# EDLD 651 Final Project Draft

Anwesha Guha[1], Heidi Iwashita[1], Christopher Loan[1], Adam Nielsen[1], & Aaron Rothbart[1]

[1] University of Oregon

## Author Note

All work done herein represents contributions from all authors equally. Author order is alphabetical.

Abstract

FILL IN ABSTRACT IF WANTED

*Keywords:* keywords

Word count: X

EDLD 651 Final Project Draft

## Introduction

We explore proportion of graduation (outcome), across several categorical variables. In particular, we plan to focus on comparisons of two groups who have historically had unequal access to resources: English language learners (ELL) vs. English proficient (EP) students & Special Education (SPED) status vs. non-SPED status.

Not only will we report these outcomes across different groups, we will also explore these across boroughs, too, to see if these groups are succeeding equally across boroughs—-as measured by graduation outcomes—-compared to the English proficient students in their boroughs.

## Methods

We retrieved the data collected by the Department of Education from

Information about variables, how they were measured here

Information about regents examinations here

### Participants

Explain participants' from what we have in data.

First, we import and clean our data:

```
raw_grad <- import(here("data", "2005-2010__Graduation_Outcomes_-__By_Borough.csv"))
grad <- raw_grad %>%
  clean_names() %>%
  as_tibble()


summary(grad$cohort) # we see here that 'Aug 2006' needs to be changed to '2006' for c
```

```
##    Length     Class      Mode
##       385 character character
```

```
grad$cohort <-  as.numeric(sub("Aug 2006", "2006", grad$cohort))


head(grad)#need to change var names to make legible, perhaps subset data to only inclu
```

```
## # A tibble: 6 x 22
##   demographic borough cohort total_cohort total_grads_n total_grads_per~
##   <chr>       <chr>    <dbl>        <int>         <int>            <dbl>
## 1 Borough To~ Bronx     2001        11453          4913             42.9
## 2 Borough To~ Bronx     2002        12032          5328             44.3
## 3 Borough To~ Bronx     2003        13632          6389             46.9
## 4 Borough To~ Bronx     2004        14364          7448             51.9
## 5 Borough To~ Bronx     2005        15175          8229             54.2
## 6 Borough To~ Bronx     2006        15579          8524             54.7
## # ... with 16 more variables: total_regents_n <int>,
## #   total_regents_percent_of_cohort <dbl>,
## #   total_regents_percent_of_grads <dbl>, advanced_regents_n <int>,
## #   advanced_regents_percent_of_cohort <dbl>,
## #   advanced_regents_percent_of_grads <dbl>, regents_w_o_advanced_n <int>,
## #   regents_w_o_advanced_percent_of_cohort <dbl>,
## #   regents_w_o_advanced_percent_of_grads <dbl>, local_n <int>,
## #   local_percent_of_cohort <dbl>, local_percent_of_grads <dbl>,
## #   still_enrolled_n <int>, still_enrolled_percent_of_cohort <dbl>,
## #   dropped_out_n <int>, dropped_out_percent_of_cohort <dbl>
```

```
# Do we want to use recode() or rename()? Also, does it make more sense to leave all o
```

## PIVOTS

The data we are starting with are already tidy, but for the purposes of demonstrating our rather acute proficiency in our *ability* to tidy data, in this segment will make the data

