Quantifying the Role of Sex and Ethnicity on the Relationship between Teacher Judgement and Student Performance on Standardized Exams in Mathematics and Reading

Sareh Meshkinfam, Julie Ivy and Amy Reamer
North Carolina State University

Abstract

It is imperative that aspiring engineering students develop a solid foundation of mathematics early. Educators are critical in influencing a student's future performance as well as the student's perception of their own abilities. Teacher judgment may influence a student's future academic placement, underscoring the importance of examining the relationship between teacher judgement and student performance. This paper evaluates these relationships longitudinally and measures the effect of grade level using standardized End-of-Grade test scores for third-to-eighth grades from 2006-2013 in North Carolina, including 2,542,451 students, via correlation analyses and multiple regression models in mathematics and reading. Regression models suggest the relationship between teacher judgment and student performance in reading and mathematics is influenced by sex, ethnicity and their intersection over time and by grade level. Disparities in the correlation between teacher judgment and student performance by sex and race may translate to later disparities in participation in STEM degree programs.

Keywords

End-of-Grade, Teacher Judgment, Student Demographics (e.g. Gender, Ethnicity), Regression Models, Standardized Tests, Student Performance, Mathematics Education

Introduction

Proficiency in mathematics is essential for students aspiring to obtain an engineering degree or take an engineering course in college. Elementary school mathematics and science construct the basis for Science, Technology, Engineering, and Mathematics (STEM) learning in higher education[1]. Research suggests that mathematics is "the key academic hurdle" to the successful graduation of engineering students[2]. Mathematics is often a roadblock to entry into engineering degree programs, specifically, "students' difficulty with higher level school mathematics is often blamed for the declining number of entrants to engineering degree courses" (Croft and Grove 2006; King 2008; Prieto et al. 2009)[2]. Student performance in higher level mathematics courses begins with their preparation in elementary and middle school. In order to understand students' preparation in mathematics, it is critical to understand the role of teachers in early education.

Teacher judgment may influence a student's future academic placement, underscoring the importance of examining the relationship between teacher judgement and student performance[3,4,13–18,5–12]. A teacher's judgment may influence not only the teacher's anticipation of the student's competence[10,19–21] , but the student's perception of his or her own academic

ability[10,16,22–26] . In the literature, teacher judgment is defined as a teacher's evaluation of a student's expected performance. The accuracy of teacher judgment is his/her capability to assess student performance correctly on a specific subject[27].

A number of studies have assessed a teacher's judgment of student performance as compared to the students' actual academic achievements as a function of demographic characteristics (e.g. gender, ethnicity, socioeconomic status)[9,12,35,16,28–34]. It is widely believed that students' grades are not only an exhibition of their academic performance, but may also be affected by inherent characteristics like gender, ethnicity, behavior, and socioeconomic status[36]. Correlation analysis (i.e. Pearson), multi-level and hierarchical regression models have been used to assess the strength of this relationship[9,16,42–44,18,32,35,37–41]. Some studies suggest that teacher judgment is influenced by student gender[12,16,39,45]; while, others fail to demonstrate significant differences in the relationship[6,8,9,43,46]. Very few studies have identified differences in this relationship based on race and/or ethnicity; some found that the teacher underestimated the minority students[39,47], others believe teachers overestimate a minority group[16] , still others found no significance differences based on ethnicity[35]. One study, Ready and Wright (2011), stated that teachers have a significantly negative prejudice against male / minority / low socioeconomic students[47]. Some reports concluded that teacher judgment was less accurate for low-achieving, lower socioeconomic status students[9,17,46,47,28–30,32,33,40,41,43]. Several of these studies found in the literature are limited in scope and/or scale. Many analyzed small samples, e.g., students from one or two grade levels in one school[12,28,33,41], while others considered large samples including data from all public schools in one or multiple states[9,16,18,30,32,42] , but they either focused on a single cohort or one grade in multiple periods, ignoring the longitudinal relationships[9,16,17,30,32,42].

In this research, we use data obtained from the North Carolina Education Research Data Center (NCERDC) to track three different cohorts of North Carolina students from 3rd to 8th grade from 2006 to 2013. This paper evaluates the relationship between students' performance and teacher judgment longitudinally and measures the effect of students' demographics within grade level using the standardized End-of-Grade (EOG) test scores in mathematics and reading comprehension, referred to hereafter as "reading". We conduct statistical analyses for significance of the correlation and develop regression models to assess the relationship between students' EOG test performance and teacher judgment as a function of student demographics over time. The following research questions are assessed: 1) What is the relationship between student EOG performance and the corresponding teacher judgment? 2) How is the relationship affected by student demographics, such as gender and ethnicity, and grade level over time?

Methods

Data Overview

The NCERDC dataset includes information on more than 600,000 students and teachers per year from North Carolina's public schools from 1993 to 2015. School-level data is gathered from the North Carolina Department of Public Instruction (NCDPI), which records students' performance on the EOG tests annually. The North Carolina EOG tests are administered in reading and mathematics and are state-mandated, curriculum-based, multiple-choice tests, that evaluate students' performance and assess how well students meet grade-level expectations at the end of a

given academic year. Raw test scores are scaled and converted to one of four annually defined scaled categories, known as achievement levels, for each grade level[48,49]. Achievement levels I and II are defined as "non-proficient" in a given subject by the state of North Carolina, while levels III and IV indicate "proficiency"[49]. Each year, teachers are asked to predict each student's achievement level score (I, II, III or IV) and this data is recorded in the NCERDC datasets.

