the way it reports applicants per seat for the 2023 admission season to make it clearer. The school admission cutoffs have not been disclosed.

> *than a simple extraction of data from a single computer storage system. Rather, it would require matching records across more than one computer storage system, and require extensive coding and programming, which the FOIL does not obligate an agency to do.[...] "*

This claim is concerning for many reasons: by necessity, the matching process has all the data in one system, and it should be easy to retrieve. Furthermore, school cutoffs should be logged as a by-product of the matching verification process. How can the DOE (or its third-party vendor) ensure there were no errors in the match without such verification? Errors in the match have been reported in past years [6, 33], yet the DOE has always declined to provide any information on their validation and verification processes.[3] But what is worrisome is that the FOIL denial implies that the DOE is not compiling the cutoff for every school. How can the DOE train its admission staff and guidance counselors to offer accurate and useful advice to students on their chances of admission if they do not have that information? How can the DOE assess the impacts of changes in admission policy without understanding their outcomes on admission odds?

Families are not powerless in the process. They have information about their own students' applications and match and have pushed for more transparency by comparing notes on social media groups and through advocacy groups and elected parent leaders. I worked with these groups to coordinate a crowdsourcing experiment to identify school cutoffs and to perform some verification on the outcome of the match. I report on this effort in the rest of this paper.

4.2 Crowdsourcing Survey Methodology

To identify the cutoffs to gain admissions to various middle- and high-schools, I used a crowdsourcing survey where families were asked to enter their lottery number (only the first two characters to preserve anonymity), the school to which they matched, and the schools they ranked higher than their match. They were also asked to enter information on which priority group their student qualified for. The survey was done via Google Forms, answers were kept completely anonymous, and no identifying (or re-identifying) information was requested. The survey was approved by Rutgers' Institutional Review Board. Participants were recruited through messages on social media parent groups.

The survey result data was used to identify the "worst" lottery number that received an offer to the school, and the "best" that did not. The corresponding percentile in the theoretical distribution was used to get a lower and upper bound of the odds of matching to a given school. Note that these odds represent chances of matching to a given school *or to a school ranked higher on the student list, as the matching algorithm will only consider a student for a school if they haven't matched to a school they ranked higher.*

The full results of the surveys were published online [22], allowing families applying in later years to have a better idea of the odds of admissions at their preferred schools and to adjust their ranked list accordingly. For the 2021 admission season, the survey received 136 (107 with lottery numbers) answers from high-school applicants and 125 (70 with lottery numbers) answers from middle-school applicants. Note that in 2021 the DOE did not provide families with their lottery number, accessing it required parents to file a FOIL request. In 2022, the DOE started including the lottery number in the students' admission portal, but it did so only a week before applications were due. For the 2022 admission season, the survey received 253 answers from high-school applicants and 102 answers from middle-school applicants.

After a manual data cleaning step to remove obvious errors in data entry, the highest (worst) lottery number of all students who received an offer to a school was extracted: this identifies a lower bound for the school cutoff. To identify the higher bound, the process is similar but needs to consider only students who did not receive an offer to a school they ranked higher than the one to which they matched.

Survey participants were asked whether they qualify for FRL (free and reduced lunch, the DOE measure of low-income status) diversity in admissions (DIA) set-asides, and whether their student had a student with disability (SWD) designation. They were given the choice to decline to answer these questions, in which case their answers were processed in the most conservative way to determine cutoffs: for instance, for a student whose FRL status is unknown, their acceptance to a school can only be used to increase the (less restrictive) DIA cutoff lower bound, while their rejection from a school can only be used to lower the (more restrictive) non-DIA cutoff upper-bound. Note that SWD are handled through a separate process, and different designations will have access to different programs. The data analysis does not differentiate among SWD statuses.

It is worth noting that families get no direct benefit from participating in the survey as information is collected *after* admission placements are communicated to families. Anecdotal information from parent groups' discussions shows that participants are motivated by wanting to understand the process better and helping future applicants. After the 2023 admission season, in which lottery numbers were made available to all applicants from the beginning, families requested that the survey be run again, as they had found the previous years' results useful to create their choice lists. As of the writing of this article (May 2023), the high-school survey for the 2023 admission season has received 647 answers.

