



University of Kentucky
UKnowledge

Theses and Dissertations--Education Science

College of Education


2020

ADVANCED MULTILEVEL MODELS FOR COMPARING GROUP CHARACTERISTICS: THE CASE OF SEX DIFFERENCES IN READING ACHIEVEMENT

Rongxiu Wu

University of Kentucky, rwu227@g.uky.edu

Author ORCID Identifier:

 <https://orcid.org/0000-0003-0457-2738>

Digital Object Identifier: <https://doi.org/10.13023/etd.2020.069>

[Right click to open a feedback form in a new tab to let us know how this document benefits you.](#)

Recommended Citation

Wu, Rongxiu, "ADVANCED MULTILEVEL MODELS FOR COMPARING GROUP CHARACTERISTICS: THE CASE OF SEX DIFFERENCES IN READING ACHIEVEMENT" (2020). *Theses and Dissertations--Education Science*. 57.

https://uknowledge.uky.edu/edsc_etds/57

This Doctoral Dissertation is brought to you for free and open access by the College of Education at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Education Science by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Rongxiu Wu, Student

Dr. Xin Ma, Major Professor

Dr. Michael D. Toland, Director of Graduate Studies

ADVANCED MULTILEVEL MODELS
FOR COMPARING GROUP CHARACTERISTICS:
THE CASE OF SEX DIFFERENCES IN READING ACHIEVEMENT

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in the
College of Education
at the University of Kentucky

By
Rongxiu Wu
Lexington, Kentucky
Director: Dr. Xin Ma, Professor of Educational, School, and Counseling Psychology
Lexington, Kentucky
2020

Copyright © Rongxiu Wu 2020

ORCID ID: <https://orcid.org/0000-0003-0457-2738>

ABSTRACT OF DISSERTATION

ADVANCED MULTILEVEL MODELS FOR COMPARING GROUP CHARACTERISTICS: THE CASE OF SEX DIFFERENCES IN READING ACHIEVEMENT

To help improve and advance research methodology when comparing the group characteristics, two advanced multilevel models were developed and introduced, which would allow a deeper and more refined look at the issue of sex differences in reading achievement.

The first model is a restricted multilevel model for the examination of institutional effects on multiple groups of individuals. The goal of this multivariate multilevel model with individuals nested within institutions was to estimate the institutional effects on multiple groups of individuals. With the employment of 2009 OECD Programme for International Student Assessment (PISA) data, an application was illustrated to examine whether school reading environment had the same effect on reading achievement between boys and girls. In this two-level model, the level 1 was a multivariate model highlighting students' average reading achievement for each sex group (two dichotomous variables) and level 2 was two linear regression equations, one for boys and one for girls. The effects of five school reading environment variables (*diversity of reading, enjoyment of reading, stimulators of reading, daily reading hours, and online reading hours*) were constrained respectively to be the same for both boys and girls. A significance test was performed to examine whether this restriction held true. It was found that the effects of *enjoyment of reading* and *online reading hours* were statistically different on reading achievement between boys and girls based on PISA 2009 dataset. The model is an effective omnibus statistical technique to examine the institutional effects on multiple groups of individuals, which unmasked the specific group dynamics concerning institutional effects with a broad applicability as well as convenient execution.

The second model was a multilevel model with heterogeneous sigma squared function to compare distributional properties of multiple groups. A good understanding of the distributional properties across groups is an essential part of making group comparisons. The combination of central tendency and variability is the preferred way to describe (and compare) distributions across groups. An advanced multilevel model with

an embedded analytic function referred to as heterogeneous sigma squared was developed to perform statistical tests of significance to compare means and variances across multiple groups at the same time, which made it convenient to examine the distributional properties comprehensively and simultaneously. With the employment of 2009 OECD PISA data, an application was illustrated to examine the distributional properties concerning reading achievement for boys and girls. In the two-level model, the level one had sex as the categorical independent variable (dummy coded as boys = 0 and girls = 1) and level two had the random intercept modeled by school background variables. It was found that girls performed significantly better than boys in reading achievement, but boys and girls share similar variance in reading achievement. A violin plot revealed that girls had higher mean and occupied the very top distribution of reading achievement, while boys had a lower mean and occupied the very bottom of reading achievement. The distribution for girls was near normal, but there were two peaks for boys indicating that the distribution for boys was not normal. The full model explained a total of nearly a third of the variance in reading achievement.

The above advanced multilevel models can be easily extended to examine other equity issues in education. It is the hope of the author that these advanced multilevel models would inspire statistical efforts in developing other advanced models. The results of similar models may promote more credible educational reforms through a revisit to educational policies and practices concerning equity issues in education (based on more robust and precise empirical evidence).

KEYWORDS: Sex Differences, School Effect, Reading Achievement, Distributional Properties

Rongxiu Wu

03/30/2019

ADVANCED MULTILEVEL MODELS
FOR COMPARING GROUP CHARACTERISTICS:
THE CASE OF SEX DIFFERENCES IN READING ACHIEVEMENT

By

Rongxiu Wu

Dr. Xin Ma

Director of Dissertation

Dr. Michael D. Toland

Director of Graduate Studies

March 30th, 2020

DEDICATION

To my parents who have supported me all along and myself who choose to continue learning and serving as a way of life.

ACKNOWLEDGMENTS

This dissertation, while an individual work, benefitted greatly from the wisdom and direction of my Dissertation Chair, Dr. Xin Ma. His specialty instruction in the methods of multilevel modeling has inspired me to do this research; his ongoing patient assistance was even more critical in developing the project and preparing the dissertation itself. In addition, I also want to thank Dr. Michael Toland for his timely and instructive comments and evaluation at each stage of dissertation process, allowing me to keep reflecting on my work. In addition to Dr. Ma and Dr. Toland, I also wish to thank Drs. Shake and Dr. Burns as members of my committee for their insights on sharpening the understanding of differences between sex groups on reading achievement in the research. Thanks as well to Dr. Christopher Burns for his willingness to engage in this process as an outside examiner.

In addition to the essential support provided by the five academic experts above, I must acknowledge with profound gratitude the fellow students at the College of Education with whom I have been privileged to learn over the past five years. I am very grateful to them for being my patient audience again and again.

Lastly, as a first-generation college student, I am glad that my parents have never stopped me to move forward from my rural roots in China to a major university in the U.S. and have been pleased with the academic opportunities I have had the privilege to take advantage of in my life.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	iii
LIST OF TABLES	vi
CHAPTER 1: Statement of the Problem.....	1
1.1 Introduction to Study 1 on a Restricted Multilevel Model for Examining the Institutional Effects on Multiple Groups of Individuals	1
1.2 Introduction to Study 2 on a Multilevel Model with Heterogeneous Sigma Squared Function to Compare the Distributional Characteristics of Multiple Groups	5
1.3 Methodological Significance of the Studies	7
1.4 Practical Significance of the Studies.....	8
1.5 Organization of the Study	9
CHAPTER 2: A Restricted Multilevel Model for Examination of Institutional Effects on Multiple Groups of Individuals.....	10
2.1 The Model.....	10
2.2 The Assumptions	13
2.3 The Estimation.....	13
2.4 The Application	14
2.4.1 Model Specification	14
2.4.2 Literature review	17
2.4.3 Data Source.....	19
2.4.4 Analytical Procedure.....	24
2.4.5 Results.....	25
2.5 Final Remarks on the Restricted Multilevel Model	30
2.5.1 Model Performance.....	31
2.5.2 Model Extension	31
2.5.3 Model Limitation	32
CHAPTER 3: A Multilevel Model with Heterogeneous Sigma Squared Function to Compare Distributional Properties of Multiple Groups	37
3.1 The Model.....	37

3.2 The Assumptions	40
3.3 The Estimation	41
3.4 The Application	41
3.4.1 Model Specification	41
3.4.2 Literature Review.....	43
3.4.3 Research Questions	47
3.4.4 Data Sources	47
3.4.5 Analytical Procedures	50
3.4.6 Results.....	51
3.5 Final Remarks	53
CHAPTER 4: Summary.....	58
4.1 Motivation for Methodological Advancement.....	58
4.2 Methodological Advancement	59
4.3 Applications of Advanced Multilevel Models	61
4.4 Tentative Practical Contributions.....	63
4.5 Limitations and Suggestions	64
Appendix A.....	66
Appendix B.....	68
Appendix C.....	70
Appendix D.....	71
Appendix E	72
References.....	73
Vita.....	80

LIST OF TABLES

Table 2.1	33
<i>Estimates of Absolute Effects of School Reading Environment on Reading Achievement of Boys and Girls</i>	33
Table 2.2	33
<i>Estimates of Relative Effects of School Reading Environment on Reading Achievement of Boys and Girls, Controlling for Student Characteristics</i>	33
Table 2.3	34
<i>Estimates of Relative Effects of School Reading Environment on Reading Achievement of Boys and Girls, Controlling for Student Characteristics and School Characteristics</i>	34
Table 2.4	34
<i>Estimates of Variance Components and Proportion of Variance Explained for Enjoyment of Reading</i>	34
Table 2.5	35
<i>Estimates of Variance Components and Proportion of Variance Explained for Online Reading Hours</i>	35
Table 3.1	55
<i>HLM Models of Heterogeneous Sigma Squared Comparing Means and Variances between Boys and Girls in Reading Achievement</i>	55
Table 3.2	56
<i>Estimates of Variance Components and Proportion of Variance Explained for Reading Achievement</i>	56

CHAPTER 1: Statement of the Problem

1.1 Introduction to Study 1 on a Restricted Multilevel Model for Examining the Institutional Effects on Multiple Groups of Individuals

Institutions have indispensable effects on groups of individuals. One such example is the effect of schools on students' academic achievement. Schools have been recognized as non-negligible institutions in impacting students' academic achievement (Ma, Ma, & Bradley, 2008; Marks, 2008; Walkerdine, 1988). The "added-value" of the schools to the academic achievement of students cannot be overlooked (Everson & Millsap, 2004; Lee, Zuze & Ross, 2005; Opdenakker & Dammer, 2006). But how do we usually examine the institutional effects on multiple groups of individuals? Study 1 aims to examine this issue. The goal of Study 1 is to propose a general statistical model that can be used to address this issue and to apply this model to the examination of school effects on sex groups in the area of reading education.

Due to the obvious hierarchical structure of social institutional systems (e.g. patients nested in clinics nested in states; students nested in classes nested in schools), multilevel modeling has become a required and popular methodology in the field of institutional effectiveness, such as school effectiveness research in which the hierarchical structure of student-level and school-level variables are included in the model (Lee & Bryk, 1989; Goldstein, 1995; Snijder & Bosker, 1999; Raudenbush & Bryk, 2002; Ma et al., 2008; Opdenakker & Dammer, 2000). Being regressive in nature, multilevel modeling techniques are excellent and powerful ways to establish relationships, which are far more credible than any traditional ways (e.g., multiple regression) for the same purpose (Goldstein, 1995; Rasbash et al., 2000; Raudenbush & Bryk, 2002).

The common way to study the institutional effect on multiple groups of individuals is, initially, to take the group variable such as sex as a dummy variable (boy = 0 and girl = 1 or inversely) or groups of dummy variables when there are multiple groups. This common use of dummy coding for group variables to mimic one-to-one group comparison has been criticized for covering up important information about group dynamics (Ma, 1999), which could come to light as a result of the decomposition of interaction effects among independent variables. The dummy system has an inevitable disadvantage. The following is a typical multilevel model to examine institutional effects (IE) on sex groups (female is coded as 1 and male is coded as 0).

$$Y_{ij} = \beta_{0j} + \beta_{1j} * Female_{ij} + \varepsilon_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01} * IE_j + U_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} * IE_j + U_{1j}$$

where β_{0j} is the intercept, which in fact is the mean achievement for males in institution j ; β_{1j} is the sex gap in institution j (i.e. the mean difference of achievement between boys and girls); ε_{ij} is an error term assumed to be normally distributed with mean of zero and homogeneous (constant) variance σ^2 across institutions, that is, $\varepsilon_{ij} \sim N(0, \sigma^2)$; γ_{00} is the average mean of achievement across the institutions; γ_{10} is the average sex gap across the institutions; U_{0j} is the error term unique to the intercept associated with the institution j ; U_{1j} is the error term unique to the slope associated with the institution j ; U_{0j} and U_{1j} are multivariate normally distributed, with mean of zero and variance-covariance matrix.

Institution-level variables may exert different effects on sex groups. The above multilevel model can indeed distinguish the differential effects of institutions on sex groups. Note that β_{0j} can be considered the male average measure (when female takes on

the value of 1) and, therefore, γ_{01} is the institutional effects on the male group. Also, note that β_{1j} is the difference between females and males so that $\beta_{0j} + \beta_{1j}$ is the female average measure and, therefore, $\gamma_{01} + \gamma_{11}$ is the institutional effects on the female group. However, there is no direct significance test in the above multilevel model that compares γ_{01} with $\gamma_{01} + \gamma_{11}$. For example, if neither γ_{01} nor γ_{11} is statistically significant, whether or not $\gamma_{01} + \gamma_{11}$ is statistically significant cannot be determined from the above multilevel model. As a result, whether IE exerts the same or different effects on the sex groups cannot be tested using the dummy-coding approach. In other words, whether the same institutional variable has the same strength across the groups is hidden in the model.

If student-level control variables such as race are included in the above multilevel model, it becomes even more difficult to figure out the institutional effect on male and female groups. For example, in the case of race (coded white = 0 and non-white = 1), the intercept becomes the average measure no longer for males but actually for white males. This simple example effectively serves to illustrate that the above multilevel model cannot be used to address the issue of institutional effects on male and female groups.

This limitation has caused researchers to consider separate group comparisons (e.g. boys and girls). For example, Ma (1999) attempted to single out males and females for separate analyses. However, such a univariate approach (that examines each sex group in isolation) has its own problems. With two univariate multilevel models, Ma (1999) cannot compare whether the same institutional variable affects males and females with the same strength. As a result, Ma (1999) did not resolve the lack of test for statistical significance between the male and the female effect. Thus, the issue remains of how to effectively compare institutional effects on groups of individuals. More advanced

(and more general) analytical frameworks are needed to put the separate analyses together in one model and make the comparisons. It is a challenge, and this is where the restricted multilevel model shows its potential (Raudenbush & Bryk, 1986).

In the restricted multilevel model, the effects for groups (e.g. male effect and female effect) from the same institutional variable can be forced to be equal, and a significance test can be performed to examine if this restriction holds true. The same multilevel model can estimate the amount of the difference, if the difference really exists. The restricted multilevel model has been applied by Barnett, Brennan, Raudenbush, & Marshall (1993) to estimate the association between marital-role quality and psychological distress in a sample of 300 full-time married couples. The following was their multilevel model but modified to compare with the above multilevel model.

$$Y_{ij} = \beta_{1j} * (M_{ij}) + \beta_{2j} * (F_{ij}) + \epsilon_{ij}$$

$$\beta_{1j} = \gamma_{10} + \sum \gamma_{1q} * W_{qj} + U_{1j}$$

$$\beta_{2j} = \gamma_{20} + \sum \gamma_{2q} * W_{qj} + U_{2j}$$

where γ_{10} and γ_{20} are the intercepts for males and females respectively; γ_{1q} is the effect (equivalent to IE) of the q_{th} predictor for males; γ_{2q} is effect (equivalent to IE) of the q_{th} predictor for females. Each sex group is modeled directly from a function of predictor variables. Constrains are then made to the corresponding coefficients (γ_{1q} and γ_{2q}), and the significance test indicates whether these coefficients (again equivalent to IE) are the same for the male and female groups.