In this study, we analyze students' EOG achievement level scores in reading and mathematics, the corresponding teacher judgment, and demographic characteristics of three cohorts of students from 2006-2013, including 2,542,451 students in total. The three cohorts are defined as students who progressed from 3rd to 8th grade from 2006 to 2011 (cohort1), from 2007 to 2012 (cohort2), and from 2008 to 2013 (cohort3) without retention or skipping. We excluded records missing information relevant to this study. Table 1 presents the total number of students for each grade level by cohort and summarizes the distribution of students' demographic characteristics. The variable names related to ethnicity used in this study were defined by the NCERDC.

Table 1. Number of students in each grade and cohort for NCERDC (Population Data from 2006 - 2013), and their corresponding demographics

Total number					Demographics		
Grade \ Cohort	1	2	3	Grade Total	Gender	Male	50.34%
3	104,808	112,280	135,755	352,843		Female	49.66%
4	108,971	112,250	159,688	380,909	Ethnicity	White	54.86%
5	132,679	160,177	160,409	453,265		Asian	2.31%
6	154,715	152,570	154,070	461,355		Black	26.31%
7	157,840	158,474	159,886	476,200		Hispanic	11.16%
8	151,945	152,990	112,944	417,879		American Indian	1.53%
Cohort Total	810,958	848,741	882,752	**2,542,451**		Multi-Racial	3.83%

Statistical Analysis

Pearson correlation analysis and generalized multiple regression models were used to analyze the relationship between students' demographics and teacher judgment with EOG scores in reading and mathematics. Correlations between student's EOG and teacher judgment in each grade, cohort, and based on students' gender and ethnicity are compared with each other using hypothesis tests to evaluate the significance of the differences in each comparison. Multiple regression models were developed to assess the longitudinal and cross-sectional relationships between students' EOG scores and their corresponding teacher judgment scores based on their gender, ethnicity, grade level, year, and all possible interactions. Proc Corr and Proc Mixed in SAS®9.4 were used to calculate the correlations matrix and build the regression models for reading and mathematics, respectively.

Results

Correlation Analysis

The strength of the relationship between students' EOG scores and the corresponding teacher judgment was measured via Pearson correlation for each grade level. The correlations are all significant and their values ranged from 0.607 to 0.677. Figures 1.a to 1.c show the trend and

differences in correlations from a variety of perspectives; Figure 1.a illustrates the trend in correlation for all students in reading and mathematics and that correlation is significantly higher for mathematics in all grades, where we notice an increasing and then decreasing trend over the grades. Figure 1.b shows the correlation for male and female students by grade, where each have a similar trend as the overall population. The correlation for males are significantly higher than females, in both reading and mathematics. Similarly, Figure 1.c displays the relationship based on ethnicity group, in which the correlation is higher for Asian and White students and lower for Black and American Indian students, and the differences/gap increases with grade. Detailed information about the correlation values, and related p-values can be found in Table A.1 to A.5 in the appendix.

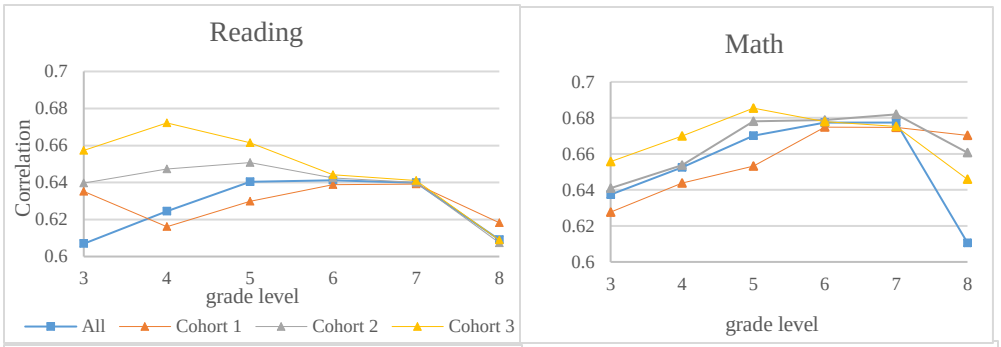

Figure 1.a. Correlation between EOG achievement score and teacher judgment in mathematics and reading by grade level for each cohort and overall data.

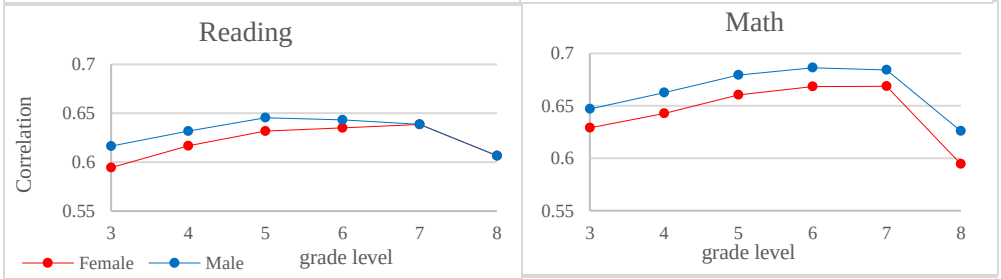

Figure 1.b. Correlation between EOG achievement score and teacher judgment in mathematics and reading by grade level based on students' gender.