untidy and then tidy it once more.

```
messy_grad <- grad %>%
  pivot_wider(names_from = borough,
              values_from = total_cohort)
head(messy_grad)
```

```
## # A tibble: 6 x 25
##   demographic cohort total_grads_n total_grads_per~ total_regents_n
##   <chr>        <dbl>         <int>            <dbl>           <int>
## 1 Borough To~   2001          4913             42.9            2644
## 2 Borough To~   2002          5328             44.3            3118
## 3 Borough To~   2003          6389             46.9            3861
## 4 Borough To~   2004          7448             51.9            4625
## 5 Borough To~   2005          8229             54.2            5618
## 6 Borough To~   2006          8524             54.7            6312
## # ... with 20 more variables: total_regents_percent_of_cohort <dbl>,
## #   total_regents_percent_of_grads <dbl>, advanced_regents_n <int>,
## #   advanced_regents_percent_of_cohort <dbl>,
## #   advanced_regents_percent_of_grads <dbl>, regents_w_o_advanced_n <int>,
## #   regents_w_o_advanced_percent_of_cohort <dbl>,
## #   regents_w_o_advanced_percent_of_grads <dbl>, local_n <int>,
## #   local_percent_of_cohort <dbl>, local_percent_of_grads <dbl>,
## #   still_enrolled_n <int>, still_enrolled_percent_of_cohort <dbl>,
## #   dropped_out_n <int>, dropped_out_percent_of_cohort <dbl>, Bronx <int>,
## #   Brooklyn <int>, Manhattan <int>, Queens <int>, `Staten Island` <int>
```

```
clean_grad <- messy_grad %>%
  pivot_longer(cols = c("Bronx":"Staten Island"),
```

```
                names_to = "borough",
                values_to = "total_cohort",
                values_drop_na = TRUE)


clean_grad <- clean_grad[, c(1,21,2,22,3:20)]
head(clean_grad)
```

```
## # A tibble: 6 x 22
##   demographic borough cohort total_cohort total_grads_n total_grads_per~
##   <chr>       <chr>    <dbl>        <int>         <int>             <dbl>
## 1 Borough To~ Bronx     2001        11453          4913              42.9
## 2 Borough To~ Bronx     2002        12032          5328              44.3
## 3 Borough To~ Bronx     2003        13632          6389              46.9
## 4 Borough To~ Bronx     2004        14364          7448              51.9
## 5 Borough To~ Bronx     2005        15175          8229              54.2
## 6 Borough To~ Bronx     2006        15579          8524              54.7
## # ... with 16 more variables: total_regents_n <int>,
## #   total_regents_percent_of_cohort <dbl>,
## #   total_regents_percent_of_grads <dbl>, advanced_regents_n <int>,
## #   advanced_regents_percent_of_cohort <dbl>,
## #   advanced_regents_percent_of_grads <dbl>, regents_w_o_advanced_n <int>,
## #   regents_w_o_advanced_percent_of_cohort <dbl>,
## #   regents_w_o_advanced_percent_of_grads <dbl>, local_n <int>,
## #   local_percent_of_cohort <dbl>, local_percent_of_grads <dbl>,
## #   still_enrolled_n <int>, still_enrolled_percent_of_cohort <dbl>,
## #   dropped_out_n <int>, dropped_out_percent_of_cohort <dbl>
```

Now that we have tidied the entire dataset, we can focus on our variables of interest: enrollment and graduation for specific boroughs, cohorts and demographics.

```
filtered_grad <- clean_grad %>%
  select(c(1:6, 16:22)) %>%
  filter(demographic == "English Language Learners" |
         demographic == "English Proficient Students" |
         demographic == "Special Education" |
         demographic == "General Education") %>%
  mutate(student_characteristic =
           factor(demographic,
                  levels = c("English Language Learners",
                      "English Proficient Students",
                      "Special Education",
                      "General Education"),
                  labels = c('ELL', 'EP', 'SPED', 'Non-SPED')
                  ))


new_grad <- filtered_grad %>%
  mutate(unclassified_n = total_cohort - (total_grads_n + dropped_out_n + still_enrolled
         unclassified_percent_of_cohort = round(unclassified_n/total_cohort * 100, 1))
```

```
head(new_grad)
```