Inspired by Barnett et al. (1993), this study will attempt to create a general multilevel platform to test the institutional effects on multiple groups where groups are

analyzed both separately and collectively in a multivariate environment so that the effects of institutional variables can be compared directly among different groups (i.e., group comparisons). This general multilevel platform can accommodate any number of groups (as to be discussed in detail in Chapter 2). This platform will then be applied to data from the Program for International Student Assessment (PISA) to examine school effects on reading achievement of males and females. Specifically, the effects of five school variables descriptive of school reading environment (i.e., diversity of reading, enjoyment of reading, online reading hours, stimulators of reading, and daily reading hours) will be examined for differences between males and females in a multilevel model with student background and school characteristics adjusted over the effects.

1.2 Introduction to Study 2 on a Multilevel Model with Heterogeneous Sigma Squared Function to Compare the Distributional Characteristics of Multiple Groups

Undoubtedly, a good understanding of the distributional properties across multiple groups of individuals is essential in making group comparisons. How do researchers usually compare the distributional properties of multiple groups of individuals? The literature on large-scale assessments indicates that the most popular method is to use basic central tendency statistics, such as differences in means and percentages (Feingold, 1992a, 1992b; Hedges & Friedman, 1993a, 1993b; Johnson, 1996; Nowell & Hedges, 1998; Litez, 2006; Shiel, 2016). Feingold (1992a, 1992b) stated that the research on sex differences in intellectual abilities has focused generally on male-female mean differences in average performance. Litez's (2006) meta-analysis of sex differences in large-scale assessments between 1970 and 2002 in the area of reading achievement confirmed this mean-based comparison.

Are mean-based comparisons adequate to capture the differences in distributional properties across multiple groups of individuals? The answer to this question may come from the cases in mathematics education. It is documented that, in general, boys' mean in mathematics achievement were higher than that of girls in mathematics achievement. Nonetheless, boys were found to occupy both the top and bottom of the achievement distribution while females were sandwiched in between (Feingold, 1992a, 1992b; Beller & Gafni, 1996; Halpern, Benbow, Geary, Gur, Hyde, & Gernsbacher, 2007; Bayes & Monseur, 2016). This case illustrates that a solo focus on differences in means across multiple groups of individuals is not adequate to capture the differences in distributional properties across multiple groups of individuals.

There is some awareness in the literature of the need to examine the variance difference in addition to the mean difference across multiple groups of individuals (e.g., Feingold, 1988; Feingold, 1992a, 1992b; Feingold, 1994; Hedges & Friedman, 1993a, 1993b; Humphrey, 1988). Lynn & Mikk (2009) found sex differences in the variance of reading achievement in all international studies they examined, in which boys showed greater variance in reading comprehension than girls in all countries with analysis of Program for International Student Assessment (PISA) and the Progress in International Reading Literacy Study (PIRLS) datasets. Hedges & Nowell (1995) looked at the trends in sex differences in academic achievement from the aspects of differences in mean, variance, and extreme score across the entire achievement distribution through 1960 to 1994.

However, some issues have still been overlooked by researchers. Most variance studies on multiple groups are operated group by group for statistical testing of variance.

So far, these tests have been performed outside of a certain statistical model that examine mean differences as a stand-alone procedure. There is a lack of credible statistical models that provide a function for tests to be performed inside or within a certain statistical model that examines mean differences. This study aims to fill in this gap, particularly in the multilevel modeling literature. Specifically, an advanced multilevel model will be developed that has the function to test for differences in variance across multilevel groups of individuals. Such a multilevel model can be referred to as a multilevel model with heterogeneous sigma squared function. This multilevel model will provide statistical tests of significance on key distributional properties including central tendency (i.e. mean) and variability (i.e. variance). The results may facilitate a graphic illustration or a visual appreciation of distributions across multiple groups of individuals. This approach will allow researchers to examine and compare variability differences on both the lower and upper end of achievement distribution across groups.

With an evaluative focus shifting to include variance, some benchmarks developed in sex difference studies may help further quantify the distributions. The variance ratio calculated by the variance of one sex in relation to that of the other sex may be useful (Glass & Hopkin, 1984; Feingold, 1992a; Ma, 1995; Nowell & Hedges, 1998; Brozo et al., 2008). The empirical benchmarks are effect sizes on the mean as well as the percentiles 5, 10, 90 and 95, which could help illustrate more substantial differences in extreme scores (Bayes & Monseur, 2016).

1.3 Methodological Significance of the Studies

The two studies of this dissertation research target the methodological weaknesses of the research literature concerning institutional effects and distributional properties on

multiple groups of individuals. For study 1, the restricted multilevel modeling has rarely been applied to the research literature on group comparisons. With the application of this methodology, multivariate analysis combining groups meets the necessary condition to conduct a credible group comparison concerning institutional effects. The advantage of carrying out multivariate analysis instead of a series of univariate statistical tests is to deflate the Type I error rate as well as gain more statistical power. For study 2, the comparison of distributional properties using heterogeneous sigma squared as an integral part of a multilevel model is even rarer in the research literature. This innovative advancement of multilevel modeling would allow researchers to compare the means and the variances in outcomes across groups simultaneously.

1.4 Practical Significance of the Studies

As a result of the application of these advanced multilevel models, Study 1 may provide empirical evidence on how school reading environment, a collective condition under which students learning about reading, affects student reading achievement between boys and girls. Study 2 intends to provide a more efficient and effective way to describe and compare distributional properties of student reading achievement between boys and girls. Together, the studies may promote an exploration in the reading literacy field to add informative insight to the literature of sex differences in reading achievement. It targets the weaknesses of the research literature on sex-related issues concerning reading achievement. In the literacy literature field, the results of these studies may promote more insightful and more credible educational reforms through revisiting educational policies and practices concerning sex differences in student reading achievement (based on more robust and precise empirical evidence). It is also the

motivation of this dissertation research to understand the mechanisms behind sex differences in student reading achievement so as to achieve better sex equality in reading education through educational reform in school reading environment.

1.5 Organization of the Study

The organization of this dissertation is twofold. Chapter 2 contains Study 1 that attempts to develop a multilevel model that identifies the extent to which institutional effects differ across multiple groups of individuals. As an application of this multilevel model, the effects of school reading environment on student reading achievement between 15-year-old boys and girls with and without controls over student and school characteristics have been examined. Chapter 3 contains Study 2 that aims to compare distributional characteristics of multiple groups of individuals by developing an advanced multilevel model to perform statistical tests of significance on distributional properties including central tendency (i.e., mean) and variability (i.e., variance). An application is made to compare, both statistically and graphically, the distributional characteristics of the reading achievement between boys and girls.

Copyright © Rongxiu Wu 2020

CHAPTER 2: A Restricted Multilevel Model for Examination of Institutional Effects on Multiple Groups of Individuals

2.1 The Model

Given the statistical structure that individuals are nested within institutions, a multilevel model is commonly employed to examine the institutional effects on multiple groups of individuals (e.g., sex groups). Due to the limitation of univariate (multilevel) analyses for group comparison that tend to mask specific group dynamics concerning institutional effects, multivariate multilevel analysis separating groups into one multivariate model becomes a necessary condition for a credible comparison between groups (Ma & Ma, 2004; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). Following this logic, the present study aims to develop a general multivariate multilevel framework (model) specifically to estimate the institutional effects on multiple groups of individuals.

The multilevel model contains two levels, with individuals nested within institutions. The first level contains a key grouping variable with a number of other variables that can function as control variables to adjust for group comparison. The key grouping variable has n categories. Instead of the common dummy coding (resulting in $N - 1$ dichotomous variables leaving out a reference category), N dichotomous variables are created to represent each group. The model at this level, thus, intends to set up a multivariate environment for the analysis, with the N dichotomous variables denoting the N groups:

$$Y_{ij} = \beta_{1j} * (X_{1ij}) + \beta_{2j} * (X_{2ij}) + \dots + \beta_{Nj} * (X_{Nij}) + \sum_{m=1}^M \beta_{(N+m)j} * \\ \textit{Individual}_{mij} + r_{ij}$$

$$X_{nij} = \begin{cases} 1, & \text{for group } n \text{ (} n = 1, 2, 3, \dots N \text{)} \\ 0, & \text{for other groups} \end{cases}$$

where Y_{ij} is the outcome score for individual i at institution j ; β_{nj} ($n = 1, 2, 3, \dots N$) is the average outcome score for group n at institution j ; X_{nij} ($n = 1, 2, 3, \dots N$) is an indicator for group n (more precisely for individual i in group n at institution j). The N average outcome scores, one for each group, can be adjusted by individual characteristics. **Individual** $_{mij}$ ($m = 1, 2, 3, \dots M$) represents these individual characteristics as controlling variables in individual level. Finally, r_{ij} is the error term specific to individual i at institution j , which is assumed to be normally distributed with a mean of zero and variance σ^2 .

$$r_{ij} \sim \text{NID}(0, \sigma^2)$$

The second level of the multivariate multilevel model includes two sets of regressions. The first set aims to model institutional effects on multiple groups of individuals:

$$\beta_{nj} = \gamma_{n0} + \gamma_{n1} * IE_j + \sum_{p=1}^P \gamma_{n(p+1)} * \text{Institution}_{pj} + U_{nj} \quad (n = 1, 2, 3, \dots N)$$

where γ_{n0} is the intercept for group n ($n = 1, 2, 3, \dots N$), which is the (adjusted) average of outcome score for group n ; γ_{n1} ($n = 1, 2, 3, \dots N$) is the coefficients of institutional characteristics for group n ($n=1, 2, 3, \dots N$), representing institutional effects, the research focus in this study. These N institutional effects, separately for each group, can be adjusted by institutional characteristics or institution. **Institution** $_{pij}$ ($p = 1, 2, 3, \dots P$) represents these institutional characteristics as controlling variables in the institutional level. Finally, μ_{nj} ($n = 1, 2, 3, \dots N$) is the error term unique to institution j concerning group n , which is assumed to be multivariate normally distributed with a full variance-

covariance structure. The full variance-covariance structure is assumed because it is reasonable to allow group means to be correlated (across institutions). The variance and covariance structure is an n by n matrix (symmetrical along the diagonal), which is represented as

$$\begin{bmatrix} \mathbf{a}_{11} & \mathbf{a}_{12} & \cdots & \mathbf{a}_{1n} \\ \mathbf{a}_{21} & \mathbf{a}_{22} & \cdots & \mathbf{a}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_{n1} & \mathbf{a}_{n2} & \cdots & \mathbf{a}_{nn} \end{bmatrix}$$

The second set of regressions for the model at the second level “descends” from the coefficients (slopes) of individual characteristics at the first level. Each coefficient is considered as fixed at the second level.

$$\beta_{(N+m)j} = \gamma_{(N+m)0} + \mathbf{U}_{(N+m)j} \quad (m = 1, 2, 3, \dots M)$$

With the above specification of the multivariate multilevel model, the coefficients representing institutional effects (at the second level), γ_{n1} ($n = 1, 2, 3, \dots N$), are restricted to be equal, meaning that institutional effects are constrained to be the same for all N groups. The significance test examines if this restriction holds true. The null hypothesis is:

$$H_0 : \gamma_{11} = \gamma_{21} = \gamma_{31} = \cdots = \gamma_{N1}$$

If the null hypothesis cannot be rejected (because of no significance), then institutional effects are statistically the same across the N groups. If the null hypothesis is rejected, then institutional effects are significantly different across the N groups. Obviously, this is an omnibus test.

2.2 The Assumptions

All statistical models including multilevel models have assumptions that need to be checked to ensure the validity of the procedures for estimating the model (Raudenbush & Bryk, 2002). The multilevel model specified above is, by nature, a regular multilevel model. For a regular multilevel model, according to McNeish, Stapleton, and Silverman (2016), the basic assumptions speak to the independence of observation at the higher level (institution in this case) and that each institution shares the same institutional characteristics. Apart from these basic assumptions, the major assumptions are normality and homogeneity of variance. Specifically, the multilevel model assumes normal distribution of both level 1 and level 2 residuals as well as equal variances (level 2 residuals) across institutions. A large sample size may make the multilevel model robust to the violation of normality, and similar sample size across institutions may make the multilevel model robust to the violation of homogeneity of variance (Raudenbush & Bryk, 2002). The present study takes advantages of the PISA dataset which is large in size for the overall sample and similar in size across school samples, making the multilevel model robust to potential violations of multilevel assumptions.

2.3 The Estimation

A multilevel model can usually be estimated by either the full maximum likelihood (FML) or the restricted maximum likelihood (RML). Firstly, the FML estimator takes in richer information with numerical integration that includes both the regression coefficients and the variance components in the likelihood function. Compared to the FML, the RML includes only the variance components in the likelihood function. Secondly, FML is widely used and strongly preferred when the importance of predictor

variables is assessed (Hox, 2010). Lastly, in practice, the differences between the two models is usually small if the sample is relatively big (Hox, 1998; Kreft & de Leeuw, 1998). The RML is more realistic, particularly when dealing with small samples in data analysis (Bryk & Raudenbush, 1992; Longford, 1993). Since in the present study, the importance of predictor variables (whether the school reading environment variables have different impacts on sex groups) is the primary research focus and the dataset is huge, it is more appropriate to employ the FML method of estimation.

2.4 The Application

In general, school effects research is a macro-level empirical investigation that focuses on the effectiveness of educational policy and practice in promoting positive educational outcomes for students (Ma, Ma & Bradley, 2008). One of the popular theoretical essentials for school effects research is the input-process-output model (Ma et al., 2008). The present study employs this model to guide the selection of variables and the specification of models. As an application, the present study examines whether school reading environment has the same effect on reading achievement in boys as in girls, with controls over student and school characteristics.

2.4.1 Model Specification

In order to illustrate the application of the above model for the examination of institutional effects on multiple groups of individuals, a special case, which can be considered as the simplest restricted multilevel model for examination of institutional effects on multiple groups of individuals, is presented. The chosen application concerns the effects of school reading environment on student reading achievement between boys

and girls. Overall, this model has two levels with students nested within schools, and the grouping variable is sex with two categories (boys and girls).