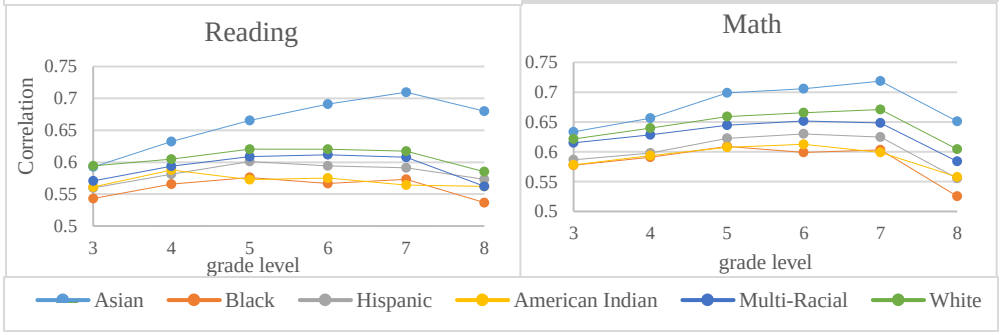

Figure 1.c. Correlation between EOG achievement score and teacher judgment in mathematics and reading by grade level based on students' ethnicity.

Regression Analysis

We modeled the relationship between student's EOG test scores and their corresponding teacher judgment score, gender, grade level, year and their corresponding interactions using samples of three cohorts of data in 2006-2011, 2007-2012, and 2008-2013. The detailed result is provided in Table A.6 in the appendix.

The regression model indicates significant relationships between students' *gender* (negative effect for male in reading and positive one in mathematics), *ethnicity* (negative effect for non-white groups who are not Asian in reading and mathematics), *grade level* (the positive effects (higher in mathematics), decreased for higher grade levels as compared with 8th grade), *year* (negative effects (effects are greater for mathematics)), and *cohorts* (large and positive effects decreased for more recent cohorts). There is a large and highly significant intercept especially for reading which indicates the general underestimation for students (by almost one achievement level) within this subject.

The regression model also shows that teacher judgment seems to have a significantly larger effect for male students in mathematics. Similarly, the results demonstrate teacher judgment differs significantly by race and ethnicity in reading and mathematics (with the exception of Hispanic and Multi-Racial in reading). Further, teacher judgment is significantly related to students' grade level (more so in reading), and year (more in mathematics).

Figure 2 shows the students' predicted achievement level scores based on regression models for a given teacher judgment level (here, a teacher judgment level of 4 is shown (used TJ abbreviation in figures)) in reading and mathematics. There is variability for reading by grade level based on students' gender and ethnicity for each cohort. While in mathematics, students' achievement level scores are higher for the more recent cohort, and increase across the grade levels. There is a noticeable decline in students' scores in 2013, when there was a significant change to the EOG exam and grade level standards[50] which was expected to result in lower test scores. Plots for predicted scores for each value of teacher judgment (1 to 4), for each cohort in reading and mathematics are provided in Figure A.1 in the appendix.

Conclusion

This research explores the vital role an educator, attuned to each individual student's performance, plays in the evolution of a student's cognition over time. A review of current literature concludes that primary and middle school educators influence a student's future performance and the student's perception of his or her own academic abilities. In this research, we studied the relationship between students' mathematics and reading EOG test scores and teacher judgment as a function of student demographics from 3rd to 8th grade for three cohorts of students in North Carolina. In general, teacher judgment and student performance are positively correlated, and these correlations are generally higher in mathematics than reading. This result is expected given that reading can be assessed more subjectively than mathematics. When teacher judgment is documented as level 3 or 4 (the state's definition of "proficient" in a subject in a given grade level), predicted student performance is lower, but, for teacher judgment levels of 1 and 2 ("non-proficient") the predicted student performance is higher. There is more variability in prediction of student performance as a function of teacher judgment in reading than in mathematics. Across cohort (an indicator of time), the predicted student performance pattern is roughly the same. Furthermore, the relationship between teacher judgment and student performance is influenced by students' demographics. Correlation is higher for male students than females, and for Asian and White students as compared to other ethnicities in both reading and mathematics. It appears that for teacher judgment levels of 3 or 4, the predicted student performance matches that level for Asian and White students, but not for students of other

ethnicities studied. When teachers project a score of 1 or 2 ("non-proficient") for a student, the predicted EOG test score is higher for all racial and ethnic groups, but is highest for Asian and White students. The predicted performance increases across grade level for mathematics especially.

This research develops models to assess student performance on standardized exams in mathematics and reading as a function of teacher judgment as well as student demographics via correlation and regression analyses. Disparities in the correlation between teacher judgment and student performance by sex and race may translate to later disparities in participation in STEM degree programs. Teacher judgement may influence a student's appreciation of their own skills and this judgment may be influenced by a student's race and ethnicity. This type of research may offer an opportunity to remove early barriers to participation in engineering by underrepresented groups. Our analysis of student socioeconomic status (SES) was limited in this study, due to lack of access to consistent variables representing SES across all years and cohorts. Areas of future work include further analysis of the role of ethnicity and SES on student performance. This research is an initial phase in establishing a basis of exhaustive statistical indication for detecting points along a student's learning trajectory where teacher judgment proves especially impactful. Thus, further research can be done on causal analysis, decision trees, and logistic analysis for the placement of well-timed learning interventions.