4.3 Limitations of the Study

The crowdsourcing survey's goal is to gain a better understanding of the admission odds and to potentially identify issues with the process. It is not meant to be exhaustive, the recruitment of participants was done through online parent groups and mailing list and as a consequence skews toward families that speak English and are well-informed. Survey answers are not representative of the whole applicant population: respondents are clustered in some geographical areas, skew higher income and higher-achieving than the DOE student population. For instance, in 2022, some high schools were using a coarse academic priority grouping to select students with a minimum GPA. All answers to the HS survey reported their

[3] The errors were only caught because well-informed parents understood how the matching should work and were willing to share information with each other. They coincided with the switch to a new third-party vendor in charge of the matching process. The DOE declined to explain what checks and quality controls, if any, were imposed on the new vendor.

	2021 Survey	2022 Survey	2023 Survey	Actual Applicant Pool
Number of Applicants	136	253	647	[70,000-75,000]
Percentage of Low-Income	7%	7%	9%	≈72%
Percentage of Students with Disabilities	3.5%	5.5%	11.5%	≈20%
Percentage of Unmatched Students	22%	8%	1.5%	7% (2023 data not available)
Number of Schools	76	106	100	≈436

Table 1: Survey Data Statistics

student being placed in academic Group 1 for screened school admissions (63% of students citywide were placed in Group 1, from data received by CCHS, the Citywide Council on High Schools –an elected parent body). This does not impact the correctness of the cutoff information derived from the surveys but does impact the completeness of the information. As such, the results are bound to be incomplete and to miss a large number of schools.

Data, such as priority group eligibility, is self-reported by parents, and errors in data entry are possible. The data is sparse, data is missing for many schools, and it is skewed toward some schools and districts (where the survey was shared). In many cases, the cutoff could not be narrowed precisely, so the results only identified a range of lottery numbers where the cutoff falls.

Table 1 reports on the statistics of the data generated by each year's survey. The overall school systems statistics are given in the last column but only include data from 2021 and 2022 [11] as the 2023 data was not available as of the writing of this article. As noted above, the demographics of survey participants is skewed towards higher-income students and students who do not have disabilities, although the representation of these two categories improved from year to year. Approximately 436 high schools participate in the match (the exact number is not clear as some schools are often added or removed at the last minute; the count was generated from the most recent (2021) High School Directory data [24]. Some schools have several programs to which students can apply, the survey results differentiate among these programs [22]. The survey data contains information from about a quarter of the schools.[4].

An interesting data point from Table 1 is the percentage of unmatched students in the dataset. In 2021 and 2022, 7% of students citywide were unmatched. The rate of unmatched survey participants for 2022 is close to the actual citywide rate, but the rate of unmatched survey students for 2021 is much higher. This is likely due to a few reasons:

(1) In 2021, families had to FOIL their lottery number to gain access to it. Those who did go through the process were likely to be the ones who were upset at their outcomes and were motivated to understand what happened. In later years, all applicants could see their lottery numbers in the applicant portal.
(2) Survey respondents from 2021 were typically from higher-income districts where the rate of unmatched students was higher than the city average (18% in District 2, 12% in District 3, 9% in District 15 [11]).[5]

The unmatched rate for the 2023 survey is much lower. Unfortunately, the actual unmatched rate is not available as of the writing of this article (May 2023). It is unclear whether the citywide unmatched rate declined, or if survey participants, a self-selected group, most of whom had access to past survey results and were able to tailor their choice list based on their lottery number and past cutoff information, modified their behavior in a way that minimized their odds of being unmatched. More analysis on possible changes in behavior due to having access to lottery numbers is needed (see Section 6).

4.4 Crowdsourced Data Analysis

Full results and analysis were shared online [22], and made available to NYC families to plan their application. As the details of the odds of admissions at specific schools are not directly relevant to the reader, I will only highlight the main lessons learned from the analysis in this section.

