

Key Demographics of High School Students that passes the Mandatory Ontario Secondary School Literacy Test (OSSLT) on First Attempt*

An Analysis Based on the Ontario School and Student Demographic Data published by the
ministry of Education

Charles Lu

29 April 2022

Abstract

Using data extracted from school and student demographic report by Education, this analysis aims to study Ontario high school's student passing rate on OSSLT test by analysing the profile and demographics of each school specifically in school type and ratio statistics like ESL students, special education students, gifted students, low-income household, and parents with no education degree. The report concludes that, with deep understanding of the context and proper interpretations, all demographics shows an significant amount of effect on students ability to pass the OSSLT. In the future, this report can give educators a better understanding of causes and reasons of the OSSLT results, reseach further into the correlation between the test results and student demographics, eventually imrove on the current education system and hopefully raises overall literacy skills of all students.

Contents

| | | |
|----------|--------------------------------|----------|
| 1 | Introduction | 2 |
| 2 | Data | 3 |
| 2.1 | Data overview | 3 |
| 2.2 | The cleaning process | 3 |
| 2.3 | Visualization | 4 |
| 3 | Model | 5 |
| 3.1 | Model construction | 5 |
| 3.2 | Model Validation | 6 |
| 4 | Results | 6 |

*https://github.com/R300G/Ontario_School_Data

| | |
|--|-----------|
| 5 Discussion | 8 |
| 5.1 Interpretation of results and findings | 8 |
| 5.2 Weaknesses | 9 |
| 5.3 Future directions | 10 |
| Appendix | 11 |
| .1 Additional figures | 11 |
| .2 Datasheet | 15 |
| References | 20 |

1 Introduction

Canada is a multicultural country, Canadians come from a vast range of nations, races, religions, and heritage and together created an open and diversified community. This multicultural diversity comes from centuries of immigration. To work together as a whole community with many cultural differences, possessing the ability to communicate using Canada’s first official language English is crucial. As the most populated province in Canada, the Ontario government have taken this very seriously, and they tackles the communication issue by focusing on improving the literacy skills of all school-age children, since students are the future of this society. Launched in 1999, the OSSLT (Ontario Secondary Literacy Test) is a standardized test that measure whether grade 10 Ontario high school students are meeting the minimum standard for literacy across all subjects up to the end of grade 9. Ontario ministry of education have included the successful completion of this literacy test as one of the requirements to earn an Ontario Secondary School Diploma. (“Ontario Secondary School Literacy Test (OSSLT)” 2022) Exploring school and student demographics that can potentially influences the results of OSSLT test is crucial as the findings can guide educators and school officials to revise and improve Ontario’s education program and help all students to achieve provincial academic standards on literacy skills.

This paper aimed to determine the effect of various school and students demographics on the school’s students ratio that passed OSSLT on their first attempt. The demographics that the paper will be examined are school type (public or catholic), English as second language students ratio, special education students ratio, gifted students ratio, percentage of students living in low-income household, and percentage of students whose parent’s have no education degree. A linear mixed effect regression model was constructed to fit the school demographics data and tasked to estimate the percentage of students that passed OSSLT on first try. The results provide valuable insights. The models indicates that students living in low-income household and/or have parents with no education are less likely to pass OSSLT. On the other hand, students in catholic school and/or identified as gifted are more likely to pass OSSLT. Although ESL and special education also showed significant amount of influences on the chances of students succeeding in OSSLT, it is important to note that ESL and special education students receives special accommodations that minimizes the difficulty of the test and greatly increases the chances of them passing OSSLT. Thus, results for ESL and special education students should not be easily interpreted without context. With these results in mind, the paper suggest educators and school officials to study and learn from catholic and gifted education programs where students are more likely to achieve provincial literacy standards and focuses its attention and resources on special education students, and students coming from a family with low-income and no education.

This paper utilizes analytic programming language R (R Core Team 2020). R packages tidyverse (Wickham et al. 2019), janitor (Firke 2021) and dplyr (Wickham et al. 2021) are used to clean, organize and manipulate the data. Furthermore, R packages ggplot2 (Wickham 2016), kableExtra (Zhu 2021) and stargazer (Hlavac 2022) are used to create figures and tables. R package lme4 (Bates et al. 2015) are used to create linear mixed effect model. Lastly, R packages bookdown (Xie 2021a) and tinytex (Xie 2021b) are used to generate the R markdown Report.

This remaining paper will be separated into different sections. Firstly, an overview of the original data, its sources and collection methodology will be introduced. The paper will later on discuss the detailed process of cleaning and tidying the data, and visualizations of the data will also be included to better comprehend the structure and distribution of the data. Secondly, the model section will thoroughly discuss the construction of the model used in this paper, model validity and assumptions will also be provided and explained. Then, the results section discuss the interpretation of the significance and estimated coefficients from the model and its effect on the OSSLT pass ratio. Lastly, interpretations, key findings, future suggestions of the research results under the given context will be discussed, limitations of this paper and future suggestions on future research will also be included in the end.

2 Data

2.1 Data overview

This report utilizes data regarding school information and student demographic, published by the Ontario Ministry of Education. (“School Information and Student Demographics - Ontario Data Catalogue” 2022) The dataset is publicly available through the Ontario Data Catalogue Portal. The data collection provides board information, school information, students EQAO and OSSLT result, as well as student demographic percentages on student parents, special education, first language spoken and new students to Canada. This report consolidates information from various databases, including Board School Identification Database (BSID), Ontario School Information System (OnSIS), OnSIS Preliminary 2019-2020, Education Quality and Accountability Office Database, and 2016 Census. This information is also available on the Ministry of Education’s School Information Finder website by individual schools. The information in the School Information Finder is the most up-to-date that the Ministry of Education has at this time, based on reports from schools, school boards, EQAO, and Statistics Canada. The information is updated on a monthly basis, and is last refreshed on April 12, 2022. The population of this data set is all public, Roman Catholic, and provincial schools in Ontario. The data set does not include private schools, Education and Community Partnership Programs (ECPP), summer and night schools.