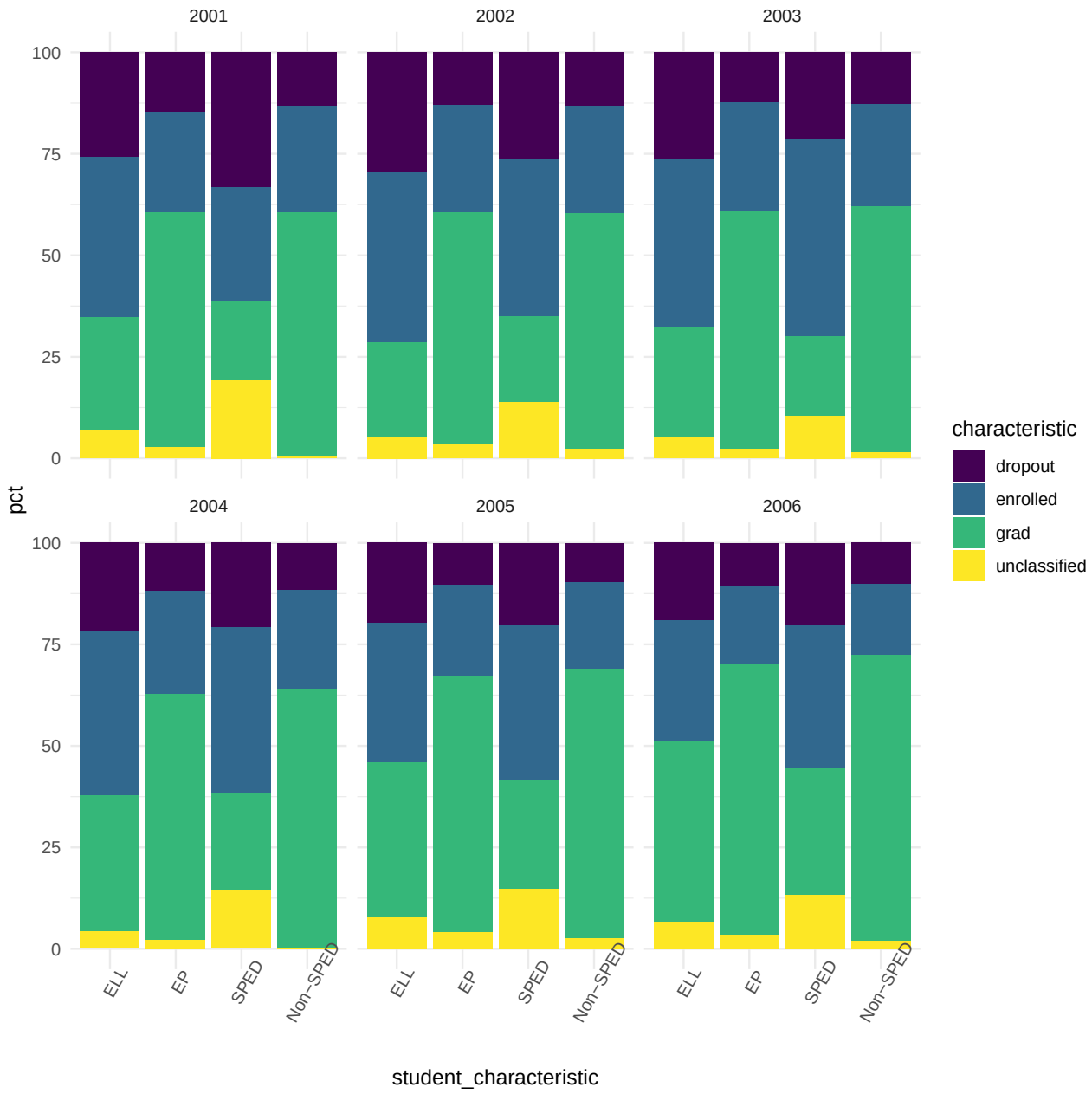
```
## # A tibble: 6 x 16
##   demographic borough cohort total_cohort total_grads_n total_grads_per~ local_n
##   <chr>       <chr>    <dbl>        <int>         <int>            <dbl>   <int>
## 1 English La~ Bronx     2001         1984           388             19.6     311
## 2 English La~ Bronx     2002         1693           333             19.7     257
## 3 English La~ Bronx     2003         1905           391             20.5     296
## 4 English La~ Bronx     2004         1894           640             33.8     426
```

```
## 5 English La~ Bronx      2005          1940            694              35.8      377
## 6 English La~ Bronx      2006          2143            791              36.9      395
## # ... with 9 more variables: local_percent_of_cohort <dbl>,
## #   local_percent_of_grads <dbl>, still_enrolled_n <int>,
## #   still_enrolled_percent_of_cohort <dbl>, dropped_out_n <int>,
## #   dropped_out_percent_of_cohort <dbl>, student_characteristic <fct>,
## #   unclassified_n <int>, unclassified_percent_of_cohort <dbl>
```

```
# group by relevant demographics (ELL & EP, GE & SPED)
demographic_data <- new_grad %>%
  group_by(student_characteristic, cohort) %>%
  summarize(mean_grad_pct = mean(total_grads_percent_of_cohort),
            mean_dropout_pct = mean(dropped_out_percent_of_cohort),
            mean_enrolled_pct = mean(still_enrolled_percent_of_cohort),