The level 1 model is a multivariate model highlighting students' average reading achievement for each sex group:

$$Y_{ij} = \beta_{1j} * (BOY_{ij}) + \beta_{2j} * (GIRL_{ij}) + \sum_{m=1}^M \beta_{(m+2)j} * StuC_{mij} + \varepsilon_{ij}$$

where Y_{ij} is the score of the reading achievement for student i in school j ; BOY_{ij} is an indicator for boys (equal to 1 if Y_{ij} is a score for a boy and equal to 0 if Y_{ij} is a score for a girl); $GIRL_{ij}$ is an indicator for girls (equal to 1 if Y_{ij} is a score for a girl and equal to 0 if Y_{ij} is a score for a boy). β_{1j} is the average reading achievement for boys in school j and β_{2j} is the average reading achievement for girls in school j . Both β_{1j} and β_{2j} can be adjusted by student characteristics or StuC ($m = 1, 2, 3, \dots M$). Finally, ε_{ij} is the error terms which is assumed to be normal in distribution with a mean of zero and variance σ^2 .

$$\varepsilon_{ij} \sim \text{NID}(0, \sigma^2)$$

The level 2 model is two linear regression equations, one for boys and one for girls, with both boys' average reading achievement and girls' average reading achievement in school j as outcomes to be modeled by the variables representing institutional effects with control over school characteristics or SchC ($p = 1, 2, 3, \dots P$).

$$\beta_{1j} = \gamma_{10} + \gamma_{11}D_Read_j + \sum_{p=1}^P \gamma_{1(p+1)} * SchC_{pj} + U_{1j}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}D_Read_j + \sum_{p=1}^P \gamma_{2(p+1)} * SchC_{pj} + U_{2j}$$

where γ_{10} is the intercept for boys, which is the adjusted mean of reading achievement for boys; γ_{20} is the intercept for girls, which is the adjusted mean of reading achievement for girls. *D_Read* is *diversity of reading*, which represents institutional effects coming from an element of school reading environment. γ_{11} measures the institutional effects concerning *diversity of reading* for boys across schools and γ_{21} measures the institutional effects concerning *diversity of reading* for girls across schools. U_{1j} is the error term unique to school j for boys, and U_{2j} is the error term unique to school j for girls. Errors are assumed to be normally distributed with variances τ_{11} and τ_{22} and covariance τ_{12} .

$$\begin{bmatrix} \tau_{11} & \tau_{12} \\ & \tau_{22} \end{bmatrix}$$

The covariance indicates that, among schools, boys' average (adjusted) achievement scores and girls' average (adjusted) achievement scores are correlated.

With the above specification of the multivariate multilevel model, the coefficients representing school effects (at the second level), γ_{11} and γ_{21} , are restricted to be equal, meaning that the effects of school reading environment (e.g., *diversity of reading*) are constrained to be the same for both boys and girls across the schools. The significance test examines if this restriction holds true. The null hypothesis is:

$$H_0 : \gamma_{11} = \gamma_{21}$$

If the null hypothesis cannot be rejected (because of no significance), then school effects are statistically the same between boys and girls. If the null hypothesis is rejected, then school effects are statistically significantly different between boys and girls. In the latter case, a new model without restriction is specified, and the two resulting coefficients are statistically significantly different between boys and girls, indicating that school effects on reading achievement vary between boys and girls.

2.4.2 Literature review

Input-Process-Output Model of School Effects. The input-process-output model illustrates how school experiences influence students' academic outcome with respect to differing family backgrounds as well as different cognitive and affective conditions. Inputs such as family characteristics, home influences, and family social and cultural values are what students bring into their schools. Schools then process students with different inputs into different categories of educational outcomes (output), such as performance, course selection, attitude and interest. With the various input that students bring into the schools, educational outcome (output) will produce different school characteristics, which are school contexts and climates (Ma et al., 2008). Under school characteristics, school contextual and climatic variables are two classified types of characteristics. Context variables describe the "hardware" of the school, with characteristics descriptive of the material resources of a school, the student body and the teacher body. Climate variables describe the "software" of the school, which includes characteristics descriptive of the learning environment, such as how students are organized for instruction, academic students' expectations for principals and teachers, principal leadership style, decision-making processes, teacher classroom practices, and ways that a school is operated (Mullis, Martin, Foy, & Drucker, 2012; Organization for Economic Co-operation and Development (OECD, 2013). School context and climate variables have long been used to examine school effects on academic and non-academic outcomes and how they promote different learning environments for various students (Ma, 2002).

In the research literature on reading education, school reading environment is generally thought to affect student reading achievement (Lenkeit, Chan, Hopfenbeck, & Baird, 2016). For example, Costa & Araujo's (2017) study takes school characteristics into account in measuring the students' reading achievement in Danish, Swedish and French schools and offers compelling evidence of the influence of school climate variables in the development of reading ability with the database from Progress in International Reading Literacy Study (PIRLS). The importance was also recognized by Brozo and colleagues, who held that school-level programs provide students with strategies that enabled them to read with purpose and understanding while monitoring their own learning (Brozo, Shiel, & Topping, 2008; Brozo, Sulkunen, Shield, Garbe, Pandian, & Valtin, 2014).

Sex Differences in Reading Achievement. Lietz (2006) conducted a meta-analysis on 139 large-scale studies between 1970 and 2002, using a two-level hierarchical linear model (HLM) to address the questions of the extent of sex differences in reading across countries. The analysis indicated that female secondary students performed 0.19 *SDs* above males when taking age and language of instruction into account. With evidence from 31 countries, Marks (2008) concluded the average sex gap among these countries was 32 score points higher for girls in reading. Lynn & Mikk (2009) revealed that the advantages in reading achievement for 10-year-old girls was 0.23 *SDs* and 0.42 *SDs* for 15-year-old girls, with the analysis of recent international assessment PISA 2000, 2003 and 2006 and the PIRLS 2001 and 2006 dataset. In 2010, the Center on Education Policy reported that in all 50 state-level tests of reading, girls significantly outperformed boys (Chudowsky & Chudowsky, 2010). When comparing the magnitude of differences on

these assessments, Lietz (2006) concluded that the sex differences in favor of girls was more pronounced in studies conducted by the National Assessment of Educational Progress (NAEP) and Program for International Student Assessment (PISA). Solheim & Lundetra (2018) compared the impact of sex on reading literacy in PIRIL 2011 (10-year-olds), PISA 2009 (15-year-olds) and Program for the International Assessment of Adult Competencies (PIAAC) 2012 (16-24-year-olds), respectively, across the Nordic countries and found similar patterns of sex differences, with the largest effect sizes in PISA and the smallest in PIAAC. This study features much research on sex differences in reading achievement with a large dataset, however, the study is scarce in measuring whether the school effect exerts statistical differences in sex group in a multilevel setting.

As a demonstration of the application of the restricted multilevel model, this study examines the effects of school reading environment on student reading environment between boys and girls with and without controls over student and school characteristics. The research question is: Does school reading environment have the same effect on reading achievement in boys as in girls, with and without controls over student and school characteristics?

2.4.3 Data Source

With measures of students' reading achievement and individual background as well as school context and climate variables, PISA dataset is appropriate for the present study. As an international assessment, PISA measures 15-year-old students' reading, mathematics and science literacy every three years. First conducted in 2000, the major domain of study rotates between reading, mathematics and science in each cycle. All achievement measures in PISA have a standardized mean score of 500 points and a

standard deviation of 100 points (Adams & Wu, 2002). It employed a two-stage stratified random sampling procedure in each participating country or region (OECD, 2007a). In the first stage, PISA randomly selected a sample of schools from a national list of eligible schools. In the second stage, PISA randomly selected a sample of students (35 students) from sampled schools. When a school had fewer than 35 students, all students were sampled. To make the sample reflective of the population, PISA used normalized student weights and school weights. Data for the present study came from the 2009 national sample of the United States. The 2009 PISA was the latest PISA cycle that emphasized reading. The data contained 5,233 students (2,727 boys and 2,506 girls) enrolled in 165 schools.

2.4.3.1 Outcome Measure

The outcome variable was student reading achievement. PISA measured reading achievement as reading literacy, defined as the ability to extract the relevant information from texts and also to understand, use and reflect on written texts. To reduce testing time, PISA employed the matrix sampling technique (i.e., using short and different booklets of items), resulting in five plausible values for reading (OECD, 2002a). Plausible values are not test scores (in the traditional sense) and they need to be integrated together to produce a test score for each student (OECD, 2010b). After the integration, PISA scaled students' reading achievement to have a mean of 500 and a standard deviation of 100 (Adams & Wu, 2002).

2.4.3.2 Independent Variables

There were independent variables at both student and school levels in the present study. Student-level variables included student characteristics of sex, age, socioeconomic

status (SES), family structure (single-parent family vs. both-parent family), immigration status (yes or no), and home language (English vs. others). These student-level variables have long been used to explain individual differences in academic achievement (Ma et al., 2008). Specifically, *sex* contained two categories, boys and girls. *Age* was a continuous variable, measured as the number of months since birth. *SES* was a standardized index of family economic, social, and cultural status. *Family structure* was used to derive a dichotomous variable of both-parent family vs. single-parent family. *Immigration status* was used to derive a dichotomous variable of native-born student vs. immigrant student. *Home language* was used to derive a dichotomous variable of English-speaking family vs. non-English-speaking family. The only composite (index) variable at the student level was *SES*, and Appendix A presents the construction of this composite variable.

School-level variables included school contextual variables and school climate variables. As school contextual variables, *school size* was the number of enrolled students, and *school location* produced two dichotomous variables with urban schools as the baseline category against which suburban and rural schools were compared. Other school contextual variables were *proportion of girls* and *proportion of certified teachers* (measuring teacher quality). Finally, *teacher shortage* measured the adequate employment of teachers in the school, *teacher-student ratio* measured the ratio between teacher and students, and *quality of educational resources* measured school material resources, such as the conditions of buildings (as well as heating, cooling, and lighting systems), instructional space, instructional resources (computers, instructional materials

in the library, multi-media resources, science laboratory equipment, facilities for the fine arts.

Characteristics of the school reading environment were the key school climate variables. PISA created a set of five scale or index variables to measure various reading behaviors of students. These variables were aggregated across students within each school to form five measures (indicators) of school reading environment. Specifically, *diversity of reading* measures the extent to which students read six types of text including magazines, comics, fiction books, nonfiction books, e-mail and webpages. Students were asked to rate their level of diversity of reading by answering the question “How often do you read these materials because you want to?” They were asked to use a five-point scale, with 1 indicating “Never or almost” and 5 indicating “Several times a week.” *Enjoyment of reading* refers to reading for pleasure. Students were asked to rate their level of enjoyment of reading by responding to 11 questions measuring, “How much do you agree or disagree with these statements about reading.” Students used a four-point scale, with 1 indicating “Strongly disagree” and 4 meaning “Strongly agree.” They could choose statements such as, “I read only if I have to,” “Reading is one of my favorite hobbies,” and so on. *Stimulators of reading* measures the extent to which teachers stimulate students for reading engagement and work with students on reading skills (e.g., the teacher helps students relate the stories they read to their lives and encourages students to express their opinions about a text). Students were asked to rate their level of stimulators of reading through the question, “In your lesson, how often does the following occur?” Students used a four-point scale, with 1 indicating “Never or hardly ever” and 4 indicating “in all lessons,” and statements such as “The teacher asks

students to explain the meaning of a text.” *Daily reading hours* is a sum of reading activities in which students engage each day. Students were asked to rate their level of daily reading hours through the question, “How much time do you spend reading for enjoyment?” using a five-point scale with 1 indicating “zero hours” and 5 indicating “more than two hours.” *Online reading hours* refers to the process of extracting meaning from a text that is in a digital format. These variables were coded in such a way that a higher value indicated a more positive school reading environment. Students were asked to rate their level of online reading hours by responding to the question “How often are you involved in the following reading activities?” (reading emails or chatting online, for example), and they were asked to use a five-point scale where 1 indicated “I do not know what it is” and 5 indicated “Several times a day.”

There were other school climate variables used as adjustments for school reading environment. *Teacher participation* was a composite variable, measuring the extent to which student learning is supported by teachers’ responsibility for decisions regarding the management of the school (e.g., admitting students to the school and determining course content). *Teacher behavior* was a composite variable, measuring the extent to which student learning is hindered by some behaviors of teachers in relation to their students, such as holding low expectations for students and having a poor relationship with students. *Student behavior* was a composite variable, measuring the extent to which student learning is hindered by some disruptive behaviors in school (e.g., student absenteeism, disruption of classes by students, and student use of alcohol or illegal drugs). *School leadership* was a composite variable, measuring the extent to which student learning is supported by the making or altering school policy (e.g., activities and

behaviors of the principal, principal observation of classroom instruction). The composite variables at the school level included *diversity of reading*, *enjoyment of reading*, *stimulator of reading*, *daily reading hours* and *online reading hours*. *Ability grouping* was a dichotomous variable, denoting whether a school groups students according to ability for instruction. Appendix A presents the construction of these composite variables. Appendix B contains descriptions of all student-level variables and school-level variables. Appendix C and D present descriptive statistics and Spearman's correlation on all student-level variables and school-level variables (to check multicollinearity).

2.4.4 Analytical Procedure

A two-step procedure was carried out. In the first step, the effects of the five schools' reading environment variables (*diversity of reading*, *enjoyment of reading*, *stimulators of reading*, *daily reading hours*, and *online reading hours*) were examined individually, without the control for student characteristics and school characteristics, by means of the above restricted multilevel model. In the second step, student characteristics and school characteristics were added to adjust for the effects of the school reading environment variables. Whether these school reading environment variables were statistically different between boys and girls would be tested.

When the school reading environment variables showed statistically the same effects between boys and girls in the restricted multilevel model, the restricted multilevel model would become the final model. When the school reading environment variables showed statistically different effects between boys and girls in the restricted model, the non-restricted model, which was the conventional model, would be used to derive the

final model and show the extent of differences in the effects of the school reading environment variables between boys and girls.

The alpha level for all statistical tests was set as .05. The HLM7.03 software provided the analytical platform for the present study. The full maximum likelihood estimation procedure was applied for all multilevel analyses. A full variance-covariance structure was estimated for each multilevel model.

2.4.5 Results

A two-step procedure has been carried out. In the first step, the effects of the five school reading environment variables (diversity of reading, enjoyment of reading, stimulators of reading, daily reading hours, and online reading hours) were examined individually through the absolute model. The goal of the absolute model was to test whether the effects of these school reading environment variables were statistically different between boys and girls in the absence of student and school characteristics and, if yes, by how much (Table 2.1); in the second step, the relative model was produced by adding student and school characteristics. The goal of the relative model was to examine whether the effects of school reading environment variables across the boys and girls would change in the presence of student and school characteristics (Tables 2.2 and 2.3). With the application of HLM7.03 software, an effect was considered statistically significant if the p value was below 0.05 at the school level throughout the analysis.

2.4.5.1 Absolute Model of School Reading Environment on Reading Achievement of Boys and Girls, without Control for Student and School Characteristics.

For the absolute model (Table 2.1), the two school reading environment variables *diversity of reading* and *online reading hours* have been shown to exert statistically

different effects on reading achievement between boys and girls. The effects of *diversity of reading* on reading achievement for boys ($\beta = -19.77$, SE = 18.51) was statistically significantly different from the effects of *diversity of reading* on reading achievement for girls ($\beta = -29.47$, SE = 13.08). A one-unit increase (out of a measurement scale of 1 to 5) in diversity of reading collectively in a school was associated with a decrease in student individual reading achievement of 19.77 for boys and 29.47 for girls.

Meanwhile, the effects of *online reading hours* on reading achievement for boys ($\beta = -24.79$, SE = 16.55) was statistically different from the effects of *online reading hours* on reading achievement for girls ($\beta = -27.73$, SE = 9.66). A one-unit increase (out of a measurement scale of 1 to 5) in online reading hours collectively in a school was associated with a decrease in student individual reading achievement of 24.79 for boys and 27.73 for girls.