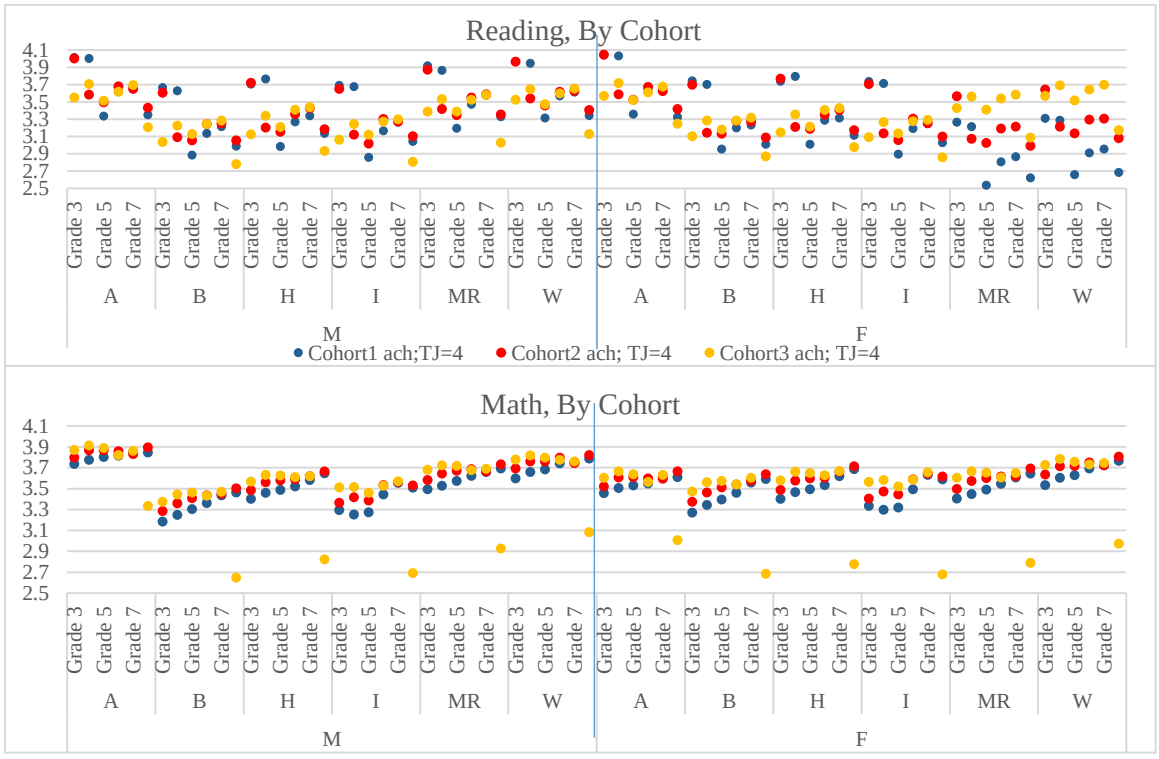

Figure 2. Predicted students' achievement level score, using the regression model when teacher judgment is level 4 in mathematics and reading for each cohort by grade level based on students' gender, and ethnicity.

References

1. Epstein D, Miller RT. Slow off the Mark. *Cent Am Prog*. 2011;(May):1-21. www.americanpress.org.

2. Goold E, Devitt F. The role of mathematics in engineering practice and in the formation of engineers BT - SEFI 40th Annual Conference 2012, September 23, 2012 - September 26, 2012. 2012;1(June).

3. Alvidrez J, Weinstein RS. Early Teacher Perceptions and Later Student Academic Achievement. *J Educ Psychol*. 1999;91(4):731-746. doi:10.1037/0022-0663.91.4.731

4. Clark, C. M.; Peterson PL. *Teachers' Thought Processes. In M. C. Wittrock (Ed.)*. New York, NY: Macmillan; 1986.

5. Hoge RD. {P}sychometric {P}roperties of {T}eacher-{J}udgement {M}easures of {P}upil {A}ptitudes, {C}lassroom {B}ehaviros, and {A}chievement {L}evels. *J Spec*. 1983;17(4):401-429.

6. Hoge, R.; Butcher R. Analysis of teacher judgments of pupil achievement level. *J Educ Psychol*. 1984;26:777-781.

7. Hoge RD, Coladarci T. Teacher-Based Judgments of Academic Achievement: A Review of Literature. *Rev Educ Res*. 1989;59(3):297. doi:10.2307/1170184

8. Sharpley, C. F.; Edgar E. Teachers' ratings vs. standardized tests: An empirical investigation of assessment between two indices of achievement. *Psychol Sch*. 1986;23(`):106-111.

9. Demaray MK, Elliott SN. Teachers' judgments of students' academic functioning: A comparison of actual and predicted performances. *Sch Psychol Q*. 1998;13(1):8-24. doi:10.1037/h0088969

10. Südkamp, A.; Kaiser, J.; Möller J. Accuracy of Teachers' Judgments of Students' Academic Achievement: A Meta-Analysis. *J Educ Psychol*. 2012;104(3):743-762.

11. Bailey AL, Drummond K V. Who Is at Risk and Why? Teachers' Reasons for Concern and Their Understanding and Assessment of Early Literacy. *Educ Assess*. 2006;11(3-4):149-178. doi:10.1080/10627197.2006.9652988

12. Beswick, J. F.Willms, J. DouglasSloat EA. A Comparative Study of Teacher Ratings of Emergent Literacy Skills and Student Performance on a Standardized Measure. *Education*. 2005;126(1):116-137. http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=18360670&site=ehost-live.