2.2 The cleaning process

Variables from the school information and students demographic data were cleaned and filtered in 01-data_cleaning.R under the scripts folder. Firstly, the paper aims to research the effect of school and students demographics on the Ontario Secondary Literacy Test pass rate. Thus, I filtered all the Ontario school observations that were not a high school, and also removed all of the schools with any NA (not available), SP (suppressed), N/D (no data), or N/R (not reported) in their recorded information. I then remove all the variables that are irrelevant to high school, such as grade range, school level, percentage of grade 3 students achieving provincial writing standards and much more. In order to continue create and maintain a tidy data set, I made following choices to select variables that provide similar information with other variables.

- school_number, school_name was removed, as they both serve similar purpose to the index number of each observation.
- board_type was removed and kept school type, as they both shows whether the school is catholic or public.
- Only board_name was kept for observation’s region information as I think this is the best way to group schools by region because schools under the same school board shares similar education curricular and programs.

The variables this paper have selected to be included in the data set are school type, board name, percentage of students whose first language is not English, percentage of students receiving special education services,

Table 1: The first 10 Rows of cleaned Dataset

| school_type | board_name | osslt_passes | esl_students | se_students | gifted_students | low_income_household | no_degree_parents |
|-------------|------------------------------|--------------|--------------|-------------|-----------------|----------------------|-------------------|
| public | algoma dsb | 71 | 0 | 25 | 0 | 15 | 10 |
| public | algoma dsb | 53 | 0 | 20 | 0 | 10 | 10 |
| public | algoma dsb | 61 | 0 | 30 | 0 | 15 | 10 |
| public | algoma dsb | 68 | 3 | 19 | 0 | 12 | 6 |
| public | algoma dsb | 57 | 5 | 30 | 0 | 5 | 0 |
| public | algoma dsb | 72 | 2 | 20 | 1 | 16 | 6 |
| public | algoma dsb | 57 | 0 | 40 | 0 | 20 | 10 |
| public | algoma dsb | 58 | 10 | 35 | 0 | 15 | 0 |
| catholic | algonquin and lakeshore cdsb | 77 | 5 | 25 | 0 | 8 | 1 |
| catholic | algonquin and lakeshore cdsb | 86 | 5 | 20 | 0 | 20 | 0 |

percentage of students identified as gifted, percentage of students that passed the grade 10 OSSLT on their first attempt, percentage of school-aged children who live in low-income households, and percentage of students whose parents have no degree, diploma or certificate. The variables were cleaned and renamed to school_type, board_name, esl_students (English as second language), se_students (special education), gifted_students, osslt_passes, low_income_household, no_degree_parents. To clarify, ESL students are English language learners whose first language is not English. Special education students are students who possess learning disabilities and requires tailored special education. Gifted students are identified intellectual students in the top 2% of the Ontario student population.

Lastly, cell values within the data were cleaned as some of them are inconsistent with the others. All characters were lower cased so that board_name like “TORONTO DSB” and “Toronto DSB” will be considered as the same thing. All percentage numeric values are in the same format, values like 33.33% or 0.3333 were all converted to 33.33.

A table with the first 10 rows of the filtered and cleaned data was shown in table 1.

2.3 Visualization

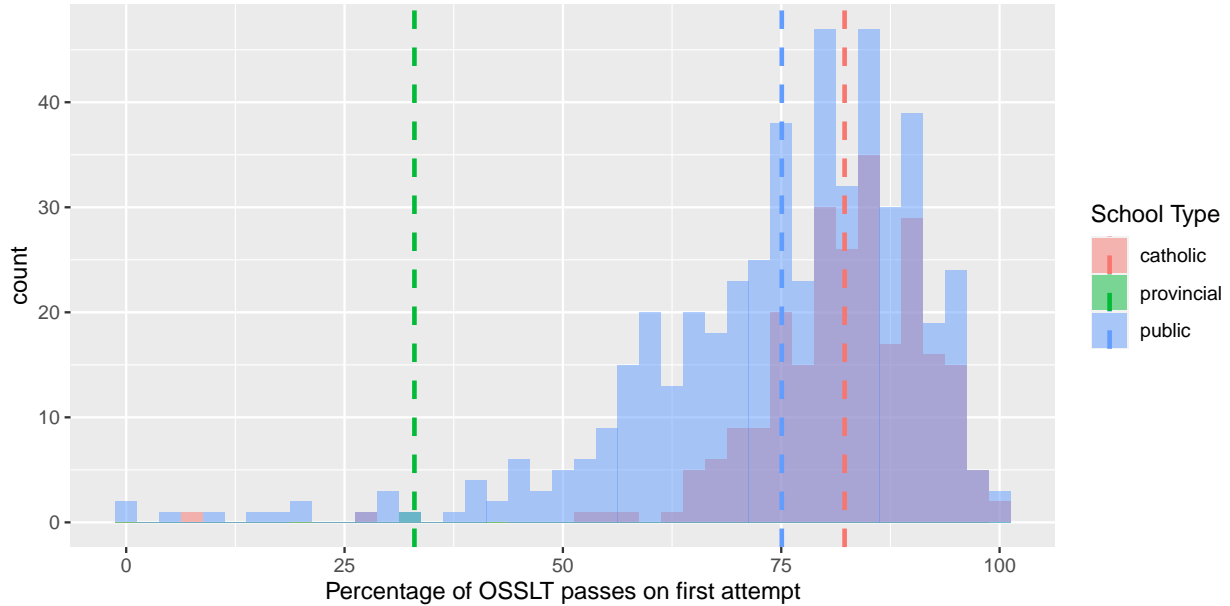


Figure 1: The distribution of OSSLT pass rate by school type

To have a deeper understanding of the data, figure 1 represent the visualizations of the distribution of OSSLT pass ratio based on the type of high school. In doing so, figure 1 showcases the relationship between OSSLT passes and school type, and how each school type influences the chances of students passing OSSLT. Firstly, the single green bin in the figure represents the OSSLT pass rate of the only one provincial and demonstrations high school in Ontario. The school provides educations for special education students who are deaf, blind, deafblind or have severe learning disabilities, and it is reasonable to see that the school had a 33% pass ratio. Moreover, between public and catholic school, there are way more public schools than catholic schools in Ontario, as the blue bin surpasses the red bin across all OSSLT pass ratio. However, catholic school's OSSLT pass rates shares similar distribution shape with public school's but its mean and overall distribution are more right to the relative position of public school's. Thus, it is reasonable to interpret that catholic school students overall are a little bit more likely to pass OSSLT than public school students.