The other three school reading environment variables—*enjoyment of reading*, *stimulators of reading* and *daily reading hour*—did not show any statistically different effects on students' reading achievement between boys and girls. There was no difference in the effects of *enjoyment of reading* on reading achievement between boys and girls ($\beta = 86.13$, SE = 47.04, $p > .05$); no difference in the effects of *stimulators of reading* on reading achievement between boys and girls ($\beta = 37.88$, SE = 21.10, $p > .05$); and no difference the effects of *daily reading hour* on reading achievement between boys and girls ($\beta = -18.68$, SE = 32.11, $p > .05$).

2.4.5.2 Relative Model of School Reading Environment on Reading Achievement of Boys and Girls, with Control for Student Characteristics.

Two school reading environment variables, *enjoyment of reading* and *online reading hours*, have indicated statistically different effects on reading achievement between boys and girls after student background variables were added into the model to adjust the effects.

The effect of *enjoyment of reading* on reading achievement for boys ($\beta = 63.37$, $SE = 38.06$) was statistically different from the effects of *enjoyment of reading* on reading achievement for girls ($\beta = 77.01$, $SE = 30.12$), controlling for student characteristics. A one-unit increase (out of a measurement scale of 1 to 5) in enjoyment of reading collectively in a school was associated with an increase in student individual reading achievement of 63.37 for boys and 77.01 for girls, when student characteristics were controlled.

The effects of *online reading hours* on reading achievement for boys ($\beta = -20.00$, $SE = 15.83$) was statistically significantly different from the effects of *online reading hours* on reading achievement for girls ($\beta = -23.64$, $SE = 8.85$), controlling for student characteristics. A one-unit increase (out of a measurement scale of 1 to 5) in diversity of reading collectively in a school was associated with a decrease in student individual reading achievement of 20.00 for boys and 23.64 for girls, when student characteristics were controlled.

The other three school reading environment variables did not show any statistically different effects on students' reading achievement between boys and girls. There was no difference in the effects of *diversity of reading* on reading achievement between boys and girls ($\beta = -22.71$, $SE = 12.32$, $p > .05$); no difference for the effects of *stimulator of reading* on reading achievement between boys and girls ($\beta = 26.02$, $SE =$

17.95, $p > .05$) and no difference for the effects of *daily reading hour* on reading achievement between boys and girls ($\beta = -6.14$, $SE = 24.89$, $p > .05$).

Overall, after the control of student characteristics, the different absolute effects of diversity of reading between boys and girls disappeared, but the different absolute effects of online reading hours between boys and girls remained. Meanwhile, after the control of student characteristics, the different effects of enjoyment of reading between boys and girls appeared.

2.4.5.3 Relative Model of School Reading Environment on Reading Achievement of Boys and Girls, with Control for Student and School Characteristics. Two school reading environment variables, *enjoyment of reading* and *online reading hours*, consistently showed statistically different effects on reading achievement between boys and girls, after student characteristics and school characteristics variables were added in the model to adjust for the effects.

The effect of *enjoyment of reading* on reading achievement for boys ($\beta = 49.19$, $SE = 23.00$) was statistically different from the effects of *enjoyment of reading* on reading achievement for girls ($\beta = 58.77$, $SE = 16.37$), with student background and school characteristics controlled in the model. A one-unit increase (out of a measurement scale of 1 to 5) in enjoyment of reading collectively in a school was associated with an increase in student individual reading achievement of 49.19 for boys and 58.77 for girls, when student background and school characteristics were controlled.

The effects of *online reading hours* on reading achievement for boys ($\beta = -18.33$, $SE = 16.99$) was statistically significantly different from the effects of *online reading hours* on reading achievement for girls ($\beta = -21.69$, $SE = 8.38$), controlling for student

and school characteristics. A one-unit increase (out of a measurement scale of 1 to 5) in *online reading hours* collectively in a school was associated with a decrease in student individual reading achievement of 18.33 for boys and 21.69 for girls, controlling for student and school characteristics.

The other three school reading environment variables did not show any statistically different effects on students' reading achievement between boys and girls. After controlling for student and school characteristics, there was no difference for the effects of *diversity of reading* on reading achievement between boys and girls ($\beta = -20.47$, $SE = 10.58$, $p > .05$); no difference for the effects of *stimulator of reading* on reading achievement between boys and girls ($\beta = 11.85$, $SE = 11.29$, $p > .05$), and no difference for the effects of *daily reading hour* on reading achievement between boys and girls ($\beta = -16.23$, $SE = 12.04$, $p > .05$).

2.4.5.4 Variance Components and Proportion of Variance. Although the variance components did not directly help address the research questions, their estimations were used to calculate the proportion of variance accounted for by models involving statistically significant school environment variables, *enjoyment of reading* and *online reading* (Table 2.4 and Table 2.5). For *enjoyment of reading*, 71% of the variance in boys' reading achievement has been accounted for by the overall model, while 80% of the variance in girls' reading achievement has been accounted for by the overall model. For variable *online reading hours*, 72% of the variance in boys' reading achievement has been accounted for by the overall model, and 81% of the variance in girls' reading achievement has been accounted for by the overall model. The explained proportions for both boys and girls indicated that these two school reading environment variables *online*

reading hours and *enjoyment of reading* each played an important role (i.e., explained substantial amount of variation) in its specific overall model.

It appears that enjoyment of reading associates positively with reading achievement, and this association is stronger for girls than boys based on the analysis of PISA 2009 reading achievement dataset. This finding implies that helping students, particularly girls, enjoy reading is an effective educational strategy to improve reading achievement. For boys to improve their reading achievement, the promotion of enjoyment of reading would not be sufficient if the educational goal is to have them achievement as much as girls. Other educational interventions need to be considered. Meanwhile, it appears that online reading actually harms reading achievement with a negative association for both boys and girls, but the negative effects are stronger on girls compared to boys. Because online reading can be irrelevant to schoolwork, educators and parents are suggested to monitor the content that students, particularly girls, spend online for reading. For girls to overcome stronger negative effects, it may also be necessary to limit online hours that they spend.

2.5 Final Remarks on the Restricted Multilevel Model

As a family of advanced multilevel model techniques, the restricted multilevel model has a list of advantages over the traditional multilevel ones. It is an effective omnibus statistical technique to examine the institutional effects on multiple groups of individuals. It has a broad applicability and is a convenient tool to see the impact of higher-level institutional effects on lower-level groups of individuals, such as the effects of school reading environment variables on sex groups demonstrated as an example in the current study.

2.5.1 Model Performance

As we demonstrated in the study, the restricted multilevel model was convenient to perform and execute. First, the data is easy to prepare in the model. Instead of the common dummy coding, which resulted in $N - 1$ dichotomous variables leaving out a reference category, the restricted multilevel model created N dichotomous variables to present each group. This representation of categorical variables makes much more sense to readers who do not have substantial statistical background. All the other controlling variables in the first or second level were prepared in the same way as the traditional multilevel model, with no differences in how researchers would prepare for multiple regression analysis. This familiarity allows them to set up their database for analysis quickly and easily.

Second, the model is easy to specify. It is a relatively straightforward way to set up the equation in the model. Any researcher with basic knowledge and skills on multiple regression analysis can specify the model effortlessly. Third, the modeling result was easy to show and interpret. Just like the PISA example employed in the study, the final result would be interpreted as to whether the school-level reading environment variables had the same or different effects on the two sex groups. Lastly, the restricted multilevel model can be run in different analytical platforms, including the HLM employed in the current study, as well as MLwin, Mplus or R.

2.5.2 Model Extension

First, besides the simplified version of the two-group comparison the model demonstrated, it can extend the comparisons from two to multiple groups. Multiple pairs of coefficients can be constrained in the model when more groups are involved, creating

an ANOVA-like data analysis. This situation will allow many researchers to easily work with the complex model because of their familiarity with ANOVA. Second, not only can the groups themselves be constrained in the model, the interactions between the level-2 variables can also be constrained so that researchers can examine whether groups share the same interactive pattern regarding the outcome measure. For instance, using sex groups as an example, L2A and L2B are level-2 variables, and L2A*L2B are their interaction. This interaction can be constrained for male and female groups. In this way, researchers can see if the two sex groups share the same interaction pattern in regard to their outcome measure. Such an extension opens doors to many research possibilities that would be very difficult to imagine with traditional statistical approaches.

2.5.3 Model Limitation

However, even though the model can specify multiple groups for categorical variables with categories more than two, the model does not directly generate post-hoc analyses for researchers who would like to rank order categories based on outcome measures. In other words, the model can generate an effect for each group and can perform an omnibus test on whether these effects are all the same. However, when the omnibus test is statistically significant, the model cannot perform subsequent post-hoc analysis. Note that this limitation concerns the software, not necessarily the model. A more precise statement is that the current software packages cannot perform post-hoc analysis for restricted multilevel models with multiple constrained groups. Researchers need to write program codes (e.g., in R) to extend the analytical function of the restricted multilevel model presented in this study.

Table 2.1

Estimates of Absolute Effects of School Reading Environment on Reading Achievement of Boys and Girls

	Effect	SE
Diversity of Reading		
Boys	-19.77	18.51
Girls	-29.47	13.08
Enjoyment of Reading		
Boys	86.13	47.04
Girls	86.13	47.04
Stimulators of Reading		
Boys	37.88	21.10
Girls	37.88	21.10
Daily Reading Hours		
Boys	-18.68	32.11
Girls	-18.68	32.11
Online Reading Hours		
Boys	-24.79	16.55
Girls	-27.73	9.66

Note. Statistically significantly different effects between boys and girls are bold ($p < 0.05$).

Table 2.2

Estimates of Relative Effects of School Reading Environment on Reading Achievement of Boys and Girls, Controlling for Student Characteristics

	Effect	SE
Diversity of Reading		
Boys	-22.71	12.32
Girls	-22.71	12.32
Enjoyment of Reading		
Boys	63.37	38.06
Girls	77.01	30.12
Stimulators of Reading		
Boys	26.02	17.95
Girls	26.02	17.95
Daily Reading Hours		
Boys	-6.14	24.89
Girls	-6.14	24.89
Online Reading Hours		
Boys	-20.00	15.83
Girls	-23.64	8.85

Note. Statistically significantly different effects between boys and girls are bold ($p < 0.05$). Student characteristics include age, socioeconomic status, family structure (single-parent family vs. both-parent family), immigration status (yes vs. no), and home language (English vs. others).

Table 2.3

Estimates of Relative Effects of School Reading Environment on Reading Achievement of Boys and Girls, Controlling for Student Characteristics and School Characteristics

	Effect	SE
Diversity of Reading		
Boys	-20.47	10.58
Girls	-20.47	10.58
Enjoyment of Reading		
Boys	49.19	23.00
Girls	58.77	16.37
Stimulators of Reading		
Boys	11.85	11.29
Girls	11.85	11.29
Daily Reading Hour		
Boys	-16.23	12.04
Girls	-16.23	12.04
Online Reading Hours		
Boys	-18.33	16.99
Girls	-21.69	8.38

Note. Statistically significantly different effects between boys and girls are bold ($p < 0.05$). Student characteristics include age, socioeconomic status, family structure (single-parent family vs. both-parent family), immigration status (yes vs. no), and home language (English vs. others). School characteristic includes school size, school location (suburban and rural vs. urban), proportion of girls, proportion of certified teachers, teacher-student ratio, teacher shortage, quality of educational resources, teacher participation, teacher behavior, student behavior, school leadership, and ability grouping (yes vs. no).

Table 2.4

Estimates of Variance Components and Proportion of Variance Explained for Enjoyment of Reading

Variance Components	M0	M1	M2
---------------------	----	----	----

Boys	2728.90	1444.24	796.22
Girls	2218.44	998.86	440.63
Proportion of Variance Explained			
Boys			0.71
Girls			0.80

Note. M0 = absolute model (without student and school characteristics). M1 = relative model with student characteristics. M2 = relative model with student and school characteristics. Student characteristics include age, socioeconomic status, family structure (single-parent family vs. both-parent family), immigration status (yes vs. no), and home language (English vs. others). School characteristic includes school size, school location (suburban and rural vs. urban), proportion of girls, proportion of certified teachers, teacher-student ratio, teacher shortage, quality of educational resources, teacher participation, teacher behavior, student behavior, school leadership, and ability grouping (yes vs. no).

Table 2.5

Estimates of Variance Components and Proportion of Variance Explained for Online Reading Hours

Variance Components	M0	M1	M2
Boys	3261.28	1896.51	916.64
Girls	3015.81	1614.09	587.62
Proportion of Variance Explained			
Boys			0.72
Girls			0.81

Note. M0 = absolute model (without student and school characteristics). M1 = relative model with student characteristics. M2 = relative model with student and school characteristics. Student characteristics include age, socioeconomic status, family structure (single-parent family vs. both-parent family), immigration status (yes vs. no), and home language (English vs. others). School characteristic includes school size, school location (suburban and rural vs. urban), proportion of girls, proportion of certified teachers, teacher-student ratio, teacher shortage, quality of educational resources, teacher

participation, teacher behavior, student behavior, school leadership, and ability grouping
(yes vs. no).

Copyright © Rongxiu Wu 2020

CHAPTER 3: A Multilevel Model with Heterogeneous Sigma Squared Function to Compare Distributional Properties of Multiple Groups

3.1 The Model

A good understanding of the distributional properties across groups is an essential part of making group comparisons. The most popular way to make group comparisons is by considering means across the groups. Although this approach focusing on the central tendency is important, the other critical element in describing the distributions across groups has been generally ignored. That critical element is the variability. Overall, the combination of central tendency and variability is the preferred way to describe (and compare) distributions across groups (Bayes & Monseur, 2016; Halpern et al., 2007; Ma, 2008). The present study aims to fill in this gap in the quantitative research literature. Specifically, the goal is to propose an advanced multilevel model with an embedded analytic function, referred to as heterogeneous sigma squared, that can perform statistical tests of significance to compare variances across multiple groups. As a result, this multilevel model is able to examine the distributional properties, including central tendency and variability, simultaneously. The term “simultaneously” is worth emphasizing. It implies a multivariate treatment of central tendency and variability. In contrast, the separate testing of central tendency and variability constitutes a univariate approach. Obviously, the multilevel model with heterogeneous sigma squared function provides a more efficient and effective way to describe and compare distributional properties across multiple groups.