13. Teisl JT, Mazzocco MMM, Myers GF. The utility of kindergarten teacher ratings for predicting low academic achievement in first grade. *J Learn Disabil*. 2001;34(3):286-293. doi:10.1177/002221940103400308

14. Shepard LA. *Classroom Assessment. In R. L. Brennan (Ed.)*. Westport, CT; 2006.

15. Stiggins R, Chappuis J. Using {Student}-{Involved} {Classroom} {Assessment} to {Close} {Achievement} {Gaps}. *Theory Pract*. 2005;44(1):11-18. doi:10.1207/s15430421tip4401_3

16. Martínez JF, Stecher B, Borko H. Classroom assessment practices, teacher judgments, and student achievement in mathematics: Evidence from the ECLS. *Educ Assess*. 2009;14(2):78-102. doi:10.1080/10627190903039429

17. Eckert, T. L.; Dunn, E. K.; Codding, R. S.; Begeny, J. C.; Kleinmann AE. Assessment of mathematics and reading performance: An examination of the correspondence between direct assessment of student

Eperformance and teacher report. *Psychol Sch*. 2006;43:247-265.

18. Meisels SJ, Bickel DD, Nicholson J, Xue Y, Atkins-Burnett S. Trusting teachers' judgments: A validity study of a curriculum-embedded performance assessment in kindergarten to grade 3. *Am Educ Res J*. 2001;38(1):73-95. doi:10.3102/00028312038001073

19. Brophy, J.; Good T. *Teacher Behavior and Student Achievement. In M. C. Wittrock (Ed.)*. New York, NY: McMillan.; 1986.

20. de Boer H, Bosker RJ, van der Werf MPC. Sustainability of Teacher Expectation Bias Effects on Long-Term Student Performance. *J Educ Psychol*. 2010;102(1):168-179. doi:10.1037/a0017289

21. Jussim, L.; Eccles JS. Teacher expectations II: Construction and reflection of student achievement. *J Pers Soc Psychol*. 1992;63:947-961.

22. Marsh HW. A multidimensional, hierarchical model of selfconcept: Theoretical and empirical justification. *Educ Psychol Rev*. 2:77-172.

23. Möller, J.; Pohlmann, B.; Köller, O.; Marsh HW. A metaanalytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Rev Educ Res*. 2009;79:1129-1167.

24. Black, P.; Wiliam D. Assessment and classroom learning. *Assess Educ*. 1998;5(1):7-71.

25. Brookhart MS. A theoretical framework for the role of classroom assessment in motivating student effort and achievement. *Appl Meas Educ*. 1997;10(2):161-180.

26. Rodriguez MC. The role of classroom assessment in student performance on TIMSS. *Appl Meas Educ*. 2004;17(1):1-24.

27. Anders, Y.; Brunner, M.; Krauss K. Diagnostic skills of mathematics teachers and the performance of their students. *Psychol Educ*. 2011;3(2):175-193.

28. Coladarci T. Accuracy of Teacher Judgments of Student Responses to Standardized Test Items. *J Educ Psychol*. 1986;78(2):141-146. doi:10.1037/0022-0663.78.2.141

29. Feinberg AB, Shapiro ES. Teacher Accuracy : An Examination of Teacher-Based Judgments of Students '. 2010;(October 2012):37-41.

30. Valdez A. Teacher Judgment of Reading Achievement: Cross-Sectional and Longitudinal Perspective. *J Educ Learn*. 2013;2(4):186-200. doi:10.5539/jel.v2n4p186

31. Rausch T, Karing C, Dörfler T, Artelt C. Personality similarity between teachers and their students influences teacher judgement of student achievement. *Educ Psychol*. 2016;36(5):863-878. doi:10.1002/zamm.19750551002

32. Mowrey, Sascha C.; Farran DC. Performance and Preparation: Alignment between student achivement, teacher ratings, and parent perceptions in urban middle-grades mathematics classrooms. *J Urban Learn Teach Res*. 2016;12:61-74.

33. Begeny JC, Eckert TL, Montarello SA, Storie MS. Teachers' Perceptions of Students' Reading Abilities: An Examination of the Relationship Between Teachers' Judgments and Students' Performance Across a Continuum of Rating Methods. *Sch Psychol Q*. 2008;23(1):43-55. doi:10.1037/1045-3830.23.1.43

34. Meissel K, Meyer F, Yao ES, Rubie-Davies CM. Subjectivity of teacher judgments: Exploring student

characteristics that influence teacher judgments of student ability. *Teach Teach Educ*. 2017;65:48-60. doi:10.1016/j.tate.2017.02.021

35. Gallant DJ, Moore JLIII. Ethnic-Based Equity in Teacher Judgment of Student Achievement on a Language and Literacy Curriculum-Embedded Performance Assessment for Children in Grade One. *Educ Found*. 2008;22(1/2):63-77.

36. Reeves D. *Elements of Grading*.; 2011.

37. Webster, R. E; Hewett, B.; Crumbacker HM. Criterion- related validity of the WRAT–R and K-TEA with teacher estimates of actual classroom academic performance. *Psychol Sch*. 1989;26:243-248.