            mean_unclassified_pct = mean(unclassified_percent_of_cohort))


# group by borough, look at % of local students
borough_data <- new_grad %>%
  group_by(borough, cohort) %>%
  summarize(mean_local = mean(local_percent_of_cohort),
            mean_grad_pct = mean(total_grads_percent_of_cohort),
            mean_dropout_pct = mean(dropped_out_percent_of_cohort),
            mean_enrolled_pct = mean(still_enrolled_percent_of_cohort),
            mean_unclassified_pct = mean(unclassified_percent_of_cohort))
```

```
demographic_bar <- demographic_data %>%
  pivot_longer(cols = contains("mean"),
               names_to = c("characteristic", ".value"),
```

```
                names_prefix = "mean_",
                names_sep = "_")


# which makes more sense?
# Option 1 - cohort as factor(), faceted by characteristic
# demographic_bar %>%
#   ggplot(aes(fill = factor(cohort), x = student_characteristic, y = pct)) +
#   geom_bar(position = "stack", stat = "identity") +
#   theme(axis.text.x = element_text(angle = 60)) +
#   facet_grid(~characteristic + cohort) +
#   scale_fill_viridis_d()


# Option 2 - cohort and characteristic switched
demographic_bar %>%
  ggplot(aes(fill = characteristic, x = student_characteristic, y = pct)) +
  geom_bar(position = "stack", stat = "identity") + # do we want stack or dodge for pos
  theme(axis.text.x = element_text(angle = 60)) +
  facet_wrap(~cohort) +
  scale_fill_viridis_d()
```