In most cases of application, the multilevel model has two levels, with individuals nested within institutions. The categorical variable of interest is at the level 1. The

variable has N groups (categories) (i.e., $G_1, G_2, G_3, \dots, G_N$), and the goal is to compare the distributional properties across the groups. As usual, $N - 1$ dummy variables are created through dummy coding to represent the categorical variables. The level 1 model can also incorporate other independent variables, often as control variables, to adjust for the differences among groups. The level one model is expressed as:

$$Y_{ij} = \beta_{0j} + \beta_{1j}G1_{ij} + \beta_{2j}G2_{ij} + \beta_{3j}G3_{ij} + \dots + \beta_{(N-1)j}G(N-1)_{ij} + \sum_{p=1}^P \beta_{(N-p+1)} \mathbf{individual}_{p_{ij}} + \varepsilon_{ij}$$

where Y_{ij} is the outcome for individual i in institution j ; β_{0j} is the average outcome for school j after adjusting for individual characteristics and group differences; β_{nj} ($n = 1, 2, 3, \dots, N - 1$) is the within-institution group difference in outcome for institution j ; ε_{ij} is the error term at the individual level and assumed to be normally distributed with a mean of zero and a variance component.

$$\varepsilon_{ij} \sim \text{NID}(0, \sigma^2)$$

The level 2 model is:

$$\beta_{0j} = \gamma_{00} + U_{0j}$$

$$\beta_{1j} = \gamma_{10} + U_{1j}$$

$$\beta_{2j} = \gamma_{20} + U_{2j}$$

$$\beta_{3j} = \gamma_{30} + U_{3j}$$

...

$$\beta_{(N-1)j} = \gamma_{(N-1)0} + U_{(N-1)j}$$

$$\beta_{(N-p+1)j} = \gamma_{(N-p+1)0} \quad (p = 1, 2, 3, \dots, P)$$

where $\boldsymbol{\gamma}_{00}$ is the grand outcome mean and $\boldsymbol{\gamma}_{10}$ to $\boldsymbol{\gamma}_{(N-1)0}$ are the averages within-institution slope (e.g., a gap of some kind). \boldsymbol{U}_{0j} to $\boldsymbol{U}_{(N-1)j}$ are error terms at the institution level unique to each institution, assumed to be multivariate normally distributed with a full variance-covariance structure. The full variance-covariance structure is assumed because it is reasonable to allow groups to be correlated (across institutions). The variance and covariance structure is an n by n (symmetrical) matrix, which is represented as

$$\begin{bmatrix} \mathbf{a}_{11} & \mathbf{a}_{12} & \cdots & \mathbf{a}_{1n} \\ \mathbf{a}_{21} & \mathbf{a}_{22} & \cdots & \mathbf{a}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_{n1} & \mathbf{a}_{n2} & \cdots & \mathbf{a}_{nn} \end{bmatrix}$$

On the basis of this two-level model, variances across the groups can be assumed to be different, and a structural specification on the variance can be added. The add-on equation is:

$$\sigma^2 = \exp \{ \alpha_0 + \alpha_1 \mathbf{G1} + \alpha_2 \mathbf{G2} + \alpha_3 \mathbf{G3} + \cdots + \alpha_{(N-1)} \mathbf{G(N-1)} \}$$

where α_0 is the intercept in estimating the log form of σ^2 and α_n ($n = 1, 2, 3, \dots, N-1$) are the slopes of groups (categories) in estimating the log form of σ^2 . The α coefficients are estimated through full maximum likelihood and are tested for statistical significance by means of z statistic under large-sample theory. The null hypothesis is that the population variances from which groups are drawn are equal (homogeneity of variance).

The multilevel model with heterogeneous sigma squared can be compared easily with the multilevel model with homogeneous sigma squared utilizing a likelihood-ratio test. For each model, a deviance statistic is computed. The higher the deviance, the poorer the fit (Raudenbush & Bryk, 2002). The difference between the deviance statistics (from the two models) is then used to test the hypothesis. In sum, the performance of the

multilevel model with heterogeneous sigma squared is evaluated in comparison with the performance of the multilevel model with homogeneous sigma squared.

3.2 The Assumptions

All statistical models including multilevel models have assumptions that need to be met to ensure the validity of the procedures for estimating the model (Raudenbush & Bryk, 2002). The multilevel model specified above is, by nature, a regular multilevel model. For a regular multilevel model, according to McNeish, Stapleton, and Silverman (2016), the basic assumptions speak to the independence of observation at the higher level (institution in this case) and that each institution shares the same institutional characteristics. Apart from these basic assumptions, the major assumptions are normality and homogeneity of variance. Specifically, the multilevel model assumes normal distribution of both level 1 and level 2 residuals as well as equal variance (level 2 residuals) across institutions. Large sample size may make the multilevel model robust to the violation of normality, and similar sample size across institutions may make the multilevel model robust to the violation of homogeneity of variance (Raudenbush & Bryk, 2002). The present study takes advantages of the PISA data, which are large in size for the overall sample and similar in size across school samples, making the multilevel model robust to potential violations of multilevel assumptions.

There are additional assumptions that often draw less attention from the analysts but may need to be shown in the present study. The errors at the higher level are assumed to be independent from the errors at the lower level, that is $\text{Cov}(\boldsymbol{\varepsilon}_{ij}, \mu_j) = 0$. Specific to the multilevel model in the present study, the predicted categorical variables (\mathbf{Gn}) do not covary with the residuals at any other level, which is $\text{Cov}(\mathbf{Gn}, \boldsymbol{\varepsilon}_{ij}) = 0$, $\text{Cov}(\mathbf{Gn}, \mathbf{u}_j) =$

0. With the add-on specification, heterogeneous level-1 variance is hypothesized across categories (groups) and is modeled by the predictor variables of ***Gn***.

3.3 The Estimation

A multilevel model usually can be estimated by either the full maximum likelihood (FML) or the restricted maximum likelihood (RML). Firstly, the FML estimator takes in richer information with numerical integration that includes both the regression coefficients and the variance components in the likelihood function. Compared to the FML, the RML includes only the variance components in the likelihood function. Secondly, FML is widely used and strongly preferred when the importance of predictor variables is assessed (Hox, 2010). Lastly, in practice, the differences between the two models are usually small if the sample is relatively big (Hox, 1998; Kreft & de Leeuw, 1998). The RML is more realistic, particularly when dealing with small samples in data analysis (Bryk & Raudenbush, 1992; Longford, 1993). Since, in the present study, the importance of predictor variables (whether the sex groups have different distributions) is the primary research focus and the dataset is huge, the FML is more appropriate to be employed.

3.4 The Application

3.4.1 Model Specification

To illustrate the application of the multilevel model with heterogeneous sigma squared, the distributional characteristics across two groups are examined in relation to sex differences in reading achievement. This multilevel model has students nested within schools. The level 1 model has SEX as the categorical independent variable (dummy

coded as boys = 0 and girls = 1). Student background variables are added as control variables.

$$Y_{ij} = \beta_{0j} + \beta_{1j}SEX_{ij} + \sum_{p=1}^P \beta_{(p+1)j}X_{pij} + \varepsilon_{ij}$$

where Y_{ij} is the score of the reading achievement for student i in school j ; β_{0j} is the average reading achievement for school j after adjusting for sex differences and other student-level variables; and β_{1j} is the within-school sex gap in reading achievement for school j . β_{pj} is the slope for student-level variable X_{pij} ($p = 1, 2, 3, \dots, P$) measuring the effects of each student-level variable on reading achievement. ε_{ij} is the error term at the student level and assumed to be normally distributed with a common variance.

The level 2 model has two random components, and they are modeled by school background variables. The equations are:

$$\beta_{0j} = \gamma_{00} + \sum_{q=1}^Q \gamma_{0q}W_{qj} + U_{0j}$$

$$\beta_{1j} = \gamma_{10} + \sum_{q=1}^Q \gamma_{1q}W_{qj} + U_{1j}$$

where γ_{00} is the adjusted grand mean of reading achievement; γ_{0q} is the slope for school-level variable W_{qj} ($q = 1, 2, 3, \dots, Q$) measuring the effects of each school-level variable on the school average reading achievement; and U_{0j} is the error term corresponding to the intercept at the school level unique to each school. Meanwhile, γ_{10} is the average within-school sex gap; and γ_{1q} is the slope for school-level variable W_{qj} ($q = 1, 2, 3, \dots, Q$) measuring the effects of each school-level variable on the within-school sex gap in

reading achievement. U_{1j} is the error term corresponding to the slope at the school level unique to each school.

Finally, the variance specification $\sigma^2 = \exp \{ \alpha_0 + \alpha_1 SEX \}$ is added to the multilevel model so that the heterogenous-sigma-squared procedure can be performed to compare boys and girls in terms of variance in reading achievement.

3.4.2 Literature Review

Sex Differences in Mean of Reading Achievement. Sex differences in reading achievement have been commonly studied in means by employing different large datasets through multiple research methods. Generally, pronounced sex differences in mean of reading achievement in favor of girls were found in all participating countries in the PISA surveys carried out in 2000 and 2009 (Langen, Boskers, & Dekkers, 2006; Liu & Wilson, 2009; OECD, 2001, 2010b) and averaging difference was more than 0.3 *SDs* (OECD, 2009). A meta-analysis on 139 large-scale studies between 1970 and 2002 that applied a two-level hierarchical linear modeling (HLM) indicated that female secondary students performed 0.19 *SDs* above males when taking age and language of instruction into account (Lietz, 2006). Lynn & Mikk (2009) revealed that the advantages in reading achievement for 10-year-old girls was 0.23 *SDs* and that for 15-year-old girls, it was 0.42 *SDs*, with the analysis of recent international assessment PISA 2000, 2003 and 2006 and the PIRLS 2001 and 2006 dataset. Compared with the raw scores, Marks (2008) concluded the average sex gap among these countries was 32 score points higher for girls in reading, based on evidence from 31 countries. Additionally, sex differences generally increased over PISA cycles, with the average sex difference across OECD countries increasing from 20 points in 2000 to 39 points in 2009 (Brozo et al., 2014). Additionally,

slight differences in effect size could be found in various large datasets. Solheim & Lundetra (2018) compared the impact of sex on reading literacy in PIRIL 2011 (10-year-olds), PISA 2009 (15-year-olds) and Program for the International Assessment of Adult Competencies (PIAAC) 2012 (16 to 24-year-olds) respectively across the Nordic countries and noticed similar patterns of sex differences, with the largest effect sizes in PISA and the smallest in PIAAC. However, in general, the findings regarding sex differences were remarkably similar and complementary in most large-scale assessment programs, which found that girls perform relatively higher in reading outcomes than boys.

Sex differences in Variance in Reading Achievement. Sex comparison could not be directly considered as one sex being superior to the other or equivalent to the other based only upon the mean differences (Lafontaine & Monseur, 2009; Schwabe, McElvany & Trendtel, 2015). Monseur (2016) objected to utilizing central tendency statistics only, saying it was misleading for sex comparisons. It could lead to an overly optimistic evaluation of the actual sex differences in reading achievement the study concluded. However, most researchers still viewed the whole picture from a “mean” perspective (based on gender equality on average).

Multiple empirical benchmarks have been encouraged to interpret the data in a more comprehensive way and with more insights than just the mean estimate (Hill, Bloom, Black, & Lipsey, 2008). Looking at the extreme tails and the variability helps to nuance the outcomes on sex differences, and it is more substantial than the sex differences at the mean. Comparison of groups at the extreme tails of the distribution

could be quite different from what is observed with central tendency indices (Bays & Monseur, 2016).

Though not a core concern in research and development, within-sex variability was noted more than a century ago in research. In the area of mathematics achievement research, Ellie (1894) put forward the “greater male variability hypothesis,” whereby, on one hand, male students possessed greater average math achievement than female students, but on the other hand, male students dominated both the top and bottom of the distribution while female students occupied the middle in mathematics achievement distribution. This hypothesis has been confirmed in other studies in mathematics research (e.g., Feingold, 1992; Beller & Gafni, 1996; Halpern et al., 2007). Baye & Monseur (2016) indicated that males’ scores vary more compared to females’ scores, and the difference was larger between males and females at the lower end of the distribution. They also indicated that the variability of mathematics achievement between male and female students depended at least upon age and education system.

Compared to the research on mathematics achievement by sex, the distribution and the variability of reading achievement has not been extensively researched in literacy education. It is meaningful and valuable to know the comprehensive distribution and the variability in reading achievement by sex due to the fundamental role of reading ability for students.

Factors That Affect Sex Differences. Schools are the key institutions in the lives of students and critical for overcoming sex differences (Ma, 2008; Marks, 2008; Walkerdine, 1988). The “added-value” of the schools to the academic achievement of students cannot be overlooked (Everson & Millsap, 2004; Lee, Zuze & Ross, 2005;

Opendakker & Van Damme, 2006; Willms, 1992). Under school characteristics, school contextual and climatic variables are two classified types of characteristics. Context variables describe the “hardware” of the school, with characteristics descriptive of the material resources of a school, the student body and the teacher body, and climate variables, while “software” of the school includes characteristics descriptive of the learning environment, such as how students are organized for instruction, academic students’ expectations for principals and teachers, principal leadership style, decision-making processes, teacher classroom practices, and ways that a school is operated (Ma, Ma, & Bradley, 2008; Mullis, Martin, Foy, & Drucker, 2012; OECD, 2013). School context and climate variables have long been used to examine school effects on academic and non-academic outcomes and how they promote different learning environments for various students (Ma, 2002).

School climate is usually the main research focus since it is under the direct control of parents, teachers and administrators, and it could provide more guided direction for administrators to create, amend or reform school policies and practices to provide teachers and students with a positive environment. It is imperative that studies of school effects examine how schools can use climatic characteristics to influence students’ academic performance. The disciplinary climate, academic pressure, and parental involvement, which were traditionally considered as primary measures of school climates, affect educational outcomes of students (Ma & Klinger, 2000; Ma et al., 2008; Ma & Willms, 2004; Willms, 1992).

3.4.3 Research Questions

For the application of this advanced multilevel model, this study aimed to see what distributional properties exist between boys and girls in reading achievement with control over student characteristics and school characteristics. Specifically,

1. Does one sex have a higher average reading achievement?
2. Does one sex have a large variance in reading achievement?
3. What are the unique distributional properties concerning reading achievement for each sex? For example, does one sex tend to occupy both top and bottom in the distribution of reading achievement while the other sex is sandwiched in between?

3.4.4 Data Sources

As an international assessment that measures 15-year-old students' reading, mathematics and science literacy every three years, the PISA 2009 national sample of the United States data was employed for the present study, which was the latest PISA cycle that emphasized reading. The data contained 5,121 students (2,630 boys and 2,491 girls) enrolled in 165 schools. PISA employed a two-stage stratified random sampling procedure in each participating country or region (OECD, 2007a). In the first stage, PISA randomly selected a sample of schools from a national list of eligible schools. In the second stage, PISA randomly selected a sample of students (35) from sampled schools. When a school had fewer than 35 students, all students were sampled. All achievement measures in PISA have a standardized mean score of 500 points and a standard deviation of 100 points (Adams & Wu, 2002). To make the sample reflective of the population, PISA used normalized student weights.

3.4.4.1 Outcome Measure

The outcome variable was student reading achievement, which was defined as the ability to extract the relevant information from texts and also to understand, use and reflect on written texts in PISA. To reduce testing time, PISA employed the matrix sampling technique (i.e., using short and different booklets of items), resulting in five plausible values for reading (OECD, 2002a). Plausible values are not test scores (in the traditional sense), and they integrate together to produce a test score for each student (OECD, 2009). The outcome score has a mean of 500 and a standard deviation of 100.

Regular multilevel models can directly take in plausible values for data analysis. Nonetheless, due to the software limitation, plausible values and heterogeneous sigma squared function cannot be specified at the same time.

3.4.4.2 Independent Variables

There were independent variables at both student and school levels in the present study. Student-level variables included student characteristics of age, socioeconomic status, family structure (single-parent family vs. both-parent family), immigration status (yes vs. no), and home language (English vs. others). These student-level variables have long been used to explain individual differences in academic achievement (see Ma et al., 2008). Specifically, *sex* contained two categories, boys and girls. *Age* was a continuous variable, measured as the number of months since birth. *Socioeconomic status* was a standardized index of family economic, social, and cultural status. *Family structure* was used to derive a dichotomous variable of both-parent family versus single-parent family. *Immigration status* was used to derive a dichotomous variable of native-born student versus immigrant student. *Home language* was used to derive a dichotomous variable of

English-speaking family versus non-English-speaking family. The only composite (index) variable at the student level was *socioeconomic status*, and Appendix A presents the construction of this composite variable.