The following figures represent the visualizations of the linear relationship between the school's percentage of OSSLT passes on first attempt to multiple relevant school demographics such as ESL students ratio, special education students ratio, gifted students ratio, proportion of student in low income household and proportion of student's parents have no education degree. The report aims to study the relationship between school's OSSLT pass ratio and its demographic background, and how each demographic information can influence each students chances of achieving OSSLT standard. Due to the large amount of figures, this section will showcase only the scatter plot of OSSLT pass rate by special education students. The data visualization for other demographic backgrounds are available in the appendix.

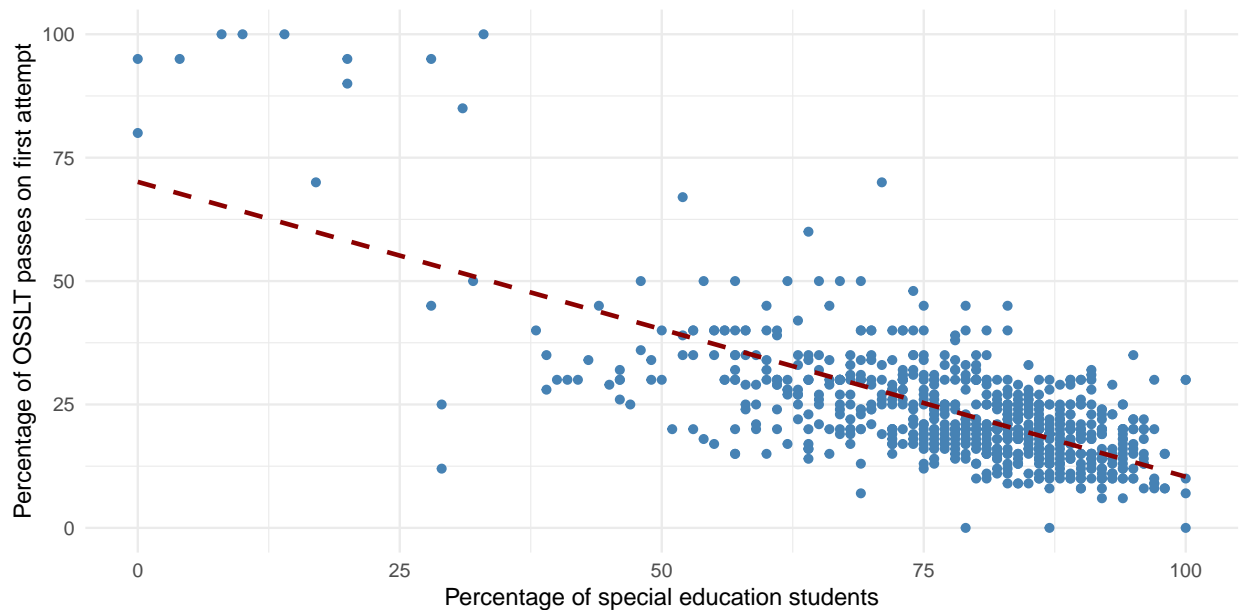


Figure 2: High school's OSSLT pass rate vs High school's percentage of SE students

3 Model

3.1 Model construction

The study wishes to investigate the effect of Ontario high schools' demographic on its academic achievement results. More specifically, OSSLTs conducted among grade 10 students every year is one of the standard measurements for academic results. Thus, `osslt_passes` representing the pass percentage of OSSLT on first attempt was used as the response variable for this model. Variables representing Ontario high school's

demographics including, *school_type*, *esl_students*, *se_students*, *gifted_students*, *low_income_household*, *no_degree_parents* were used as the explanatory variables. Since the response variable *osslt_passes* is a continuous variable, and linear relationships between *osslt_passes* and explanatory variables were shown in the data section and appendix, the study selected linear model for the analysis. Provincial and deomonstration school are removed as a school type due to its sample size of 1. Furthermore, a linear mixed effect model was used with all the explanatory variables as fixed effects and *board_name* representing the school board of the high school as a random effect. The random effect was implemented in the model due to the potential correlations of *osslt_passes* between schools within the same school board, correlations may be caused by reasons like, geographically similar demographics, or similar education curriculum and programs under the same school board. Although some correlations may exists between the explanatory variables, no interaction term was considered because intuitively none of the school demographics would influence the effect of other school demographics on OSSLT pass rates.

The model equation is as follow:

$$osslt_passes_{ij} = \beta_0 + \beta_1(school_type_public_{ij}) + \beta_2(esl_students_{ij}) + \beta_3(se_students_{ij}) + \beta_4(gifted_students_{ij}) + \beta_5(low_income_household_{ij}) + \beta_6(no_degree_parents_{ij}) + board_name_j + Residual_{ij}$$

The response for this model is the OSSLT pass rate on first attempt of high school *i* under school board *j*. Variable *school_type* is a indicator fixed effect with two levels, *catholic* and *public*, where *catholic* were selected as the reference level. Provincial was removed from the data as one of the *school_type* because there exists only one school in Ontario with such school type, the sample size was too small to be considered in the model. Variables *esl_students_{ij}*, *se_students_{ij}*, *gifted_students_{ij}*, *low_income_household_{ij}* and *no_degree_parents_{ij}* are continuous fixed effect and represents various school demographics in percentage of school *i* under the school board *j*. *board_name_j* is the random effect responsible for the correlations of OSSLT results between schools under the same school board *j*. *Residual_{ij}* is the error term account for the individual variances of each school *i* under board *j*.

3.2 Model Validation

There are 4 main assumptions for the lienar mixed effect mdoel:

1. There exists linearity between *OSSLT_passes* and other continuous fixed effect variables.
2. Random effect *board_name* follows the normal distribution with a mean of 0 and constant variance.
3. Error term *Residual* follows the normal distribution with a mean of 0 and constant variance.
4. Random effect and error term are independent.

For the first assumption, the validity of the linearity between *OSSLT_passes* and other school demographics were already mentioned in the creation of the model and the scatter plots that checks the linearity between them are available in the appendix section. Secondly, histogram and QQ plot of the random effect *board_name* were made to check validity of its normal distribution with mean of 0, and scatter plots of *board_name* against its index and fitted values of the model were also made to check the random effect's constant variance. The same 4 plots were made again for error term *Residual* to check whether it follows normal distribution with mean of 0 and its constant variance. Lastly, random effect *board_name* and error term are indeed independent to each other because error term are random values assigned each observations in the data and should be independent to all variables other than *school_id* and *school_name*.