38. Benner AD, Mistry RS. Congruence of mother and teacher educational expectations and low-income youth's academic competence. *J Educ Psychol*. 2007;99(1):140-153. doi:10.1037/0022-0663.99.1.140

39. Hinnant JB, O'Brien M, Ghazarian SR. The Longitudinal Relations of Teacher Expectations to Achievement in the Early School Years. *J Educ Psychol*. 2009;101(3):662-670. doi:10.1037/a0014306

40. Bates C, Nettelbeck T. Primary School Teachers' Judgements of Reading Achievement. *Educ Psychol*. 2001;21(2):177-187. doi:10.1080/01443410020043878

41. Begeny JC, Krouse HE, Brown KG, Mann CM. Teacher judgments of students' reading abilities across a continuum of rating methods and achievement measures. *School Psych Rev*. 2011;40(1):23-38. doi:10.1177/002076409704300306

42. McMahon S, Jones I. A comparative judgement approach to teacher assessment. *Assess Educ Princ Policy Pract*. 2015;22(3):368-389. doi:10.1080/0969594X.2014.978839

43. Zhu M, Urhahne D. Teachers' judgements of students' foreign-language achievement. *Eur J Psychol Educ*. 2015;30(1):21-39. doi:10.1007/s10212-014-0225-6

44. Eaves, R. C.; Campbell-Whatley, G.; Dunn, C.; Reilly, A. S.; Tate-Braxton C. Comparison of the Slosson Full-Range Intelligence Test and teacher judgments as predictors of students' academic achievement. *J Psychoeduc Assess*. 1994;12:381-392.

45. Bennett RE, Gottesman RL, Rock DA, Cerullo F. Influence of Behavior Perceptions and Gender on Teachers' Judgments of Students' Academic Skill. *J Educ Psychol*. 1993;85(2):347-356. doi:10.1037/0022-0663.85.2.347

46. Glanton CF. TEACHER-ASSIGNED MATH GRADES : CAN THEY ACCURATELY PREDICT STUDENT SUCCESS ON END-OF-YEAR MATH TESTS ? by Carla F . Glanton Jackson " Skot " Beazley , PhD , Faculty Mentor and Chair Lisa Reason , PhD , Committee Member Salha Jeizan , EdD , Committee Membe. 2014;(August).

47. Ready, D.D.; Wright DL. Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *Am Educ Res J*. 2011;48(2):335-360.

48. Assessment Brief: Understanding End-of-Grade Testing - Achievement Levels. Public Schools of North Carolina. http://www.ncpublicschools.org/docs/accountability/testing/eog/asb_achlev.pdf. Published 1999.

49. Achievement Level Descriptors for the EOG Mathematics Tests. Public Schools of North Carolina. http://www.ncpublicschools.org/docs/accountability/testing/eog/eogmathgr3to5ald.pdf. Published 2006.

50. STATE BOARD OF EDUCATION APPROVES HIGHER PROFICIENCY STANDARDS; NEW

ACHIEVEMENT LEVELS SET FOR EOG/EOC ASSESSMENTS. Public Schools of North Carolina. http://www.dpi.state.nc.us/newsroom/news/2013-14/20131003-01. Published 2013.

Appendix

Table A.1. Correlation and P-values between EOG achievement score and teacher judgment in Reading Comprehension for each grade level by cohorts

Grade \ Cohort	All	1	2	3
3	0.6058[a,&&]	0.6225[a,***]	0.6355[a,***]	0.6353[a,***,&&]
4	0.6257[a,&&]	0.6104[a,***,&&]	0.6273[a,&&]	0.6729[a,***]
5	0.6428[a,&&]	0.6282[a,***,&&]	0.6497[a,***,&&]	0.6668[a,***,&&]
6	0.6429[a,&&]	0.6350[a,***,&&]	0.647[a,*,&&]	0.6485[a,**,&&]
7	0.6421[a,&&]	0.6394[a,&&]	0.6428[a,&&]	0.6449[a,&&]
8	0.6124[a]	0.6200[a,***,&&]	0.6095[a,&&]	0.6129[a,&&]

[a] p-value<0.0001 related to $H_0 : r_i = 0$ for *i=A11,* and *cohort 1, 2, 3*

[***] p-value<0.001, [**] p-value<0.01, [*] p-value<0.05 related to $H_0 : r_{All} = r_c$ when cohort *c =1, 2, 3* (only illustrated in columns related to cohorts)

[&&] p-value<0.0001, [&] p-value<0.05 related to $H_0 : r_{\mathrm{Re}ading,i} = r_{Math,i}$ for *i=A11,* and *cohort 1, 2, 3*

Table A.2. Correlation and P-values between EOG achievement score and teacher judgment in Mathematics for each grade level by cohorts

Grade \ Cohort	All	1	2	3
3	0.6387[b]	0.6172[b,***]	0.637[b]	0.6648[b,***]
4	0.6541[b]	0.6382[b,***]	0.6566[b]	0.6746[b,***]
5	0.672[b]	0.6498[b,***]	0.6788[b,***]	0.6908[b,***]
6	0.6785[b]	0.6717[b,***]	0.6823[b,*]	0.6813[b]
7	0.6781[b]	0.6729[b,*]	0.6838[b,**]	0.6774[b]
8	0.6135[b]	0.6702[b,***]	0.6619[b,***]	0.6471[b,***]

[b] p-value<0.0001 related to $H_0 : r_i = 0$ for *i=A11,* and *cohort 1, 2, 3*

[***] p-value<0.001, [**] p-value<0.01, [*] p-value<0.05 related to $H_0 : r_{All} = r_c$ when cohort *c =1, 2, 3* (only illustrated in columns related to cohorts)

Table A.3. Correlation coefficients and P-values for significant differences between correlations of male and female students in mathematics for each grade level considering students' gender