School-level variables included school contextual variables and school climate variables. As school contextual variables, *school size* was the number of enrolled students, and *school location* produced two dichotomous variables with urban schools as the baseline category against which suburban and rural schools were compared. Other school contextual variables were *proportion of girls* and *proportion of certified teachers* (measuring teacher quality). Finally, *teacher shortage* measured the teacher-student ratio within a school, and *quality of educational resources* measured school material resources such as the conditions of buildings (as well as heating, cooling, and lighting systems), instructional space, instructional resources (computers, instructional materials in the library, multi-media resources, science laboratory equipment, facilities for the fine arts.)

There were other school climate variables which were used as adjustments for school reading environment. *Teacher participation* was a composite variable, measuring the extent to which the learning of students is supported by teachers' responsibility for decisions regarding the management of the school (e.g., admitting students to the school and determining course content). *Teacher behavior* was a composite variable, measuring the extent to which the learning of students is hindered by some behaviors of teachers in relation to their students, such as holding low expectations for students and having a poor relationship with students. *Student behavior* was a composite variable, measuring the extent to which student learning is hindered by some disruptive behaviors in school (e.g., student absenteeism, disruption of classes by students, and student use of alcohol or

illegal drugs). *School leadership* was a composite variable, measuring the extent to which student learning is supported by making and altering school policy (e.g., activities and behaviors of the principal, principal observation of classroom instruction).

3.4.5 Analytical Procedures

To test the groups (in this case, boys and girls) in both mean and variance in reading achievement, a two-level HLM model was developed with students nested within schools. The model would postulate that the two sexes have different means and variances in reading achievement scores. The first step of data analysis using HLM constituted a “null” model in which sex was the only independent variable (at the student level). This null model allowed for heterogeneity of variance between the two sexes. The corresponding technique for estimation was the heterogeneous sigma squared specified as a log linear model for testing differences in variance between the two sexes at the student level. In this model, the statistical significance on the differences concerning variance between boys and girls would be examined through the z-ratio. In the second step of data analysis using HLM, especially in the case where heterogeneity could be ascertained, a “full” model was established to investigate whether the result still held when adding the covariates from both student-level and school-level characteristics.

Overall, this HLM model would test for the statistical significance of the differences concerning both the mean of and the variance in reading achievement between boys and girls. This quantification would be accompanied by graphical illustration of the score distributions of boys and girls in reading achievement for visual appreciation. The visualization would reveal which sex had a higher mean and which sex

occupied the top and bottom distribution of reading achievement, thus creating a fuller knowledge about sex differences in reading achievement.

The alpha level for all statistical tests was set as .05. The HLM7.03 software provided the analytical platform for the present study. As mentioned earlier, the full maximum likelihood estimation procedure was applied for all multilevel analyses. A full variance-covariance structure was estimated for each multilevel model.

3.4.6 Results

The model would postulate that the two sexes have different means and variances in reading achievement scores. In the case where heterogeneity could be ascertained, the analysis proceeded to investigate whether the result would still hold when adding the controlling variables from both individual-level and school-level characteristics in the model. Table 3.1 presents the analytical results of this investigation based on three HLM models.

3.4.6.1 Baseline Model (M_0)

It could be seen from the null model that both the mean and the variance are statistically significantly different for boys and girls. On average, girls scored approximately 28.42 higher than boys on reading achievement ($Effect = 28.42, SD = 2.43, p < .05$); girls were also statistically significantly more variable than boys in reading achievement ($Z = -2.246, p < .05$). The final conclusion for the baseline model was that without any control over student and school characteristics, girls performed significantly better than boys in reading achievement, and girls varied significantly more than boys in reading achievement.

3.4.6.2 Intermediate Model (M1)

After student-level background variables were added to the baseline model, the mean difference still existed between males and females. On average, girls performed approximately 26.71 points higher than boys in reading achievement when holding student-level variables constant in the model (*Effect* = 26.71, *SD* = 2.53, $p < .05$). However, once student characteristics were controlled, there was no statistically significant difference in variance between boys and girls ($Z = -0.72$, $p = .47$). The final conclusion for the intermediate model was that, with control over student characteristics, girls still performed significantly better than boys in reading achievement, but boys and girls shared similar variance in reading achievement.

3.4.6.3 Full Model (M2)

After school level variables were added to the intermediate model, the mean difference still existed between males and females. On average, girls performed approximately 26.71 points higher than boys in reading achievement when holding student-level variables constant in the model (*Effect* = 27.31, *SD* = 2.31, $p < .05$). However, once student and school characteristics were controlled, there was no statistically significant difference in variance between boys and girls ($Z = -0.72$, $p = .47$). The final conclusion for the full model was that with control over student and school characteristics, girls still performed significantly better than boys in reading achievement, but boys and girls shared similar variance in reading achievement as seen from the analysis of PISA 2009 dataset.

Variance components were also estimated from the null, intermediate and full model. The null model revealed a statistically significant variance at the school level ($Z =$

-2.246, $p < .05$). Variance components at both student and school levels began to drop once student and school characteristics were added to the null model (see the intermediate model and the full model). Finally, concerning the full model with control over both student and school characteristics, 7% of the variance in reading achievement among students was explained by the full model, and 77% of the variance in reading achievement among schools was explained by the full model. Overall, the full model was effective in explaining a total of 31% of the variance in reading achievement.

Logically (concerning reading achievement from 2009 PISA dataset), because girls had a higher mean than boys but both boys and girls shared similar variance, girls would show a higher mean and occupied the very top distribution of reading achievement, while boys would show a lower mean and occupied the very bottom distribution of reading achievement. To provide a visual appreciation in showing these distributional properties concerning reading achievement between boys and girls, a combined violin plot was produced. Each violin plot showed the mean, interquartile range, and the extreme scores. The visualization revealed the pattern well; that is, girls had a higher mean and occupied the very top distribution of reading achievement, while boys had a lower mean and occupied the very bottom distribution of reading achievement. In addition, the distribution for girls was near normal, but the two peaks for boys indicated that the distribution for boys was not normal. The mode appeared both above and below the mean for boys, which dragged down the mean for boys.

3.5 Final Remarks

The description of distributional properties using the technique often referred to as heterogeneous sigma squared is rare in research literature. This innovative

advancement of HLM would allow researchers to compare the means and the variances between groups simultaneously in one HLM model. This is a perfect situation for making an accurate comparison among groups in terms of distributional properties. The present study purposefully aimed to explore these analytical potentials as methodological innovations for research in educational sciences. With such a statistical model, researchers can not only estimate differences among means, but they can also estimate differences among variances. The nested HLM models from the null model without any independent variables to the full model, with both student-level and school-level variables, are also a good idea to tap into the unique behaviors concerning both means and variances (as shown in the present study). The distributional characteristics also could be illustrated through graphics, which provided a fuller picture to the readers.

Practically, the present study intended to provide a more efficient and effective way to describe and compare the distributional properties of student reading achievement between boys and girls. One of the possible scenarios would be that one sex achieves higher but occupies both top and bottom in the distribution of reading achievement while the other sex achieves lower but is sandwiched in between (as reported in the literature of mathematics education). This did not happen in the present study. The tentative conclusion was that sex-related distributional properties of academic achievement can be quite different between reading and mathematics. The results of similar studies may promote more credible educational reforms through revisiting educational policies and practices concerning sex differences in student academic achievement (based on more robust and precise empirical evidence). The present study has certainly provided a

credible statistical instrument to investigate sex differences in distributional properties of academic achievement.

Finally, the present study has also shown that this statistical instrument is easy for researchers to use and easy for “consumers” to understand. Specifically, it is a relatively straightforward way to set up the model for researchers. Any researcher with basic knowledge and skills in HLM can specify the model effortlessly. In addition, for “consumers,” the final results are easy to understand with the help of the graph showing the distributional properties (i.e., mean and variance) between groups. Any consumer with basic knowledge of descriptive statistics can understand the model easily. Overall, it is the aim of the present study that this statistical instrument may move educational research to a higher level.

Table 3.1

HLM Models of Heterogeneous Sigma Squared Comparing Means and Variances between Boys and Girls in Reading Achievement

	M0		M1		M2	
	Effect	SE	Effect	SE	Effect	SE
Means						
Boys (vs Girls)	-28.42*	2.43	-26.71*	2.53	-27.31*	2.31
Variances						
Boys (vs Girls)	0.09*	0.04	0.03	0.04	0.03	0.04

Note. * $p < 0.05$. Comparisons on means are based on t test. Comparisons on variances are based on Z test. M0 = absolute model (without student and school characteristics). M1 = relative model with student characteristics. M2 = relative model with student and school characteristics. Student characteristics include age, socioeconomic status, family structure (single-parent family vs. both-parent family), immigration status (yes vs. no), and home language (English vs. others). School characteristic include school size, school location (suburban and rural vs. urban), proportion of girls, proportion of certified

teachers, teacher-student ratio, teacher shortage, quality of educational resources, teacher participation, teacher behavior, student behavior, school leadership, and ability grouping (yes vs. no).

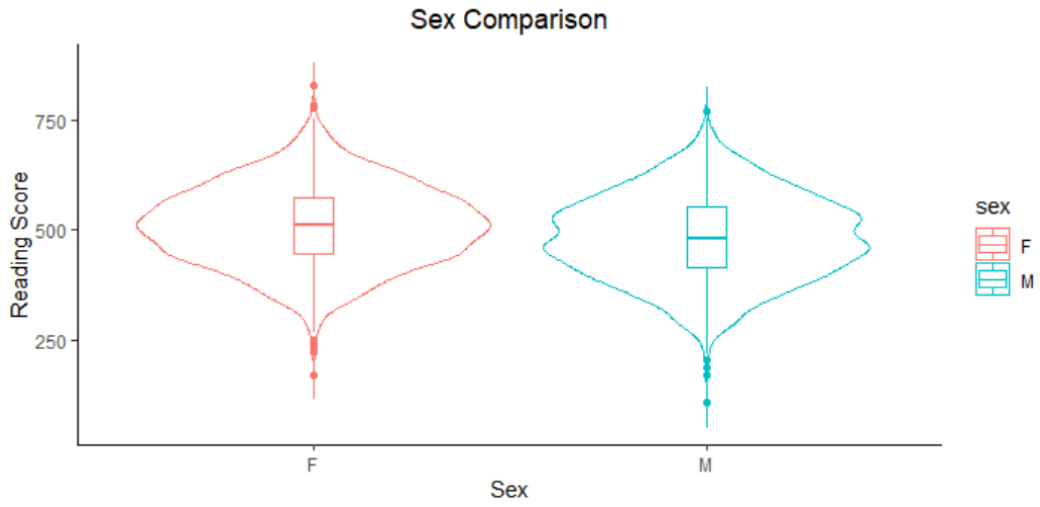
Table 3.2

Estimates of Variance Components and Proportion of Variance Explained for Reading Achievement

Variance Components	M0	M1	M2
Among Students	6345.07	5931.22	5930.55
Among Schools	3354.86	1879.84	785.79
Proportion of Variance Explained			
Among Students			0.07
Among Schools			0.77

Note. M0 = absolute model (without student and school characteristics). M1 = relative model with student characteristics. M2 = relative model with student and school characteristics. Student characteristics include age, socioeconomic status, family structure (single-parent family vs. both-parent family), immigration status (yes vs. no), and home language (English vs. others). School characteristic include school size, school location (suburban and rural vs. urban), proportion of girls, proportion of certified teachers, teacher-student ratio, teacher shortage, quality of educational resources, teacher participation, teacher behavior, student behavior, school leadership, and ability grouping (yes vs. no).

Graphic Illustration



Copyright © Rongxiu Wu 2020

CHAPTER 4: Summary

4.1 Motivation for Methodological Advancement

To help improve and advance research methodology when examining group differences in the outcome measure, two advanced multilevel models that would allow a deeper and more refined look at the issue of sex differences in reading achievement were set up as examples. It was also the motivation of this dissertation research to understand the mechanisms behind group differences in the outcome measure so as to achieve better group equalities in reading education through educational reform in school reading environment.

The traditional way to study the institutional effect on multiple groups of individuals is the dummy-coded approach, which takes the group variable, such as sex, as the dummy variable or groups of dummy variables when there are multiple groups. This approach has an inevitable disadvantage in that whether or not the same institutional variable has the same strength across the groups is hidden. If more categorical variables, such as race, acting as student-level control variables, are added in the first level of the model, the institutional effect for groups gets more complicated. Therefore, the traditional model cannot be used effectively to address the issues of institutional effects on the groups. This lack motivated Study 1 to develop a general multivariate multilevel framework (model) specifically to estimate the institutional effects on multiple groups of individuals.

A good understanding of the distributional properties across groups is an essential part of making group comparisons. The combination of central tendency and variability is the preferred way to describe (and compare) distributions across groups. Almost all

previous studies have a solo focus on differences in either means or variances across multiple groups of individuals. These tests were sometimes performed outside of a certain statistical model that examined either mean or variance differences as a stand-alone procedure. As a result, previous statistical models are not adequate to capture the differences in distributional properties across multiple groups of individuals. There was a lack of credible statistical models that provided a function for tests to be performed inside or within a certain statistical model that examined both mean and variance differences. This lack motivated Study 2 to develop a multilevel model with heterogeneous sigma squared function to compare distributional properties of multiple groups.

4.2 Methodological Advancement

The first model was a restricted multilevel model for examination of institutional effects on multiple groups of individuals, which successfully estimated the institutional effects on multiple groups of individuals. In this restricted multilevel model, the effects for groups (e.g. male effect and female effect) from the same institutional variable can be forced to be equal, and the subsequent significance test can be performed to examine if the restriction held. The same multilevel model can then estimate the amount of the difference by unrestricting the coefficients if the difference really existed. This general, multilevel platform can accommodate any number of groups.

There are several advantages that the current restricted multilevel model has over the traditional multilevel ones. This model is an effective omnibus statistical technique to examine the institutional effects on multiple groups of individuals, which unmasked the specific group dynamics concerning institutional effects with a broad applicability as well as convenient execution. Additionally, it is an easy-to-specify model that employs a

relatively straightforward way to set up the equation. Lastly, the model result was easy to show and interpret for any even entry-level statisticians. From the standpoint of a broader application of the model, firstly, the model can extend the comparison from two to multiple groups besides the simplified version of the two-group comparison, and secondly, the constraints can be applied to not only the groups themselves, but also in the interactions between the level-2 variables so that researchers can examine whether groups share the same interactive pattern regarding the outcome measure. Such an extension opens doors to many other research possibilities that would be complex and tricky to imagine with traditional statistical approaches.

The multilevel model in Study 2 directly provides statistical tests of significance on key distributional properties including both the central tendency (i.e., mean) and variability (i.e., variance). The second model was a multilevel model with heterogeneous sigma squared function to compare the distributional properties of multiple groups. An advanced multilevel model with an embedded analytic function referred to as heterogeneous sigma squared was developed to perform statistical tests of significance to compare means and variances across multiple groups at the same time, which made it convenient to examine the distributional properties comprehensively and simultaneously.