4 Results

Table 2 showcases the summary statistics of the linear mixed effect model constructed previously. Most explanatory variables are statistically significant at a 99% confidence level, while *school_type* being the only variable significant at 95% confidence level. It is important to note that, the confidence level should not be

Table 2: The summary statistic of the fitted model

| | <i>Dependent variable:</i> |
|----------------------|-----------------------------|
| | osslt_passes |
| school_typepublic | −3.590** (1.410) |
| esl_students | 0.074*** (0.021) |
| se_students | −0.736*** (0.031) |
| gifted_students | 0.352*** (0.098) |
| low_income_household | −0.228*** (0.063) |
| no_degree_parents | −0.491*** (0.069) |
| Constant | 100.878*** (1.444) |
| Observations | 735 |
| Log Likelihood | −2,658.544 |
| Akaike Inf. Crit. | 5,335.087 |
| Bayesian Inf. Crit. | 5,376.486 |
| <i>Note:</i> | *p<0.1; **p<0.05; ***p<0.01 |

used as a strict threshold to determine variable's significance to the purpose of the study, and it should be interpreted as an assessments of observations or effects relative to sampling variation.

The intercept of the model is 100.88% of OSSLT pass rate, however it is inappropriate to interpret the intercept as a measure of OSSLT passes for schools with reference level demographics as it is impossible to have a pass rate higher than 100%. The constant should only be interpret as the intercept for the functionality of the model in the context of the provided data.

Table 2 also showcased that Ontario public schools on average have a lower OSSLT pass rate than catholic schools by 3.59%. Unexpectedly, English as second language students showed a slight positive correlation with the OSSLT pass rate, as an increasing percentage of ESL students results in an 0.07% increase in OSSLT pass rate under the context of provided Ontario high school data. On the other hand, percentage increase of special education students results in an 0.74% decrease in OSSLT pass rate, percentage increase of gifted students results in an 0.35 increase in OSSLT pass rate, percentage increase of students classified within low income households reduces OSSLT passes by 0.23% and percentage increase of students with parents that have no degree reduces OSSLT passes by 0.49%.

5 Discussion

5.1 Interpretation of results and findings

5.1.1 Catholic and public schools

School type is a crucial variable when it comes to the OSSLT passing percentage on student's first attempt. This can be shown from both figure 1 and the model summary table 2. In figure 1, catholic schools' OSSLT pass rates shares similar distribution shape with public schools but its mean and overall distribution are right to the relative position of public schools'. In 2, the model predicted that Ontario catholic schools overall have a higher OSSLT pass rate than public schools by 3.59%. Further research is required to understand why catholic students performs better on OSSLT than public students, relevant information on the operational differences between two type of schools in Ontario can be helpful. This is also an opportunity for public district school boards to study and learn from catholic district school boards on ways to improve students academic achievements.

5.1.2 English as second language students

Interestingly, the model summary table 2 indicates an increasing percentage of ESL students results in an 0.07% increase in OSSLT pass rate under the context of provided Ontario high school data. Intuitively, the result is inexplicable, as students who have English as their second language should face higher difficulty to pass the literacy test. However, upon further research, Ontario ESL students are eligible to request special accommodations such as additional test time with supervision, participation deferral, or enroll into a full-credit course as alternative. ("FAQ" 2022) According to the Ontario provincial literacy result report 2018-2019, only 75% of English language learners participated in the test.(2019) This heavily affect the model's ability to estimate the literacy skills on students with English as second language. ESL students who decides to participate in the OSSLT are more likely to pass the test due to their eligibility on extra test time. ESL students who decides defer or take alternative course will result in an over estimation of the school's OSSLT pass rate, as these students are less likely to pass the test and did not end up taking the test. Thus, the positive correlation between OSSLT passes and ESL students should definitely not be interpreted as effect of ESL students on academic literacy skills, it is only a representation of the correlation between a school's ESL students percentage and its OSSLT pass rate under the context of Ontario provincial rulings and accommodations regarding ESL students with OSSLT.

5.1.3 Special education students

For special education students, figure 2 shows a clear negative correlation between OSSLT pass rate and percentage of students with special education needs. The model summary table 2 also showcases an increasing percentage of special education students results in an 0.07% decrease in OSSLT pass rate. Unlike English language learner, special education students are not eligible for deferrals or alternative courses, but special accommodations such as extra time, sign language or an oral interpreter, unified English Braille version of the test, or additional assistive technology can be provided depending on the situation of each special education students. (“What Are the Key Guidelines and Details for Administering the OSSLT This Year?” 2021) Even with assistance of the accommodations, special education students still find themselves a harder time to pass the OSSLT, only 50% of participating special education students pass the OSSLT in 2019. (2019) Thus, although an over prediction exists from the model on the literacy skills of special education students in Ontario high school, they remain to struggle in their academic learning, and passing the OSSLT. Educators should consider provide more attention and funding to the special education program as the students not only require additional help to pass the test, but also a better and more effective learning process in high school.

5.1.4 Gifted students

Unsurprisingly, gifted students are more likely to pass the OSSLT as shown in both figure 4 and table 2. Ontario’s self-contained gifted education programs gives intellectually gifted students an opportunity to work with other like-minded peers and learn materials more relevant to their interests and abilities than in a regular class. (“Gifted Program” 2022) Therefore, it is reasonable to state that having more gifted students in the high school will results a higher OSSLT pass rate, and gifted students on average are likely to possess higher literacy skills than normal students. However, Ontario’s gifted program only admits 2% of the applicants every year, and gifted students make up a very small percentage of the overall student population. Thus, area of improvement on such matter is limited, since the program was designed for only limited amount of students. Learning from the gifted program, revising and improving the current education program for regular students may be a consideration for the purpose of improving over all literacy skills of Ontario high schools.