- **The lottery number had an outsized impact on admissions, including admissions to "screened" schools that considered essays and auditions in their admission process.** Several schools used both assessments and grades. Students were ranked on a composite score based on their school-specific assessments (typically essays), and grades. Many students with bad lottery numbers saw these schools as their best hope to get a desirable match, as they could compensate for bad luck in the lottery draw with a strong essay. The results show that essays were graded coarsely, or not at all, and the lottery number was used to break ties, being a major factor in admission for these schools. For instance, to gain admission to *Beacon High School*, one of the most popular high schools, students needed a perfect 400 assessment score AND a lottery number that started with '6' or lower (a 400 score with a '6f' lottery number was not given an offer). Another school, *Manhattan / Hunter Science High School* assigned the maximum essay score of 400 to *all* students who submitted the essay, resulting in students needing a good lottery number (starting with '3' or lower) to gain admission.
- **The process results in a large amount of wasted time and resources, both on the students' side and the schools' side.** The above results highlight how wasteful in terms of time and resources the process is. Because of the coarseness

[4]It is possible that the proportion of schools to which the survey participants applied is higher, however, we only count schools for which they were considered, i.e., either the school to which they matched, or the schools that they ranked higher than their match

[5]The exact geographic distribution of survey participants is not available as the geographic data was an optional question of the survey.

of the grading, students with no chance of gaining admission to either school because of their bad lottery numbers spent time writing essays and used precious choices on their list for the schools. Meanwhile, teachers spent time grading these essays (at least at *Beacon High School*), despite the grades not having any impact of these students' admission chances.

- **Students are applying to schools for which they are no seats for their priority group.** Consider the example of the *New Explorations into Science, Technology and Math High School (NEST+m)* school, a popular school with thousands of applicants. NEST+m is a K-12 school with priority given to continuing students and a 66% DIA set-aside for low-income students. Historically 53% of offers are given to continuing students. However, the demographic data of continuing students shows few students qualifying for the low-income set-aside. Once continuing students, in the top priority group, are given offers, the only remaining seats left are reserved for students who qualify for the set-aside. Non-continuing students who do not qualify for the set-aside have little to no chance of gaining admission to the school. The survey results confirm that fewer than 6 such students received an offer.[6] However, students and families are not made aware of this, so hundreds of ineligible students applied, wrote essays, and teachers spent hours grading these essays despite the students having little or no chance of receiving an offer. Had this information been available, students might have decided not to include the school as one of their 12 choices, and not to spend the time writing a series of school-specific essays for such low odds.

As discussed in Section 2.2, a large number of students ended up not getting matched to any school on their list, especially in one Manhattan district. This was in large part due to pandemic-and diversity-driven changes in admission policy, which greatly reduced the odds of admissions to a number of schools for these students compared to previous years. (Previously, Manhattan students had geographic priority in some selected high-schools, and several high-schools selected students with high GPAs and test scores.) The DOE did not communicate well the impact of the decrease in odds to families.

4.5 Waitlists: a second bite at the apple

In 2020, the DOE added waitlists to the school match following errors in the match the previous year [6, 33]. Students are automatically added to the waitlist of schools that are higher on their list than the schools to which they matched.

The process that had been communicated to families during DOE Zoom admission events, and noted on the DOE website stated that:

> *"For waitlists, each applicant receives a new random number for each waitlist they are on."*

It would be unfair for students with "bad" lottery numbers to be penalized twice. The fairest way to handle the waitlists would be to reverse lottery numbers, but redrawing them is a reasonably fair alternative [13].

[6]This was later confirmed by NYC DOE data.

Students are given the opportunity to add themselves to waitlists of schools they did not list in the main round, but they will have lower priority than students who had listed the schools originally. This unfortunately creates new strategic incentives: a student with a bad main round lottery number may be lucky in the waitlist round and receive a good lottery number at a high-demand school that they wish to attend, but to have a realistic chance of being admitted from the waitlist, they have to list the school in the main round (otherwise their waitlist position will be after all students who have listed the school). Optimizing for waitlists is risky as it requires allocating one (or more) of the 12 choices in the main round for a potential shot at a good waitlist outcome. This will mechanically increase the chances of being unmatched for students who do not have a guaranteed placement. It also requires families to apply complex game theory reasoning to their school application decisions, and evaluate the tradeoffs between the risk of being unmatched and the opportunity for an additional chance in the lottery for their most preferred school(s). Finally, there is very little information as to the odds of admissions from waitlists; many schools seldom admit any students from their waitlists, and others made extensive use of them. However, there is no transparency on the number of students who were admitted from waitlists.