This innovative advancement of HLM would allow researchers to compare the means and the variances between groups simultaneously in one HLM model. This is a perfect situation for making an accurate comparison among groups in terms of distributional properties. The present study purposefully aimed to explore these analytical potentials as methodological innovations for research in educational sciences. With such a statistical model, researchers can not only estimate differences among means, but they

can also estimate differences among variances. The nested HLM models from the null model, without any independent variables to the full model, with both student level and school level variables are also a good idea to tap into the unique behaviors concerning both means and variances (as shown in the present study). The distributional characteristics were also illustrated through graphics, which provided a full picture to the readers.

The two studies of this dissertation research targeted the methodological weaknesses of the research literature concerning institutional effects and distributional properties on multiple groups of individuals. For Study 1, the restricted multilevel modeling rarely has been applied in the research literature on group comparisons. With the application of this methodology, multivariate analysis combining groups meets the necessary condition to conduct a credible group comparison concerning institutional effects. The advantage of carrying out multivariate analysis instead of a series of univariate statistical tests is to deflate the Type I error rate as well as gain more statistical power. For Study 2, the comparison of distributional properties using heterogeneous sigma squared as an integral part of a multilevel model is even rarer in the research literature. This innovative advancement of multilevel modeling will allow researchers to compare the means and the variances in outcomes across groups simultaneously. Overall, it is the hope of the present studies that these statistical instruments may move educational research to a higher level.

4.3 Applications of Advanced Multilevel Models

Study 1 developed a general multivariate multilevel framework (model) specifically to estimate the institutional effects on multiple groups of individuals. With

the employment of 2009 Programme for International Student Assessment (PISA) data, it was an application to examine whether school reading environment had the same effect on reading achievement between boys and girls. Overall, this model has two levels with students nested within schools, and the grouping variable was sex with two categories (boys and girls). Specifically, level 1 was a multivariate model highlighting students' average reading achievement for each sex group (two dichotomous variables) and level 2 was two linear regression equations, one for boys and one for girls. The effects of five school reading environment variables (*diversity of reading*, *enjoyment of reading*, *stimulators of reading*, *daily reading hours*, and *online reading hours*) were constrained respectively to be the same for both boys and girls across the schools. A significance test was performed to examine whether this restriction held true. In the latter case, a new model without restriction was specified if statistically significant results could be deduced from the restricted model, and the two resulting coefficients showed the extent of differences in the school effects on reading achievement between boys and girls. Based on the analysis of the PISA 2009 dataset, it was found that the effect of *enjoyment of reading* and *online reading hours* on reading achievement for boys was statistically different from the effects of the same ones on reading achievement for girls, with student background and school characteristics controlled in the model. The other three school reading environment variables, *diversity of reading*, *stimulators of reading* and *daily reading hour*, did not show any statistically different effects on students' reading achievement between boys and girls.

With the similar PISA 2009 dataset, an application was illustrated to examine the distributional properties concerning reading achievement for boys and girls in a two-level

HLM model in Study 2. In the two-level model, level 1 had sex as the categorical independent variable (dummy coded as boys = 0 and girls = 1) and level 2 had the random intercept modeled by school background variables. It was found that girls performed significantly better than boys in reading achievement, but boys and girls shared similar variance in reading achievement. A violin plot revealed that girls had higher mean and occupied the very top distribution of reading achievement, while boys had a lower mean and occupied the very bottom distribution of reading achievement. The distribution for girls was near normal, but there were two peaks for boys, indicating that the distribution for boys was not normal. The full model explained a total of nearly a third of the variance in reading achievement.

4.4 Tentative Practical Contributions

Together, the studies promoted an exploration in the reading literacy field to add informative insight into the literature of sex differences in reading achievement. For Study 1, it appeared that enjoyment of reading would associate positively with reading achievement, and this association would be stronger for girls than boys. Helping students, particularly girls, enjoy reading is an effective educational strategy to improve reading achievement. Meanwhile, it showed that online reading would actually harm reading achievement with a negative association for both boys and girls, but the effects would be stronger on girls and boys. Because online reading can be irrelevant to schoolwork, educators and parents are advised to monitor and limit online hours that students, particularly girls, spend.

For Study 2, one of the possible scenarios will be that one sex achieves higher but occupies both top and bottom in the distribution of reading achievement while the other

sex achieves lower but is sandwiched in between (as reported in the literature of mathematics education). This scenario did not happen in the present study. The tentative conclusion was that sex-related distributional properties of academic achievement could be quite different between reading and mathematics. The results of similar studies may promote more credible educational reforms through revisiting educational policies and practices concerning sex differences in student academic achievement (based on more robust and precise empirical evidence). The present study has certainly provided a credible statistical instrument to investigate sex differences in distributional properties of academic achievement.

4.5 Limitations and Suggestions

For Study 1, even though the model can specify multiple groups for categorical variables with categories more than two, the model does not directly generate post-hoc analyses for researchers who would like to rank order categories based on outcome measures. In other words, the model can generate an effect for each group and can perform an omnibus test on whether these effects are all the same. However, when the omnibus test is statistically significant, the model cannot perform subsequent post-hoc analysis. The limitation concerns the software, not necessarily the model itself. Specifically, the current software packages cannot perform post-hoc analysis for restricted multilevel models with multiple constrained groups. Researchers need to write program codes (e.g., in R) to extend the analytical function of the restricted multilevel model presented in this study. Additionally, this study adopted several composite variables of student reading behaviors created in PISA to generate measures of school reading environment. These measures were given general labels such as enjoyment of

reading. As in all educational measurement, related constructs such as enjoyment of reading are specifically defined by PISA reading education experts, and these constructs were not intended to be “one-size-fits-all”. Therefore, caution is needed when implied educational policies and practices based on the results of this study. The items that formed each construct such as enjoyment of reading must be studied carefully to fully understand the aspect of, say, enjoyment of reading that PISA intended to measure. In other words, the labels of related constructs such as enjoyment of reading should be contextual to PISA but not general without limit.

Study 2 shares a similar situation. In the presence of a number of groups, the comparison in terms of mean is made between each group with the rest of the groups. The model is not capable of performing detailed post-hoc analysis to rank the order of the group means. The same is true for comparison in terms of variance. In the sigma squared (add-on) equation, it is possible to compare each group with the rest of groups in terms of variance, but the model is not capable of performing detailed post-hoc analysis to rank the order of the group variances. Again, this limitation concerns the software, not necessarily the model itself. Researchers need to write program codes (e.g., in R) to extend the analytical function of the model.

Appendix A

Description of Composite Variables

Variables	Descriptions
Diversity of reading	<p>How often do you read these materials because you want to? (1) Magazines; (2) comic books; (3) fiction (novels, narratives, stories); (4) non-fiction books; (5) newspapers.</p> <p>Response: (a) Never or almost; (b) A few times a year; (c) About once a month; (d) Several times a month; (e) Several times a week.</p>
Enjoyment of reading	<p>How much do you agree or disagree with these statements about reading? (1) I read only if I have to; (2) Reading is one of my favorite hobbies; (3) I like talking about books with other people; (4) I find it hard to finish books; (5) I feel happy if I receive a book as a present; (6) For me, reading is a waste of time; (7) I enjoy going to a bookstore or a library; (8) I read only to get information that I read; (9) I cannot sit still and read for more than a few minutes; (10) I like to express my opinions about books I have read; (11) I like to exchange books with my friends.</p> <p>Response: (a) Strongly disagree; (b) Disagree; (c) Agree; (d) Strongly agree.</p>
Online reading	<p>How often are you involved in the following reading activities? (1) Reading emails; (2) Chat online; (3) Reading online news; (4) Using an online dictionary or encyclopedia; (5) Searching online information to learn about a particular topic; (6) Taking part in online group discussions or forums; (7) Searching for practical information online (e.g. schedules, events, tips, recipes)</p> <p>Response: (a) I do not know what it is; (b) Never or almost never; (c) Several times a month; (d) Several times a week; (e) Several times a day.</p>
Stimulators of reading	<p>In your <test language lesson>, how often does the following occur? (1) The teacher asks students to explain the meaning of a text; (2) The teacher asks questions that challenge students to get a better understanding of a text; (3) The teacher gives students enough time to think about their answers; (4) The teacher recommends a book or author to read; (5) The teacher encourages students to express their opinion about a text; (6) The teacher helps students relate the stories they read to their lives; (6) The teacher shows students how the information in text builds on what they already know.</p> <p>Response: (a) Never or hardly ever; (b) in some lessons; (c) in most lessons; (d) in all lessons.</p>
Daily reading	<p>How much time do you spend reading for enjoyment?</p> <p>Response: (a) Zero hour; (b) half hour or less a day; (c) more than 30 minutes to less than 60 minutes one hour; (d) 1 to 2 hours; (e) more than 2 hours.</p>
Teacher participation	<p>Regarding your school, who has a considerable responsibility for the following task? (1) selecting teachers for hire; (2) firing teachers; (3) establishing teachers' starting salaries; (4) determining teachers' salaries increases; (5) formulating the school</p>

budget; (6) deciding on budget allocations within the school; (7) establishing student disciplinary policies; (8) establishing student assessment policies; (9) approving students for admission to the school; (10) choosing which textbooks are used; (11) determining course content; (12) deciding which courses are offered.

Response: (a) Principals; (b) teachers; (c) school governing board; (d) regional or local education authority; (e) national education authority.

Teacher behavior

In your school, to what extent is the learning of students hindered by the following phenomenon? (1) teachers' low expectation of students; (2) poor student-teacher relations; (3) teachers not meeting individual students' needs; (4) teacher absenteeism; (5) staff resisting change; (6) teachers being too strict with students; (7) students not being encouraged to achieve their full potential.

Response: (a) Not at all; (b) very little; (c) to some extent; (d) a lot.

Student behavior

In your school, to what extent is the learning of students hindered by the following phenomenon? (1) student absenteeism; (2) disruption of classes by students; (3) students skipping classes; (4) students lacking respect for teachers; (5) student use of alcohol or illegal drugs; (6) students intimidating or bullying other students.

Response: (a) not at all; (b) very little; (c) to some extent; (d) a lot.

School leadership

Below you can find statements about your management of this school. Please indicate the frequency of the following activities and behaviors in your school during the last school year. (1) I make sure that the professional development activities of teachers are in accordance with the teaching goals of the school; (2) I ensure that teachers work according to the school's educational goals; (3) I observe instruction in classroom; (4) I use student performance results to develop the school's educational goals; (5) I give teachers suggestions as to how they can improve their teaching; (6) I monitor students' work; (7) When a teacher has problems in his/her classroom, I take the initiative to discuss matters; (8) I inform teachers about possibilities for updating their knowledge and skills; (9) I check to see whether classroom activities are in keeping with our educational goals; (10) I take exam results into account in decisions regarding curriculum development; (11) I ensure that there is clarity concerning the responsibility for coordinating the curriculum; (12) When a teacher brings up a classroom problem, we solve the problem together; (13) I pay attention to disruptive behavior in classrooms; (14) I take over lessons from teachers who are unexpectedly absent.

Response: (a) Never; (b) Seldom; (c) Quite often; (d) Very often.

Appendix B

Description of Student and School Characteristics

	Description
Student-Level Variables	
Sex	Are you female or male? 1) female, 2) male. Dummy: 1) = 1; 2) = 0.
Age	When were you born?
Father (mother) socioeconomic status (SES)	What is your father's (mother's) job? 1) worker, 2) farmer, 3) self-employed, 4) service sector, 5) government employee, 6) education or medicine sector, 7) business (management) sector, 8) military sector, 9) migrant worker, 10) unemployed. Index. Continuous.
Single-parent household	What is the composition of your family? 1) both-parent household (biological parents), 2) both-parent household (stepmother or stepfather), 3) single-parent household (father or mother passed away), 4) single-parent household (parents divorced). Dummy: 1), 2) = 0; 3), 4) = 1.
Immigration status	In what country were you and your parents born? 1) United States, 2) Other country. Dummy: 1) = 1; 2) = 0.
School-Level Variables	
School (enrollment) size	What is the total number of students in your school? Continuous (in number of units with 100 as one unit).
School location	Which of the following definitions best describes the community in which your school is located? 1) A village, hamlet or rural area; 2) a small town; 3) a town; 4) a city; 5) a large city.
Percentage of girls	Number of girls divided by school (enrollment) size. Continuous (percentage).
School mean (parental) SES	Aggregation from students within a school (with father and mother SES averaged for each student). Continuous.
Percentage of teachers with at least a bachelor's degree	What is the number of teachers at each education level in your school? 1) senior high school, 2) vocational high school, 3) professional college (2 or 3 years), 4) undergraduate and higher education level. Continuous (percentage).
Teacher shortage	How do you evaluate the adequacy of (physics, biology, and geography) teachers in your school? 1) severe shortage, 2) not enough, 3) basically enough, 4) full capacity. Continuous.
Teacher quality	How do you evaluate the quality of (physics, biology, and geography) teachers in your school? 1) very low, 2) low, 3) high, (d) very high. Continuous.

Teacher leadership	Which leadership positions are you in except teaching? 1) leader of a teacher group in a subject, 2) leader of a teacher group at a grade level, 3) leader of school youth group, 4) director of an office, 5) manager of your school, 6) none. Continuous (count of selected positions).
School leadership	How often do you work on the following tasks? 1) offer opportunities for teachers to express their opinions and suggestions, 2) treat each teacher fairly, 3) offer opportunities for teachers on decision making, 4) ask for advices from teachers on problems in school management, 5) promote democratic management of teachers in school administration, 6) make school affairs transparent. Response: (a) never, (b) seldom, (c) sometimes, (d) often, (e) always. Continuous (valid average). Cronbach's alpha is .85.
Principal school management	How often do you work on the following tasks? 1) participate in various meetings on campus and off; 2) teach students; 3) observe and evaluate teachers' lessons as well as participate in teaching and research activities; 4) communicate with teachers and listen to their views and ideas; 5) cope with monitoring and assessments of a school district; 6) plan and examine educational research, teaching, and allocation of funds. Response: (a) never, (b) seldom, (c) sometimes, (d) often, (e) always. Continuous (valid average). Cronbach's alpha is .44.
Principal support for teaching	How often do you work on the following tasks? 1) allow certain autonomy for teachers to make their instructional decision, 2) support various departments to actively promote teaching and learning, 3) consider teachers' expertise and abilities when scheduling classes, 4) encourage teachers to organize research group in various subjects, 5) provide sufficient teaching materials for teachers, 6) provide teachers with effective professional guidance and assistance. Response: (a) never, (b) seldom, (c) sometimes, (d) often, (e) always. Continuous (valid average). Cronbach's alpha is .83.
Principal support for professional development	As a principal, how do you do in the following areas? 1) take the initiative to ask teachers about their training needs and provide information, materials, and channels to meet their needs; 2) give different incentives depending on the needs of professional development of teachers; 3) operate school-based career planning to promote professional development. Response: (a) never, (b) seldom, (c) sometimes, (d) often, (e) always. Continuous (valid average). Cronbach's alpha is .86.

Appendix C

Descriptive Statistics of Student Characteristics and Spearman Correlation (N = 5233)

	1	2	3	4	5	6
1. Age	1.00					
2. Sex (female =1, male =0)	.01	1.00				
3. Socioeconomic Status	.01	-.01	1.00			
4. Both-parent household (yes = 1, no = 0)	.02	.01	-.23*	1.00		
5. Native-born student (yes = 1, no = 0)	.02	.01	-.26*	-.02	1.00	
6. English-speaking family (yes = 1, no = 0)	.01	.01	.28*	.02	-.68*	1.00
<i>M</i>	15.79	.49	.15	.75	.19	.87
<i>SD</i>	.30	.50	.92	.43	.34	.33

Note. * $p < .05$.