107

```
# We can also look at the following to get a general sense of the data:
# - total cohorts/grads, facet_wrap by borough
# - grad percentage by student_characteristic, then can do a deeper dive by borough
# - the above two repeated with dropout rate
```

## Data analysis

All analysis were conducted in R, with heavy reliance upon the `{tidyverse}` packages to manipulate and visualize the data.

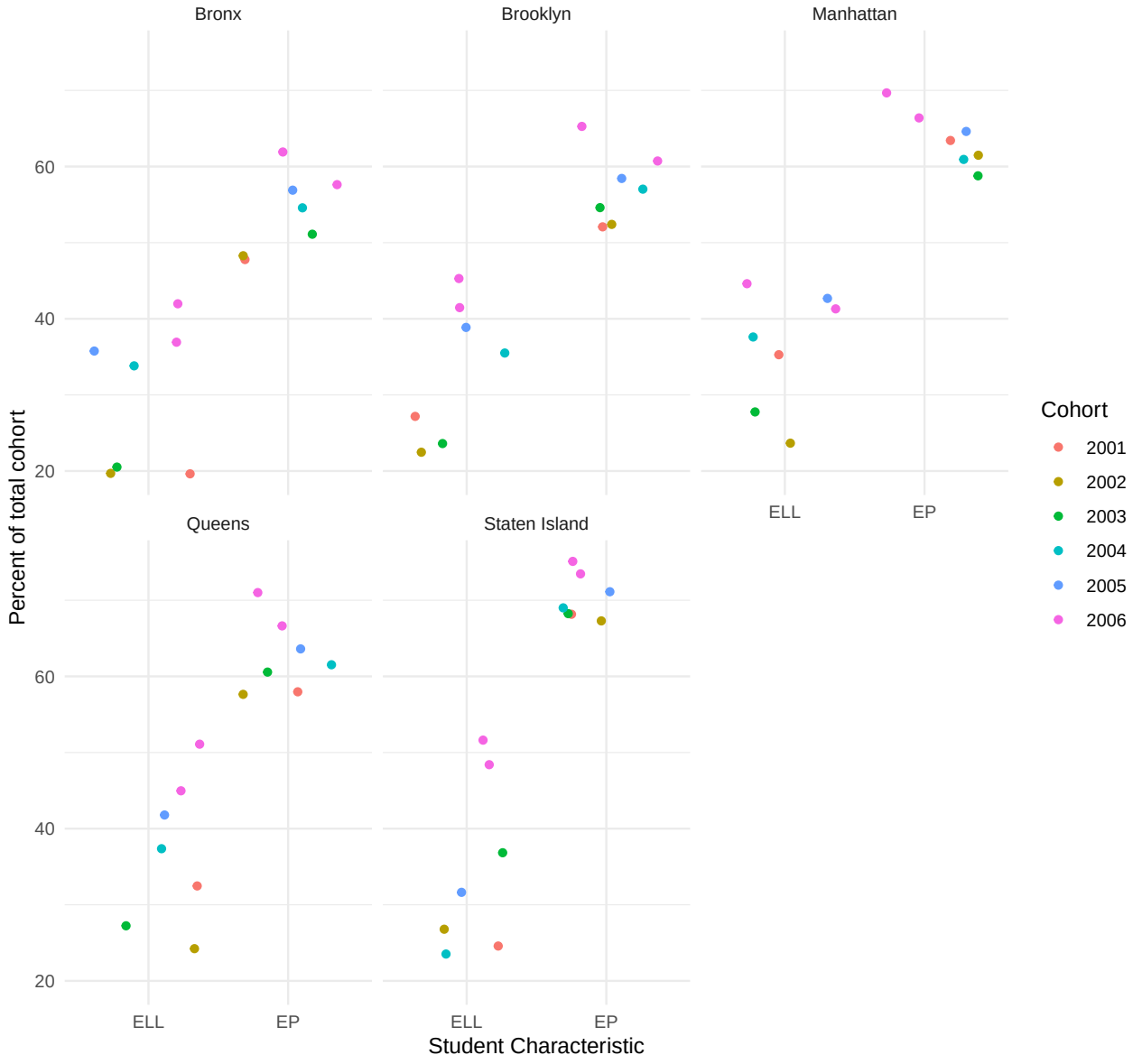
## Results

```
#report graduation by borough

#report graduation by english language status

#report graduation by SPED status

#report graduation by borough & SPED status

#report graduation by borough & english learner status
```

```
new_grad %>%
  filter(student_characteristic == "ELL" |
           student_characteristic == "EP") %>%
  mutate(Cohort = factor(cohort)) %>%
  group_by(student_characteristic, borough) %>%
  ggplot(aes(x = student_characteristic,
             y = total_grads_percent_of_cohort)) +
  geom_jitter(aes(color = Cohort)) + facet_wrap(~borough) +
  labs(title = 'Figure 1. Graduation Rates in NYC by English Learner Status',
       subtitle = 'Boroughs are reported separetely with lighter dots indicating more re
       y = 'Percent of total cohort',
       x = 'Student Characteristic')
```