Despite the published policy of redrawing lottery numbers, in July 2022 the DOE confirmed what many parents have suspected based on comparing their waitlist numbers: *the lottery numbers had not been redrawn for a subset of schools.* The reasons for this change in policy were not given. It is possible it was due to some implementation oversight. This change in published policy has two major drawbacks.

- **Unfairness.** Students with bad original random numbers received consistently bad waitlist numbers, students with good original numbers were at the top of many waitlists. This led to unfair situations where students with good placements in the main round were given multiple waitlist offers to popular schools, while students with no match in the main round had no hope to get an offer. This was an additional blow for students with bad application lottery numbers who had hoped to have better luck in the waitlist process. It also meant many families had wasted some of their choices on false hope.
- **Negative effect on diversity initiatives.** Schools do not keep separate waitlists for students eligible for the low-income set-asides. The mechanics of the set-aside means that the admission lottery cutoffs for DIA-eligible students are higher than those of non-eligible students. Because the lottery numbers are not redrawn for the waitlists, there is now a perverse effect where students at the top of the list are all non-eligible. This means that schools that admit students from their waitlists are diluting the effect of their diversity initiatives and the percentage of low-income students.

5 LESSONS LEARNED AND IMPACTS

This work, a collaboration between elected parent representatives, parent advocacy groups, and researchers has pushed the DOE to provide more, and clearer, information to families for the 2023 admission season:

- Lottery numbers are now available for all applicants on the application portal. Unfortunately, little to no information is provided as to how to interpret the 32-character random number.
- The historical number of applicants to a school has been changed to show only those who did not match higher on their list. This provides a better intuition of the popularity of a school and the odds of receiving an offer.
- Some clarification on the number of seats, and the number of offers, for each priority group is now provided in the school directory.
- Students on school waitlists can now see how many waitlist offers were made the previous year.
- Unfortunately, historical cutoffs are still not provided by the DOE.

Despite these clarifications, the implementation of the lottery lacks oversight. In 2022, some parents were surprised to encounter a "glitch" in the application portal: if they canceled and re-started their application they would be given a new lottery number [23], an obvious loophole that would have allowed gaming of the system by re-drawing bad lottery numbers. The DOE claimed that *"there is no glitch in [their] system"*, but said it will be fixed [23].

6 RELATED WORK

Hakimov and Kübler [18] survey the literature on matching markets experiments under various settings; they particularly focus on how preferences are affected by participants' understanding of the algorithms and of the information available to them such as the preferences of other market participants, or the size of the market.

In one such experiment, Bó and Hakimov [5] studied the impact of providing cutoff information to students in an iterative matching market where students apply to one school in each round. Their results show that knowledge of cutoffs leads to a significant increase in truthful behavior. While these results were observed in a laboratory experimental setting using an iterative algorithm and not the deferred acceptance algorithm used in the NYC match, it would be interesting to test the impact of knowledge of the historical cutoff information produced by the survey on truthful behavior and whether this could lead to a decrease in unmatched students.

Other works have considered the impact of the available information in settings where schools rank students using their scores on a standardized exam, which happens in several countries such as China [21, 25]. In particular, [25] shows that in the absence of prior knowledge of the exam score, choice behavior exacerbates existing bias in students' confidence, rewarding overconfidence and punishing underconfidence in testing abilities as students will select schools based on their expected performance. In our setting, lottery numbers act similarly to test scores, providing a full ranking of students. An interesting hypothesis is whether knowing their lottery numbers would change the behavior of under-privileged students by giving them knowledge of the possibility of being matched to a popular school, and combat the negative self-selection that has been observed in students from low socio-economic backgrounds [8].