5.1.5 Parents’ income and education

Utilizing the data collected from 2016 census and 2019 OSSLT results, high schools with higher percentage of low income household and higher percentage of student’s parents with no diplomacy are likely to have a lower OSSLT pass rate as shown in figure 5, figure 6, and table 2. Students from a lower income households are likely to have a less ideal study environment at home and schools that locates in poor communities are also likely to provide less ideal study environment all due to the lack of financial funding and resources. On the other hand, parent’s status of low education can possibly influences the student’s perception on the importance of education. Ministry of education and school boards can consider introducing educational financial supports for students with low income households if not already implemented, and may be reallocate more funding to school located in low income neighborhood. They can also consider focus on promoting the importance of academic learning among high schools located in neighborhood with low educations overall.

5.2 Weaknesses

There are three weaknesses of this report that should be discussed. Firstly, the purpose of this paper is to study Ontario high school student’s academic literacy skills and the background characteristics that effects it. To promote data transparency while protecting individual student’s privacy, Ontario ministry of education hid the student level data, but published the school level data with various student demographics in percentage. Modeling with school level data are less accurate than student level data when determining

key student demographics that influences their academic skills, since each observation in school level data are essentially grouped summary of student level data.

Secondly, as mentioned in the interpretations of results and findings section, English as second language and special education students are granted special accommodations like deferrals, extra time and other assistance to help them pass OSSLT, as passing this test is a high school diploma requirement for all Ontario students. Thus, students who passed OSSLT on first attempt with special accommodations does not fully represents their true literacy skills and the model's estimated effect of ESL and SE students on OSSLT pass rate should only interpreted under the condition of understanding the context of OSSLT accommodation guidelines.

Finally, the creation of the school information and student demographics data set published by Ontario ministry of Education consists of multiple sources of data. As mentioned in the data section, high school information collected from OnSIS system were updated and maintained by Education at March 2022, OSSLT results were reported by the EQAO office during 2019, and other student demographics information collected from Census published by Statistic Canada were conducted during 2016. thus, there exists inconsistencies between data variables' updated dates. Using the data, the inconsistencies will likely effect the accuracy of the fitted model. School information would not be an issue as their data are less likely to be different over time. However, student demographics during 2019 are much more likely to be a bit different from 2016 when Census was conducted, and predicting 2019 OSSLT results from 2016 student demographics can potentially leads to incorrect correlations and interpretations.

5.3 Future directions

Due to given time constraint, this study were unable to successfully request student level data from Ontario ministry of Education. However, using student level data for future studies is strongly suggested if it is accessible. Modeling with student level data will be much more accurate and will solve some of the limitations presented in this paper. Additionally, using specific OSSLT score from each students instead of OSSLT passes as response variable will also increase the effect of the model by a large margin. Moreover, although it important to update the latest data, Education should consider align the update dates with all the information, as it may cause confusion and misinterpretations if data users were not notified. Lastly, OSSLT were discontinued during the Pandemic season, it would be interesting to revisit and conduct future studies once pandemic is over. The aftermath of pandemic could be considered as one of the new student demographics when determining future results of high school student's academic skills.

Appendix

.1 Additional figures

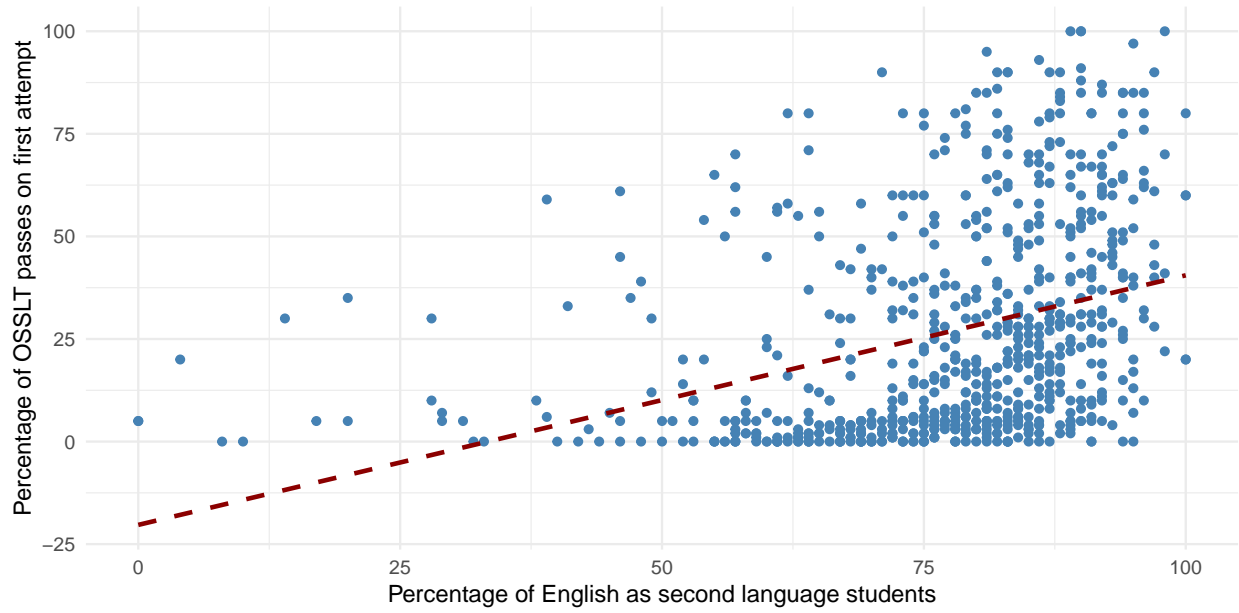


Figure 3: High school's OSSLT pass rate vs High school's percentage of ESL students