	Reading			Math		
Grade	Female	Male	p-value[#]	Female	Male	p-value[#]
3	0.5938[*a]	0.6128[*b]	<0.0001	0.6311[*]	0.6465[*]	<0.0001
4	0.6172[*a]	0.6301[*b]	<0.0001	0.6451[*]	0.6627[*]	<0.0001
5	0.6348[*a]	0.6466[*b]	<0.0001	0.6630[*]	0.6806[*]	<0.0001
6	0.6372[*a]	0.6441[*b]	<0.0001	0.6705[*]	0.6862[*]	<0.0001
7	0.6416[*a]	0.6398[*b]	0.3792	0.6695[*]	0.6843[*]	<0.0001
8	0.6102[*a]	0.6094[*b]	0.7171	0.5981[*]	0.6278[*]	<0.0001

[*] p-value<0.0001 related to $H_0 : r_i = 0$ for *i=Reading* and *Math*

[a,b] p-value<0.0001 related to $H_0 : r_{\text{Reading},i} = r_{Math,i}$ when *i=Female*([a]), *Male*([b]) for Reading and corresponding value in Math

[#] p-value related to $H_0 : r_{i,Male} = r_{i,Female}$ for *i=Reading* and *Math*, P-value>0.05 was defined as insignificant and shaded.

Table A.4. Correlation and P-values between EOG achievement score and teacher judgment in Reading Comprehension for each grade level by ethnicity

Grade \ ethnic	Asian	Black	Hispanic	American-Indian	Multi-Racial	White
3	0.6104[a,***]	0.5348[a,***]	0.5606[a,***]	0.5523[a]	0.5817[a,***]	0.5988[a,***]
4	0.6429[a,*]	0.5627[a,***]	0.5812[a,***]	0.5835[a]	0.6060[a,***]	0.6125[a,***]
5	0.6753[a,***]	0.5717[a,***]	0.6044[a,***]	0.5788[a,*]	0.6223[a,***]	0.6277[a,***]
6	0.6999[a]	0.5632[a,***]	0.5976[a,***]	0.5742[a,**]	0.6211[a,***]	0.6255[a,***]
7	0.7148[a]	0.5679[a,***]	0.5938[a,***]	0.5625[a,**]	0.6098[a,***]	0.6238[a,***]
8	0.6831[a,**]	0.5343[a,***]	0.5732[a,***]	0.5645[a]	0.5653[a,*]	0.5915[a,***]

[a] p-value<0.0001 related to $H_0 : r_i = 0$ for *i=A, B, H, I, M_R, W*

[#] p-value related to chi-square test of $H_0 : r_{i,A} = r_{i,B} = r_{i,H} = r_{i,I} = r_{i,M_R} = r_{i,W}$ for *i=Reading*

*** p-value<0.001, ** p-value<0.01, * p-value<0.05 related to $H_0 : r_{\text{Reading},i} = r_{Math,i}$ when i=A, B, H, I, MR, *W*

Table A.5. Correlation and P-values between EOG achievement score and teacher judgment in Mathematics for each grade level by ethnicity

Grade \ ethnic	Asian	Black	Hispanic	American-Indian	Multi-Racial	White
3	0.6511[b,***]	0.5737[b,***]	0.5876[b,***]	0.5735[b]	0.6168[b,***]	0.6289[b,***]
4	0.6647[b,*]	0.5872[b,***]	0.6008[b,***]	0.5871[b]	0.6384[b,***]	0.6462[b,***]
5	0.7092[b,***]	0.6058[b,***]	0.6254[b,***]	0.6054[b,*]	0.6530[b,***]	0.6656[b,***]
6	0.7095[b]	0.5967[b,***]	0.6326[b,***]	0.6121[b,**]	0.6560[b,***]	0.6700[b,***]
7	0.7225[b]	0.5991[b,***]	0.6243[b,***]	0.6026[b,**]	0.6493[b,***]	0.6749[b,***]

8	0.6523 [b,**]	0.5225 [b,***]	0.5559 [b,***]	0.5619 [b]	0.5851 [b,*]	0.6101 [b,***]

[b] p-value<0.0001 related to $H_0 : r_i = 0$ for *i=A, B, H, I, M_R, W*

[#] p-value related to chi-square test of $H_0 : r_{i,A} = r_{i,B} = r_{i,H} = r_{i,I} = r_{i,M_R} = r_{i,W}$ for *i=Math*

[***] p-value<0.001, [**] p-value<0.01, [*] p-value<0.05 related to $H_0 : r_{\text{Reading},i} = r_{Math,i}$ when *i=A, B, H, I, M_R, W*

Table A.6. Result for Multiple Regression Models for students EOG performance as a function of teacher judgment, and student's demographics in reading and mathematics

ΔR-squared*							
Reading	0.4817						
Mathematics	0. 4981						
Intercept							
Reading	0.9481***						
Mathematics	0.4187***						
Teacher Judgment							
Reading	0.4625***						
Mathematics	0.4739***						
TJ*TJ							
Reading	0.0237***						
Mathematics	0.0546***						
Sex	M				F (Base)		
Reading	-0.0388***				0		
Mathematics	0.0943***				0		
TJ*Sex	TJ*M				TJ*F (Base)		
Reading	-0.00164				0		
Mathematics	0.004003**				0		
Ethnicity	A	B	H	I	M_R	P	W (Base)
Reading	-0.3418	-0.157***	-0.2029***	-0.2268***	-0.0963***	-0.3415*	0
Mathematics	0.0362*	-0.2864***	-0.1952***	-0.2916***	-0.1819***	-0.3427**	0
TJ*Ethnicity	TJ*A	TJ *B	TJ *H	TJ *I	TJ *M_R	TJ *P	TJ *W (Base)
Reading	0.1038***	-0.0371***	0.0014	-0.0226***	0.0026	0.0624	0
Mathematics	0.0546***	-0.0289***	-0.0173***	-0.0257***	0.0070*	0.0521	0
Sex*Ethnicity	M*A	M*B	M*H	M*I	M*M_R	M*P	M*W (Base)
Reading	0.0042	-0.0429***	0.0019	-0.0059	-0.015*	-0.0222	0
Mathematics	-0.0036	-0.0335***	0.0067*	0.002	-0.0026	-0.0450	0