The work presented in this paper has helped shed light on the NYC school matching process for families but a lot is still needed: while the study allowed us to clarify how the random numbers were used, and push for more transparency from the DOE, the data gathered via crowdsourcing is very incomplete and of limited use to many families. In addition, the most vulnerable families, many of whom are non-native English speaking, are unlikely to have the resources to find online posts detailing the process. The information is therefore likely to disproportionally benefit the most privileged unless the DOE makes a significant effort in communication and outreach. Lack of information is one of the barriers to participation in school choice systems [27]. Partial or confusing information can also lead to suboptimal choice behavior and worse outcomes [17]. Promising work has shown that the quality of matches for underserved populations can be improved by providing better information to students and their families so that their choice lists better reflect their preferences [3, 9].

7 CONCLUSIONS

In this paper, I reported on an effort to add transparency in the NYC School Matching process by leveraging the information held by families, who are stakeholders in the system, through crowdsourcing. This work shed light on some of the mechanics of the implementation of the match, specifically highlighting how randomly assigned numbers are used in the process, identified some inefficiencies and inequities in the implementation, and pushed the NYC DOE to be more accountable for its processes.

This work shows that stakeholders are not powerless in the face of automated decision processes and impersonal bureaucracies, and can use their collective influence to hold public agencies accountable for their algorithmic decisions and processes. There are needed improvements in how public agencies communicate their processes and are held accountable for their decisions. Information should not only be available to those who have the time and knowledge to navigate lengthy FOIL or FOIA requests. Transparency and communication would go a great length to improve the trust of stakeholders who have no choice but to interact with automated decision systems.

ACKNOWLEDGMENTS

This work was partially supported by NSF Grant SES-2218975.

REFERENCES

[1] Atila Abdulkadiroğlu, Parag A Pathak, and Alvin E Roth. 2005. The New York City high school match. *American Economic Review* 95, 2 (2005), 364–367.
[2] Atila Abdulkadiroğlu, Parag A Pathak, and Alvin E Roth. 2009. Strategy-proofness versus efficiency in matching with indifferences: Redesigning the NYC high school match. *American Economic Review* 99, 5 (2009), 1954–78.
[3] Kehinde F Ajayi, Willa H Friedman, and Adrienne M Lucas. 2017. The importance of information targeting for school choice. *American Economic Review* 107, 5 (2017), 638–43.
[4] Itai Ashlagi and Afshin Nikzad. 2020. What matters in school choice tie-breaking? How competition guides design. *Journal of Economic Theory* 190 (2020), 105120.
[5] Inácio Bó and Rustamdjan Hakimov. 2020. Iterative versus standard deferred acceptance: Experimental evidence. *The Economic Journal* 130, 626 (2020), 356–392.
[6] Chalkbeat. 2019. *Lab Middle School experiences admissions snafu, weeks after high school problems.* https://www.chalkbeat.org/posts/ny/2019/04/25/lab-middle-school-experiences-admissions-snafu-weeks-after-high-school-problems/
[7] Chalkbeat. 2020. *NYC announces sweeping changes to middle, high school application process.* https://ny.chalkbeat.org/2020/12/18/22188384/changes-nyc-school-application-process
[8] Li Chen and Juan Sebastián Pereyra. 2019. Self-selection in school choice. *Games and Economic Behavior* 117 (2019), 59–81.

[9] Sarah Cohodes, Sean Corcoran, Jennifer Jennings, and Carolyn Sattin-Bajaj. 2017. NYC High School Admissions Study. http://www.nychighschooladmissionstudy.com.

[10] Sarah Cohodes, Sean Corcoran, Jennifer Jennings, and Carolyn Sattin-Bajaj. 2022. *When Do Informational Interventions Work? Experimental Evidence from New York City High School Choice.* Working Paper 29690. National Bureau of Economic Research. https://doi.org/10.3386/w29690

[11] NYC Department of Education. 2023. *Infohub - Admissions and Enrollment.* https://infohub.nyced.org/reports/admissions-and-enrollment

[12] Virginia Dignum. 2018. Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology* 20, 1 (01 Mar 2018), 1–3. https://doi.org/10.1007/s10676-018-9450-z

[13] Itai Feigenbaum, Yash Kanoria, Irene Lo, and Jay Sethuraman. 2020. Dynamic matching in school choice: Efficient seat reassignment after late cancellations. *Management Science* 66, 11 (2020), 5341–5361.