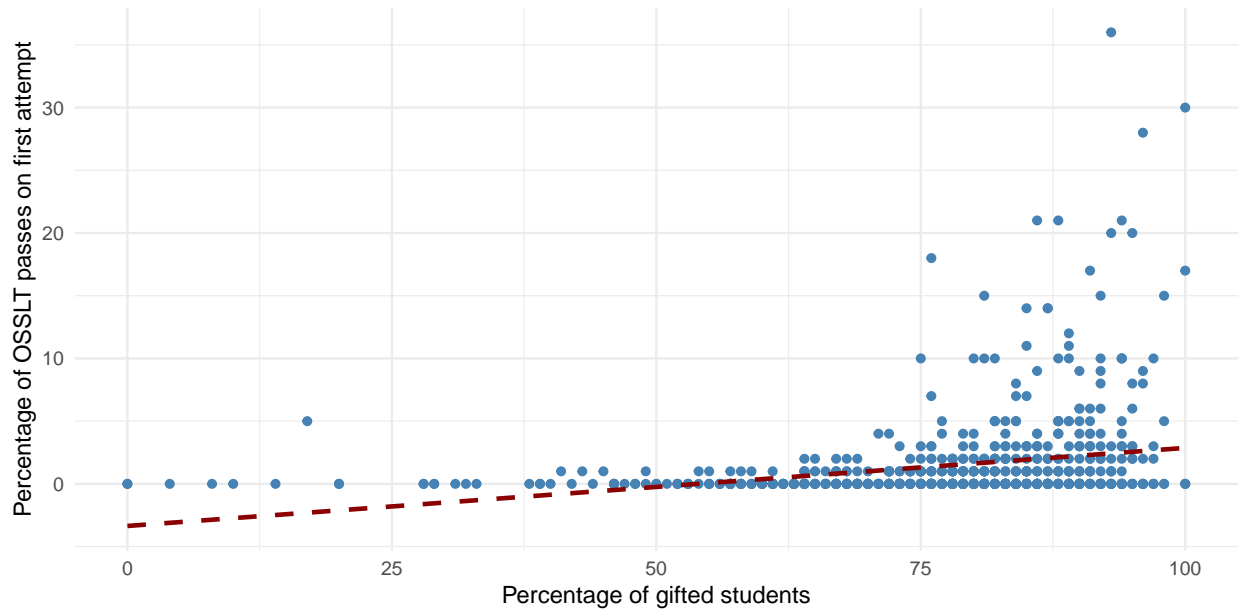


Figure 4: High school's OSSLT pass rate vs High school's percentage of gifted students

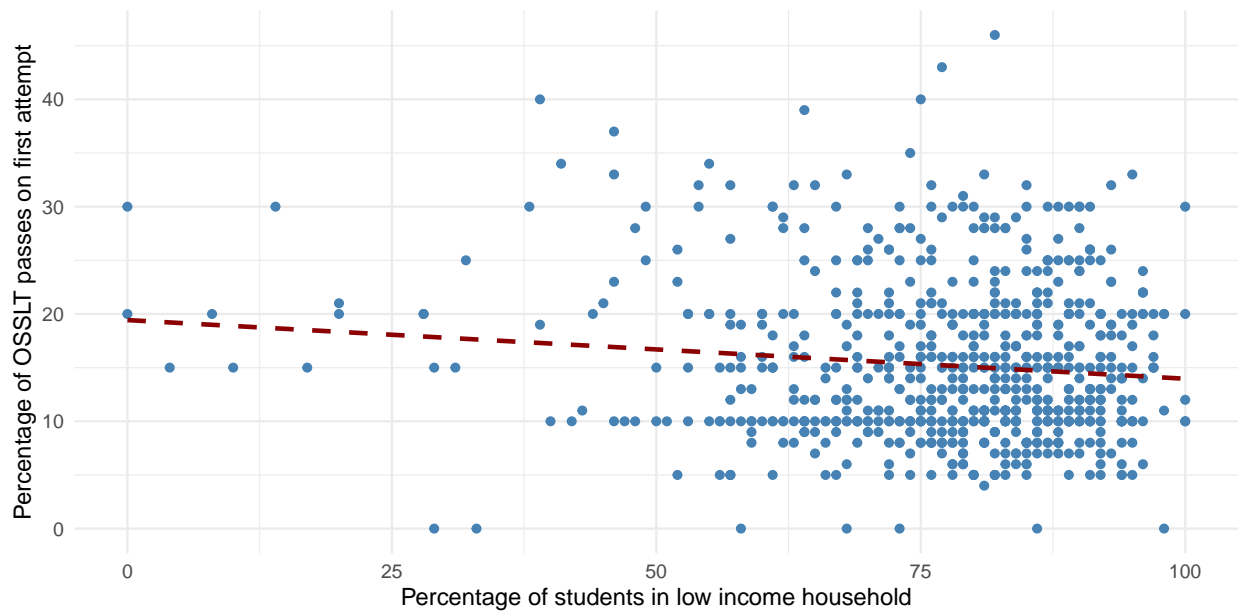


Figure 5: High school's OSSLT pass rate vs High school's percentage of students in low income household

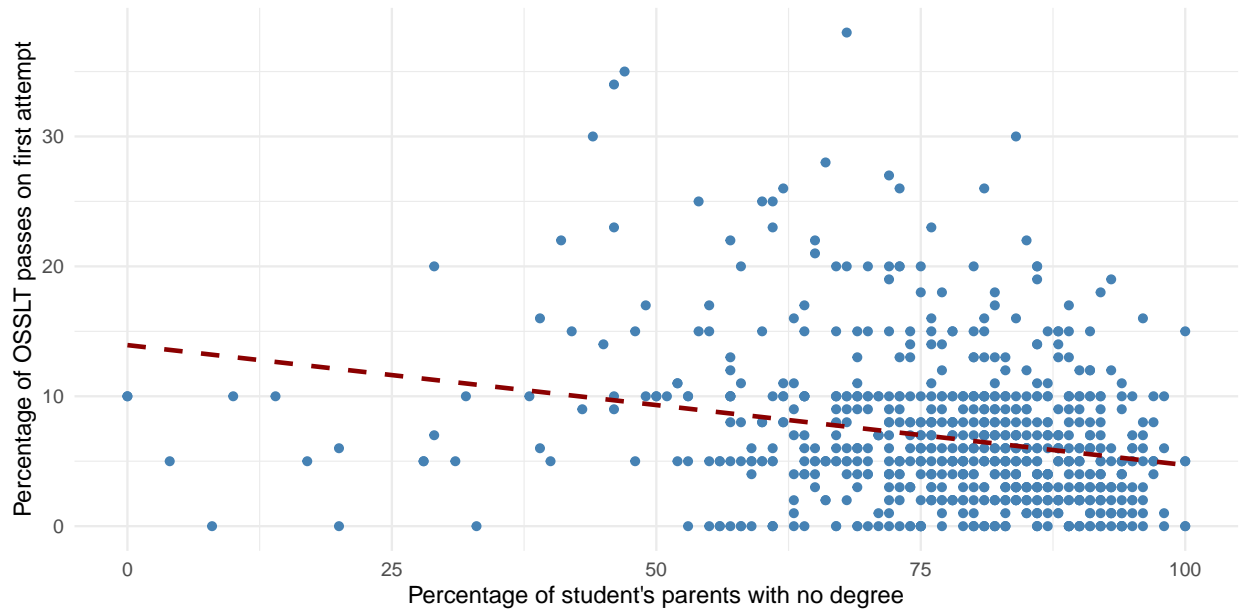


Figure 6: High school's OSSLT pass rate vs High school's percentage of student's parents with no degree