Appendix D
Spearman Correlation of School Characteristic Variables (N = 165)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1.00														
2	.32*	1.00													
3	.07*	.15	1.00												
4	.13	-.27*	-.01	1.00											
5	-.14	.02	.11	-.24*	1.00										
6	.17*	.06	-.03	-.03	-.41*	1.00									
7	.01	-.08	-.03	.21*	.04	-.10	1.00								
8	-.09	-.07	-.08	.05	-.39*	.26*	-.01	1.00							
9	-.05	-.12	.09	-.10	-.20*	.15	-.01	.54*	1.00						
10	.09	.04	-.06	-.06	-.04	.18*	-.01	.27*	.09	1.00					
11	-.08	.19*	-.01	-.09	-.01	.06	-.02	-.02	-.09	.08	1.00				
12	.01	.17*	.30*	-.07	-.15	.22*	-.05	.08	.06	-.03	.37*	1.00			
13	.05	.09*	.16*	-.01	-.19*	.17*	.09	.28*	.17*	.06	.36*	.35*	1.00		
14	-.04	.17*	.21*	-.05	-.09	.11	-.01	-.03	-.12	-.04	.28*	.67*	.11	1.00	
15	.36*	.31*	.11	-.08	-.12	.24*	-.09	.05	.13	.12	.30*	.27*	.40*	.09	1.00

Note. * $p < .05$. 1 = School (enrollment) size. 2 = School location (city=1, town=0). 3 = Proportion of girls in the school. 4 = Proportion of certified teachers. 5 = Teacher shortage. 6 = Quality of the schools' educational resources. 7 = Teacher participation. 8 = Teacher behavior. 9 = Student behavior. 10 = School leadership. 11 = Diversity of Reading. 12 = Joy of Reading. 13 = Stimulator of Reading. 14 = Reading hours. 15 = Online reading.

Appendix E

The Data Format of Category Variables in SPSS in Study 1

 Boy	 Girl
.00	1.00
1.00	.00
1.00	.00
.00	1.00
.00	1.00
1.00	.00
1.00	.00
1.00	.00
1.00	.00
.00	1.00
.00	1.00
.00	1.00
1.00	.00

Note. Different from using one dummy variable indicating two groups, in study 1 two variables were used to indicate the two groups when working on the data preparation in SPSS.

References

- Adams, R., & Wu, M. (Ed.). (2002). *Programme for international student assessment (PISA): PISA 2000 technical report*. Paris: OECD.
- Barnett, R., Brennan, R.T., Raudenbush, S.W., & Marshall, N.L. (1994). Gender and the relationship between marital-role equality and psychological distress. *Psychology of Women Quarterly*, 18, 105-127.
- Baye, A., & Monseur, C. (2016). Gender differences in variability and extreme scores in an international context. *Large-scale Assessments in Education*, 4(1), 1-16.
- Beller, M., & Gafni, N. (1996). The 1991 International Assessment of Educational Process in Mathematics and Sciences: the gender differences perspective. *Journal of Educational Psychology*, 88(2), 365-377.
- Brozo, W.G., Shiel, G., & Topping, K. (2008). Engagement in reading: Lessons learned from three PISA countries. *Journal of Adolescent & Adult Literacy*, 51(4), 304-315.
- Brozo, W., Sulkunen, S., Shield, G., Garbe, Ch., Pandian, A., & Valtin, R. (2014). Reading, gender, and engagement. *Journal of Adolescent and Adult Literacy*, 57(7), 584-593.
- Costa, P., & Araújo, L. (2018) Skilled Students and Effective Schools: Reading Achievement in Denmark, Sweden, and France. *Scandinavian Journal of Educational Research*, 62(6), 850-864.
- Chiu, M.M., & McBride-Chang, C. (2006). Gender, context and reading: A comparison of students in 43 countries. *Scientific studies of Reading*, 10(4), 331-362.
- Chudowsky, N., & Chudowsky, V. (2010). State Test Score Trends through 2007-08, Part 5: Are There Differences in Achievement between Boys and Girls?. *Center on Education Policy*.

- Cullinan, B.E. (1992). *Read to me: Raising kids who love to read*. New York: Scholastic.
- Elley, W.B. (Ed.). (1994). *The IEA study of reading literacy: Achievement and instruction in thirty-two school systems*. Exeter, UK: Pergamon.
- Everson, H., & Millsap, R. (2004). Beyond individual differences: Exploring school effects on SAT scores. *Educational Psychologist, 39*, 157-172.
- Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist, 43*, 95-103.
- Feingold, A. (1992a). Sex differences in variability in intellectual abilities: a new look at an old controversy. *Review of Educational Research, 62*(1), 61-84.
- Feingold, A. (1992b). The greater male variability controversy: Science versus politics. *Review of Educational Research, 62*, 89-90.
- Feingold, A. (1994). Gender differences in variability in intellectual abilities: A cross-cultural perspective. *Sex Roles, 30*, 81-91.
- Glass, G. V., & Hopkins, K. D. (1984). *Statistical Methods in Education and Psychology*. Englewood Cliffs, NJ: Prentice-Hall.
- Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.) London: Edward Arnold.
- Guthrie, J., & Wigfield, A. (2000). Engagement and motivation in reading. In M.L. Kamil, P. Mosenthal, P.D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 403-422). Mahwah, NJ: Erlbaum.
- Halpern, D.F., Benbow, C.P., Geary, D. C., Gur, R.C., Hyde, J.S., & Gernsbacher, M.A. (2007). The Science of sex differences in science and mathematics. *Psychological Science in the Public Interest, 8*(1), 1-51.

- Hedges, L.V., & Friedman, L. (1993a). Gender differences in variability in intellectual abilities: A reanalysis of Feingold's result. *Review of Educational Research*, 63, 94-105.
- Hedges, L. V., & Friedman, L. (1993b). Computing gender difference effects in tails of distributions: The consequences of differences in tail size, effect size and variance ratio. *Review of Educational Research*, 63, 110-112.
- Hedges, L.V., & Nowell, A. (1995). Differences in mental test scores, variability, and numbers of high-scoring individuals. *Science*, 269, 41-45.
- Hill, C., Bloom, H., Black, A. & Lipsey, M. (2008) Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, 2(3), 172-177.
- Hox, J. J., & Bechger, T. M. (1998). An introduction to structural equation modeling. *Family Science Review*, 11, 354–373.
- Hox, J. (2010). *Multilevel analysis: Technique and application*. Routledge.
- Humphreys, L.G. (1988). Sex differences in variability may be more important than sex differences in mean. *Behavioral and Brain Sciences*, 11, 195-196.
- Johnson, S. (1996). The contribution of large-scale assessment programmes to research on gender differences. *Educational Research and Evaluation*, 2(1), 25-49.
- Krashen, S. D. (2004). *The power of reading: Insights from the research: Insights from the research*. ABC-CLIO.
- Kreft, I., & de Leeuw, J.D. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.
- Lafontaine, D., & Monseur, C. (2009). Gender gap in comparative studies of reading comprehension: to what extent do the test characteristics make a difference? *European Educational Research Journal*, 8(1), 69–79.

- Lee, V.E., & Bryk, A.S. (1989). A multilevel model of the distribution high school achievement. *Sociology of Education*, 62(3), 172-192.
- Lee, V.E., Zuze, T., & Ross, K. (2005). School effectiveness in 14 sub-Saharan African countries: Links with 6th graders' reading achievement. *Studies in Educational Evaluation*, 31,207-246.
- Lenkeit, J., Chan, J., Hopfenbeck, T. N., & Baird, J. (2016). A review of the representation of PIRLS related research in scientific journals. *Educational Research Review*, 16, 102–115.
- Lietz, P. (2006). A meta-analysis of gender differences in reading achievement at the second school level. *Studies in Educational Evaluation*, 32(4), 317-344.
doi:10.1016/j.stueduc.2006.10.002
- Liu, O.L., & Wilson, M. (2009). Gender differences in large-scale math assessments: PISA trend 2000 and 2003. *Applied Measurement in Education*, 22, 164-184.
- Longford, N. T. (1993). *Random coefficient models*. Oxford, UK: Clarendon Press.
- Lynn, R., & Mikk, J. (2009). Sex differences in reading achievement. *Trames*, 13(1), 3-13.
- Ma, X. (1995). Gender Differences in Mathematics Achievement Between Canadian and Asian Education Systems. *The Journal of Educational Research*, 89(2), 118-127.
- Ma, X. (1999). Gender differences on growth in mathematical skills during secondary grades: A growth model analysis. *Alberta Journal of Educational Research*, 45, 448-466.
- Ma, X. (2002). Early acceleration of mathematics students and its effects on growth in self-esteem: A longitudinal study. *International Review of Education*, 48, 443-468.
- Ma, X., & Klinger, D. A. (2000). Hierarchical linear modeling of student and school effects on academic achievement. *Canadian Journal of Education*, 25, 41-55.

- Ma, X., & Willims, J.D. (2004). School disciplinary climate: Characteristics and effects on eighth grade achievement. *Alberta Journal of Educational Research*, 50, 169-189
- Ma, X., Ma, L., & Bradley, K. (2008). Using multilevel modeling to investigate school effects. In AA O'Connell & DB McCoach (Eds.), *Multilevel modeling of educational data* (pp. 59-110). Charlotte, NC: Information Age.
- Marks, G. (2008). Accounting for the gender gaps in student performance in reading and mathematics: Evidence from 31 countries. *Oxford Review of Education*, 34(1), 89-109.
- McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, 22(1), 114.
- Mol, S. E., & Bus, A.G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin*, 137, 267-296.
- Mullis, L.V.S., Martin, M.O., Foy, P., & Drucker, K.T. (2012). *PIRLS 2011 international results in reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Nowell, A., & Hedges, L.V. (1998). Trends in gender differences in academic achievement from 1960 to 1994: an analysis of difference in mean, variance, and extreme scores. *Sex Roles*, 39 (1), 21-43.
- Nordquist, R. (2007). Online reading: glossary of grammatical and rhetorical terms. Retrieved from <http://www.thoughtco.com/what-is-online-reading-1691357>
- OECD. (2001). *Knowledge and skills for life: First results from PISA 2000*. Paris: OECD Publishing.
- OECD. (2002a). *PISA 2000 technical report* (Paris, Organization for Economic Co-operation and Development).

- OECD. (2007a). *PISA 2006: Science Competencies for Tomorrow's World* (Vol. 1: Analysis & Vol. 2: Data). OECD, Paris.
- OECD. (2010). *PISA 2009 results: Learning to learn- Student engagement, strategies and practices* (Vol.3).
- OECD. (2010b). *PISA 2009 results: What students know and can do – Student performance in reading, mathematics and science* (Vol. I). Paris: OECD Publishing.
doi:10.1787/9789264091450-en
- OECD. (2013). *PISA 2012 Results: Ready to Learn: Students' Engagement, Drive and Self-Beliefs (Volume III)*. PISA: OECD Publishing. Retrieved from <https://doi.org/10.1787/9789264201170-en>
- Opdenakker, M.C., & Van Damme, J. (2000) The importance of identifying levels in multilevel analysis: An illustration of ignoring top or intermediate levels in school effectiveness research. *School Effectiveness and School Improvement, 11*,103-130.
- Raudenbush, S., & Bryk, A.S. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59*, 1-17.
- Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models* (2nd ed.). Newbury Park, CA: Sage.
- Rosbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Deaper, D., Langford, I., & Lewis, T. (2000). *A user's guide to MLwiN*. University of London.
- Schwabe, F., McElvany, N., & Trendtel, M. (2015). The School Age Gender Gap in Reading Achievement: Examining the Influences of Item Format and Intrinsic Reading Motivation. *Reading Research Quarterly, 50*(2), 219-232.

- Shiel, G. (2006). The PISA assessment of reading literacy. *Irish Journal of Education*, 37, 79-100.
- Shiel, G., Cosgrove, J., Sofroniou, N., & Kelly, A. (2001). *Ready for life? The literacy achievements of Irish 15-year olds with comparative international data*. Dublin, Ireland: Educational Research Centre.
- Snijders, T.A.B., & Bosker, R.J. (1999). *Multilevel analysis*. Thousand Oaks, CA: Sage.
- Solheim, O., & Lundetræ, K. (2018). Can test construction account for varying gender differences in international reading achievement tests of children, adolescents and young adults? – A study based on Nordic results in PIRLS, PISA and PIAAC. *Assessment in Education: Principles, Policy & Practice*, 25(1), 107-126.
- Stancel-Piatak, A., Mirazchiyski, P., & Desa, D. (2013). Promotion of reading and early literacy skills in schools: A comparison of three European Countries. *European Journal of Education*, 48, 449–510.
- van Langen, A., Bosker, R., & Dekkers, H. (2006). Exploring cross-national differences in gender gaps in education. *Educational Research and Evaluation*, 12(02), 155-177.
- Walkerdine, V. 1988. *Mastery of Reasons: Cognitive Development and the Production of Rationality*. London: Routledge.
- Willims, J.D. (1992). *Monitoring school performance: A guide for educators*. Washington, DC: Falmer.

Vita

Education:

- January 2020 Graduate Certificate in Applied Statistics, Department of Statistics,
University of Kentucky
- January 2020 Graduate Certificate in Developmental and Intellectual Disabilities,
Human Development Institute
- June 2015 Master of Science, Foreign Linguistics and Applied Linguistics,
Shanghai University
- June 2011 Bachelor of Science, English Literature, Tongling University, Anhui

Publications:

- Wu, R.**, Wu, X., Peabody, MR., & O'Neill, TR (2019). A Longitudinal Study on the Differences in Canadian and US Medical Student Preparation for Family Medicine. *Family Medicine*, 51.
- Wu, R.**, Corbin, A., Goldstein, P., Adams, C., Rumrill, P., Bishop, M., and Sheppard-Jones, K. (2019). Preliminary examination of return to work interests among unemployed individuals with Multiple Sclerosis. Human Development Institute Research Brief.
- Rumrill, P., **Wu, R.**, Goldstein, P., Adams, C., Sheppard-Jones, K., Lee., B., Bishop., M., Minton. D., and Leslie, M. (2020). Importance and Satisfaction Ratings on 38 Key Employment Concerns among African American Women with Multiple Sclerosis. *Journal of Vocational Rehabilitation*, 1, 1-11. DOI: 10.3233/JVR-191068
- Su, Q., Wang, L., **Wu, R.** (2019). The Concept, Goal and Support System of the Internationalization of Higher Education in Sweden-Interpretation and Reference of Swedish "Internationalization of Swedish Higher Education and Research- A Strategic Agenda". *Heilongjiang Researchers on Higher Education*, 3, 299-304.
- Wu, X. **Wu, R.** & Peabody, MR, & O'Neill, TR. Detecting cross-cultural differential item functioning for increasing validity: An example from the American board of family medicine in-training examination. *Educ Health Prof* 2018(1), 19-23.

Rongxiu Wu

March 30, 2020

Copyright © Rongxiu Wu 2020