[14] NYC ADS Task Force. 2019. New York City Automated Decision System Task Force Report. (November 2019). https://www1.nyc.gov/assets/adstaskforce/downloads/pdf/ADS-Report-11192019.pdf.

[15] Abraham Gale and Amélie Marian. 2020. Explaining monotonic ranking functions. *Proceedings of the VLDB Endowment* 14, 4 (2020), 640–652.

[16] David Gale and Lloyd S Shapley. 1962. College admissions and the stability of marriage. *The American Mathematical Monthly* 69, 1 (1962), 9–15.

[17] Yannai A Gonczarowski, Lior Kovalio, Noam Nisan, and Assaf Romm. 2019. Matching for the Israeli" Mechinot" Gap-Year Programs: Handling Rich Diversity Requirements. *arXiv preprint arXiv:1905.00364* (2019).

[18] Rustamdjan Hakimov and Dorothea Kübler. 2021. Experiments on cCentralized School Choice and College Admissions: a Survey. *Experimental Economics* 24 (2021), 434–488.

[19] C Hemphill, N Mader, M Quiroz, and L Zingmond. 2019. Screened schools: How to broaden access and diversity.

[20] AI Now Institute. 2019. Confronting Black Boxes: A Shadow Report of the New York City Automated Decision System Task Force. (December 4 2019). https://ainowinstitute.org/ads-shadowreport-2019.html.

[21] Jaimie W Lien, Jie Zheng, and Xiaohan Zhong. 2016. Preference submission timing in school choice matching: testing fairness and efficiency in the laboratory. *Experimental Economics* 19 (2016), 116–150.

[22] Amélie Marian. 2021. *Decoding the NYC School Admission Lottery Numbers.* https://medium.com/p/bae7148e337d

[23] New York Post. 2022. *Parents uncover major glitch in NYC school lottery system.* https://nypost.com/2022/04/16/parents-uncover-major-glitch-in-nyc-school-lottery-system/

[24] NYC OpenData. 2023. *NYC OpenData - Education.* https://data.cityofnewyork.us/browse?category=Education

[25] Siqi Pan. 2017. The instability of matching with overconfident agents: laboratory and field investigations. *Available at SSRN 4291709* (2017).

[26] Parag A Pathak and Jay Sethuraman. 2011. Lotteries in student assignment: An equivalence result. *Theoretical Economics* 6, 1 (2011), 1–17.

[27] Samantha Robertson, Tonya Nguyen, and Niloufar Salehi. 2021. Modeling assumptions clash with the real world: Transparency, equity, and community challenges for student assignment algorithms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–14.

[28] Julia Stoyanovich, Serge Abiteboul, and Gerome Miklau. 2016. Data, responsibly: Fairness, neutrality and transparency in data analysis. In *International Conference on Extending Database Technology.*

[29] A. Stuart and K. Ord. 1991. *Kendall's advanced theory of statistics* (fifth ed.). Vol. 2, Classical Inference and Relationship.

[30] Jim Torresen. 2018. A Review of Future and Ethical Perspectives of Robotics and AI. *Frontiers in Robotics and AI* 4 (2018), 75.

[31] Amos Tversky and Daniel Kahneman. 1971. Belief in the law of small numbers. *Psychological bulletin* 76, 2 (1971), 105.

[32] Suresh Venkatasubramanian, Nadya Bliss, Helen Nissenbaum, and Melanie Moses. 2020. Interdisciplinary Approaches to Understanding Artificial Intelligence's Impact on Society. *CoRR* abs/2012.06057 (2020). arXiv:2012.06057 https://arxiv.org/abs/2012.06057

[33] Wall Street Journal. 2019. *Top New York City School Makes Big Admissions Error.* https://www.wsj.com/articles/top-new-york-city-school-makes-big-admissions-error-11554847303