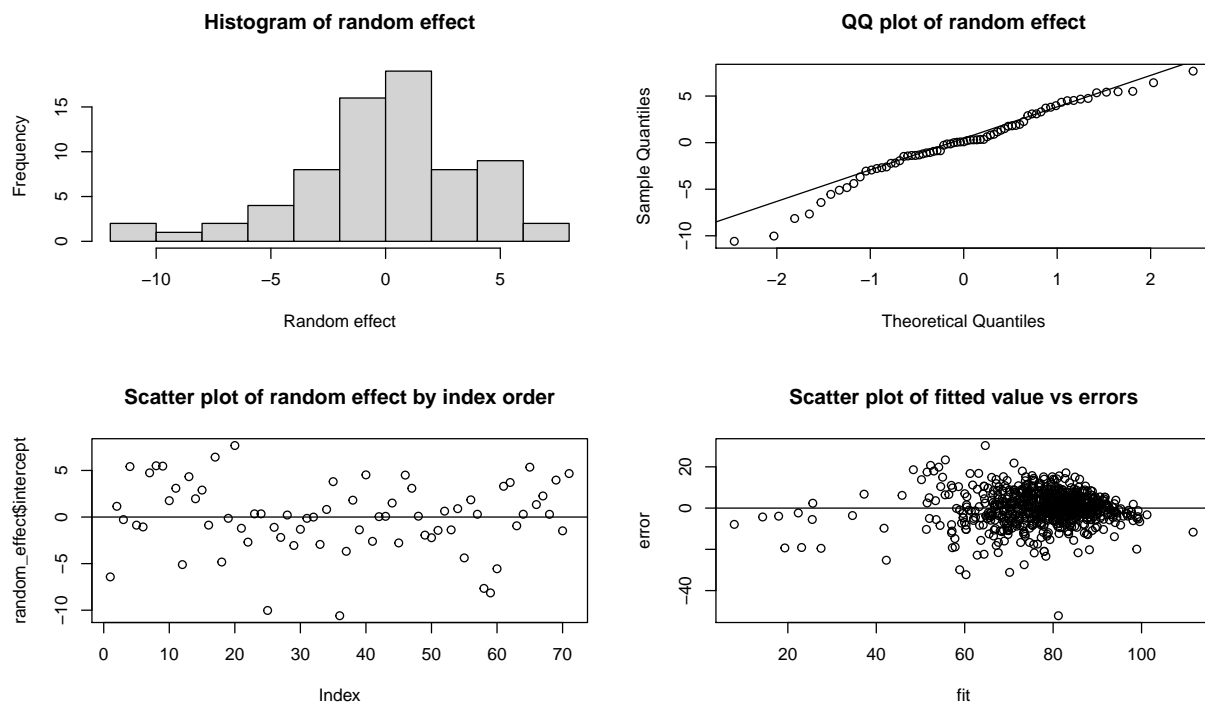


Figure 7: The graphs of checking assumptions for random effect in the model

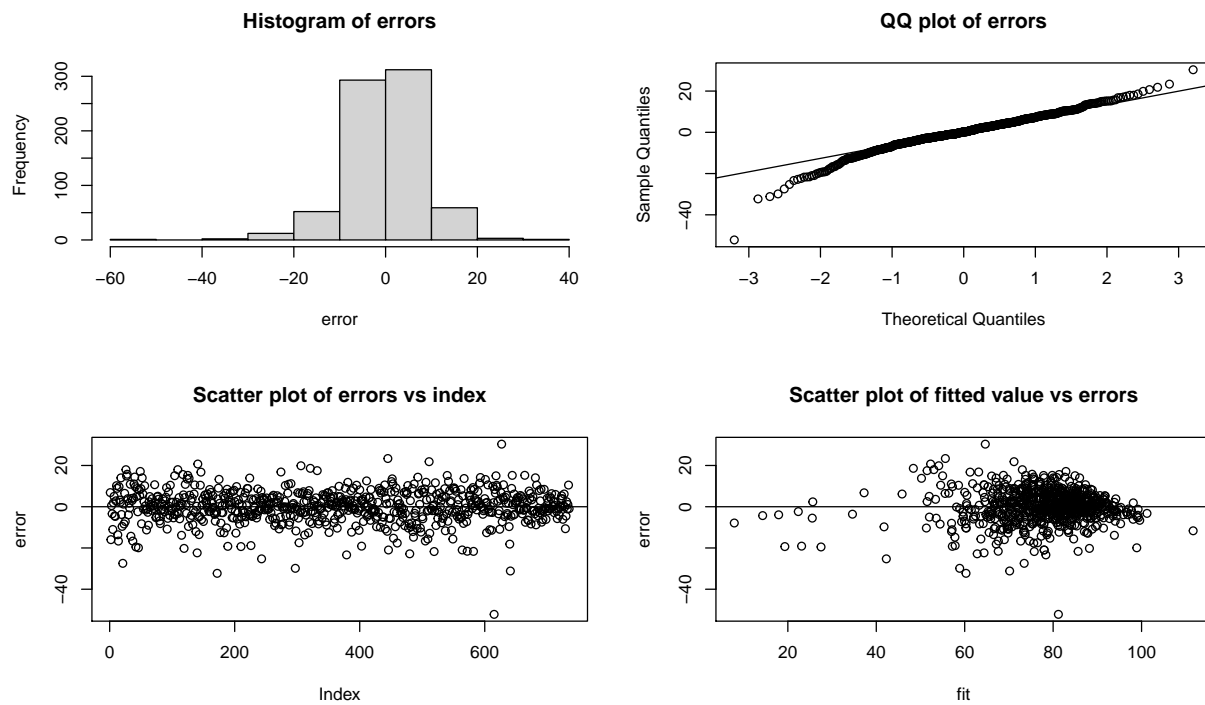


Figure 8: The graphs of checking assumptions for errors in the model

.2 Datasheet

Motivation

1. For what purpose was the dataset created?
 - Using data extracted from school and student demographic report by Education, this analysis aims to study Ontario high school's student passing rate on OSSLT test by analysing the profile and demographics of each school specifically in school type and ratio statistics like ESL students, special education students, gifted students, low-income household, and parents with no education degree.
2. Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
 - Ontario Ministry of Education
3. Who funded the creation of the dataset?
 - Ontario government funds and maintains the dataset