Figure 1. Graduation Rates in NYC by English Learner Status

Boroughs are reported separetely with lighter dots indicating more recent years



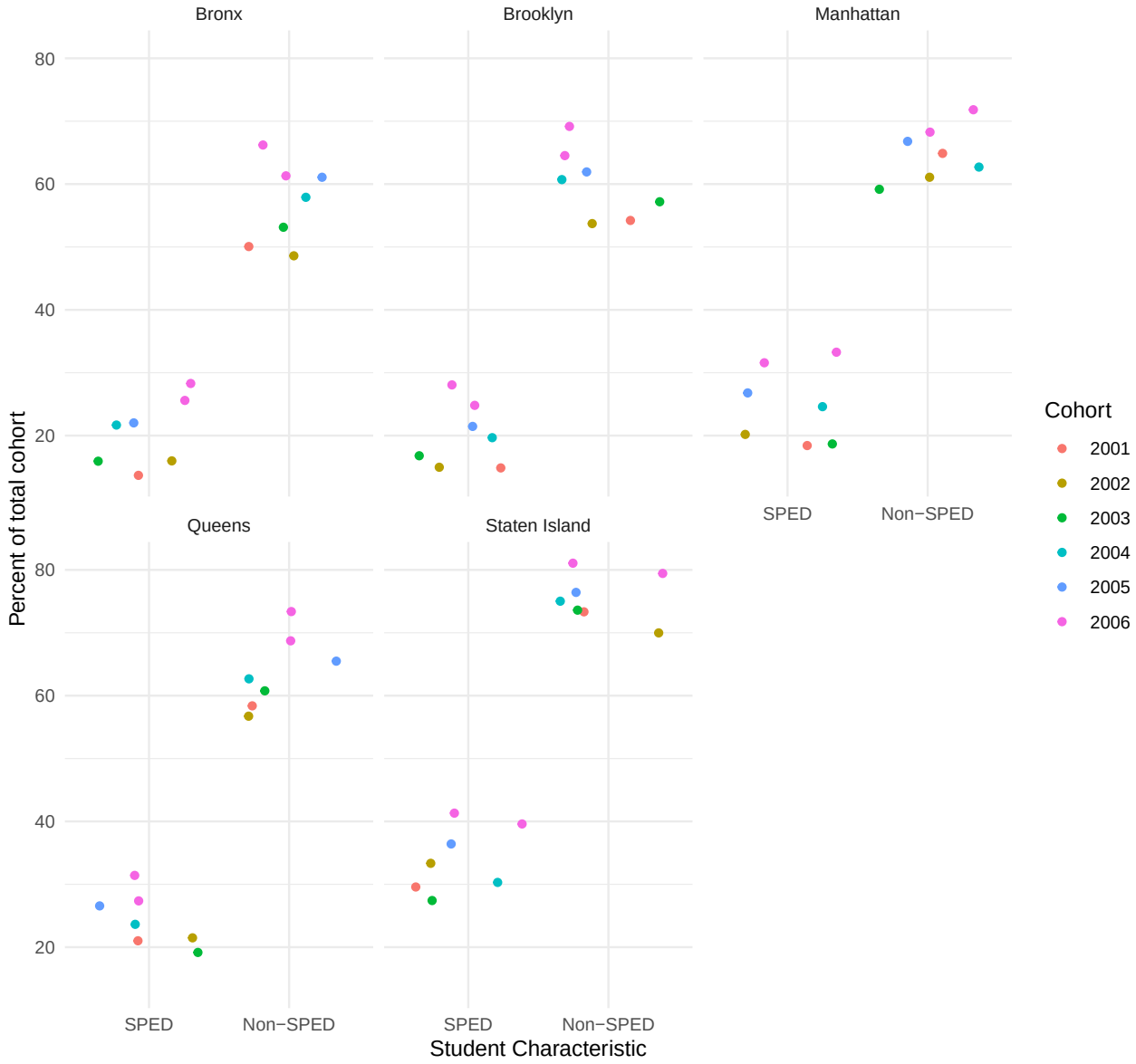

112

```
new_grad %>%
  filter(student_characteristic == "SPED" |
           student_characteristic == "Non-SPED") %>%
  mutate(Cohort = factor(cohort)) %>%
  group_by(student_characteristic, borough) %>%
  ggplot(aes(x = student_characteristic,
             y = total_grads_percent_of_cohort)) +
```

```
  geom_jitter(aes(color = Cohort)) +
  facet_wrap(~borough) +
  labs(title = 'Figure 1. Graduation Rates in NYC by English Learner Status',
       subtitle = 'Boroughs are reported separetely with lighter dots indicating more re
       y = 'Percent of total cohort',
       x = 'Student Characteristic')
```

Figure 1. Graduation Rates in NYC by English Learner Status

Boroughs are reported separetely with lighter dots indicating more recent years



## Discussion

Differences appear to be blah by blah for blah. XYZ boroughs should consider blah blah blah, based on the results. Inferential tests are recommended for next directions.

## References