Composition

1. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?
 - School information and student demographics: school name, school type, board name, location, enrollment, EQAO and OSSLT results, gifted students ratio etc.
2. How many instances are there in total (of each type, if appropriate)? -736 Ontario high school
3. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
 - dataset contains all possible instances of Ontario high schools that have full information in the data.
4. Is there a label or target associated with each instance?
 - School numbers and School names are available in the dataset.
5. Is any information missing from individual instances?
 - Multiple instances are missing all sorts of information.
6. Are there any errors, sources of noise, or redundancies in the dataset?
 - There are no errors, sources of noise or redundancies in the dataset.
7. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?
 - The dataset is self-contained.

8. Does the dataset contain data that might be considered confidential?
 - There is no confidential data, the dataset is publicly available.
9. Does the dataset identify any subpopulations (e.g., by age, gender)?
 - The dataset comprises of all Ontario high school, and high school students.
10. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?
 - It is not possible to identify individuals.
11. Does the dataset contain data that might be considered sensitive in any way?
 - No
13. Any other comments?
 - None.

Collection Process

1. How was the data associated with each instance acquired?
 - The data set was collected by the Ministry of Education in Ontario, Canada. Details can be found in the Data Methodology section of the report.
2. What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?
 - Data set was collected from Ontario Open Data Catalogue
3. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
 - It is not a sample, it includes all of the possible instances.
4. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
 - Statistic Canada for collecting Census data, EQAO office for recording OSSLT results, and Ontario Ministry of Education for recording School information
5. Over what timeframe was the data collected?
 - The data consists of data from 2016, 2019 and 2022.

6. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
 - The data is obtained from the official website of the Ontario Open Data Catalogue.
7. Were the individuals in question notified about the data collection?
 - The individuals of the question are well aware of the survey as well as the data collection.
8. Did the individuals in question consent to the collection and use of their data?
 - The individuals in question have consented to the collection and use of their data.
9. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?
 - There's no open information available to tell if the consent for revokes were granted to the survey participants.
10. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?
 - No
12. Any other comments?
 - None.

Preprocessing/ cleaning/ labeling

1. Was any preprocessing/cleaning/labeling of the data done?
 - The data was originally obtained in excel format and was converted and cleaned to a tidy csv file using R
2. Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?
 - The raw data obtained is saved in outputs/data/raw_data.csv
3. Is the software that was used to preprocess/clean/label the data available?
 - R software is available at <https://www.R-project.org/>

Uses

1. Has the dataset been used for any tasks already?
 - The dataset has not been used for other tasks yet.

2. Is there a repository that links to any or all papers or systems that use the dataset?

- https://github.com/R300G/Ontario_School_Data

3. What (other) tasks could the dataset be used for?

- The dataset would be used for understanding the demographics that effect the results of OSSLT

4. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

- The cleaning process is very generic and can used for future reports.

Distribution

1. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

- No. The data set is for personal use only.

2. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?

- The dataset can be found in the Github repository mentioned in the report.

3. When will the dataset be distributed?

- The dataset has already been distributed in April 2022.

4. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

- The report is under the MIT license.

5. Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

- There are no IP-based or other restrictions on the data.

6. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

- No.

7. Any other comments?

- None.

Maintenance

1. Who will be supporting/hosting/maintaining the dataset?

- The author Charles Lu of the report will be maintaining the dataset.
2. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
- They can be contacted via the email addresses listed in the README file on Github.
3. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?
- No dataset updates are scheduled yet.
4. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?
- The data set was collected and published as open data. There are no application limits on the retention of the data since it was collected voluntarily.
5. Will older versions of the dataset continue to be supported/hosted/maintained?
- No. Only the current data is available for now and future.
6. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?
- The extensions, augmentations and contributions from other users are not supported.

References

2019. EQAO. <https://www.eqao.com/wp-content/uploads/provincial-report-secondary-2019.pdf>.
- Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. “Fitting Linear Mixed-Effects Models Using lme4.” *Journal of Statistical Software* 67 (1): 1–48. <https://doi.org/10.18637/jss.v067.i01>.
- “FAQ.” 2022. *ERGO*. ESL/ELD Resource Group of Ontario. <https://www.ergo-on.ca/FAQ.htm>.
- Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. <https://github.com/sfirke/janitor>.
- “Gifted Program.” 2022. *DDSB*. DDSB. <https://www.ddsbc.ca/en/programs-and-learning/gifted-program.aspx#Gifted-Child-Today>.
- Hlavac, Marek. 2022. *Stargazer: Well-Formatted Regression and Summary Statistics Tables*. Bratislava, Slovakia: Social Policy Institute. <https://CRAN.R-project.org/package=stargazer>.
- “Ontario Secondary School Literacy Test (OSSLT).” 2022. *EQAO*. <https://www.eqao.com/the-assessments/osslt/>.
- R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- “School Information and Student Demographics - Ontario Data Catalogue.” 2022. *School Information and Student Demographics - Datasets - Ontario Data Catalogue*. Ontario Ministry of Education. <https://data.ontario.ca/dataset/school-information-and-student-demographics>.
- “What Are the Key Guidelines and Details for Administering the OSSLT This Year?” 2021. *EQAO*. <https://eqao-kb-osslt.lswp.vretta.com/knowledge-base/what-are-the-key-guidelines-and-details-for-administering-the-osslt-this-year/>.
- Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Golemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.
- Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. <https://CRAN.R-project.org/package=dplyr>.
- Xie, Yihui. 2021a. *Bookdown: Authoring Books and Technical Documents with r Markdown*. <https://github.com/rstudio/bookdown>.
- . 2021b. *Tinytex: Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents*. <https://github.com/yihui/tinytex>.
- Zhu, Hao. 2021. *kableExtra: Construct Complex Table with ‘Kable’ and Pipe Syntax*. <https://CRAN.R-project.org/package=kableExtra>.