Grade		3	4	5	6	7	8 (Base)
Reading		1.5368***	1.3074***	0.8696***	0.7085***	0.42***	0
Mathematics		5.8619***	4.6561***	3.4095***	2.2364***	0.9519***	0
TJ *Grade		TJ *3	TJ *4	TJ *5	TJ *6	TJ *7	TJ *8 (Base)
Reading		0.1945***	0.1393***	0.0929***	0.1099***	0.1455***	0
Mathematics		0.0034	0.0099*	0.0123**	0.0119***	0.0386***	0
Sex*Grade		M*3	M*4	M*5	M*6	M*7	M*8 & F*3-8 (Base)
Reading		-0.0262**	-0.0117	-0.0307***	-0.0171**	0.0015	0
Mathematics		0.0076	0.0021	0.0129*	0.021***	-0.0004	0
Ethnicity*Grade		*3	*4	*5	*6	*7	*8 (base)
Reading	Asian	0.0228	0.0261	0.0205	0.0283	0.0109	0
	Black	-0.0002	0.0256**	0.0612***	0.0198**	-0.0233***	0
	Hispanic	-0.0252*	0.0231*	0.0459***	0.0361***	0.0053	0
	American Indian	-0.0426	-0.0125	-0.0314	-0.0060	-0.0582**	0
	Multi-Racial	-0.0658**	-0.0661**	-0.0461**	-0.0418**	-0.0277*	0
	White	0	0	0	0	0	0
Math	Asian	-0.0063	0.0079	0.0207	-0.0129	0.0285	0
	Black	-0.0142	-0.0131	0.0088	-0.0192**	0.0292***	0
	Hispanic	0.0006	0.0099	0.0146	-0.0246**	0.0179**	0
	American Indian	0.1013**	0.0242	-0.0437*	0.0312	0.1003***	0
	Multi-Racial	0.0407*	0.0226	0.0281*	-0.0025	0.0178	0
	White	0	0	0	0	0	0

Year		2006	2007	2008	2009	2010	2011	2012	2013
Reading		-1.1275***	-0.8512***	-1.7196***	-1.2103***	-0.7793***	-0.4091***	0	0
Mathematics		-7.234***	-6.0202***	-4.7563***	-3.5467***	-2.3576***	-1.2115***	0	0
TJ *Year		TJ *2006	TJ *2007	TJ *2008	TJ *2009	TJ *2010	TJ *2011	TJ *2012	TJ *2013
Reading		-0.2066***	-0.1927***	0.0057	-0.0023	-0.0146**	-0.0198***	-0.0234***	0
Mathematics		-0.0990***	-0.0904***	-0.0909***	-0.0837***	-0.079***	-0.0777***	-0.0833***	0
Sex*Year		M*2006	M*2007	M*2008	M*2009	M*2010	M*2011	M*2012	M*2013 & F*Years (Base)
Reading		0.035**	0.023*	0.048***	0.041***	0.068***	0.066***	0.055***	0
Mathematics		-0.055***	-0.058***	-0.068***	-0.080***	-0.085***	-0.088***	-0.096***	0
Ethnicity*Year		*2006	*2007	*2008	*2009	*2010	*2011	*2012	*2013
Reading	Asian	-0.0649*	-0.0604*	-0.0958**	-0.07352**	-0.0925***	-0.135***	-0.1057***	0

	Black	0.0354**	-0.0079	-0.1599***	-0.1282***	-0.0945***	-0.0728***	-0.0548***	0
	Hispanic	-0.052**	-0.0216	-0.1985***	-0.1652***	-0.1511***	-0.0774***	-0.0803***	0
	American Indian	0.0809*	0.0530	-0.1181**	-0.0977*	-0.0344	-0.0428	-0.0318	0
	Multi-Racial	0.1053***	0.0751**	0.0106	0.0225	0.0255	0.0223	-0.0027	0
	White	0	0	0	0	0	0	0	0
Math	Asian	-0.1067**	-0.1415***	-0.1529***	-0.1644***	-0.1774***	-0.1936***	-0.1777***	0
	Black	0.0371**	0.0402***	0.0458***	0.0737***	0.094***	0.1104***	0.1171***	0
	Hispanic	0.0604***	0.0488***	0.0471***	0.0609***	0.0744***	0.1148***	0.1013***	0
	American Indian	-0.011	-0.0378	0.0252	0.0632*	0.0995***	0.1123***	0.1017***	0
	Multi-Racial	0.0122	0.0047	0.0168	0.0380*	0.0493**	0.0592***	0.0683***	0
	White	0	0	0	0	0	0	0	0

Cohort	1: (2006-2011)	2: (2007-2012)	3: (2008-2013) (Base)
Reading	0.7017***	0.3688***	0
Mathematics	2.3171*	1.1678***	0

***p-value<0.0001, **p-value<0.005, *p-value<0.05

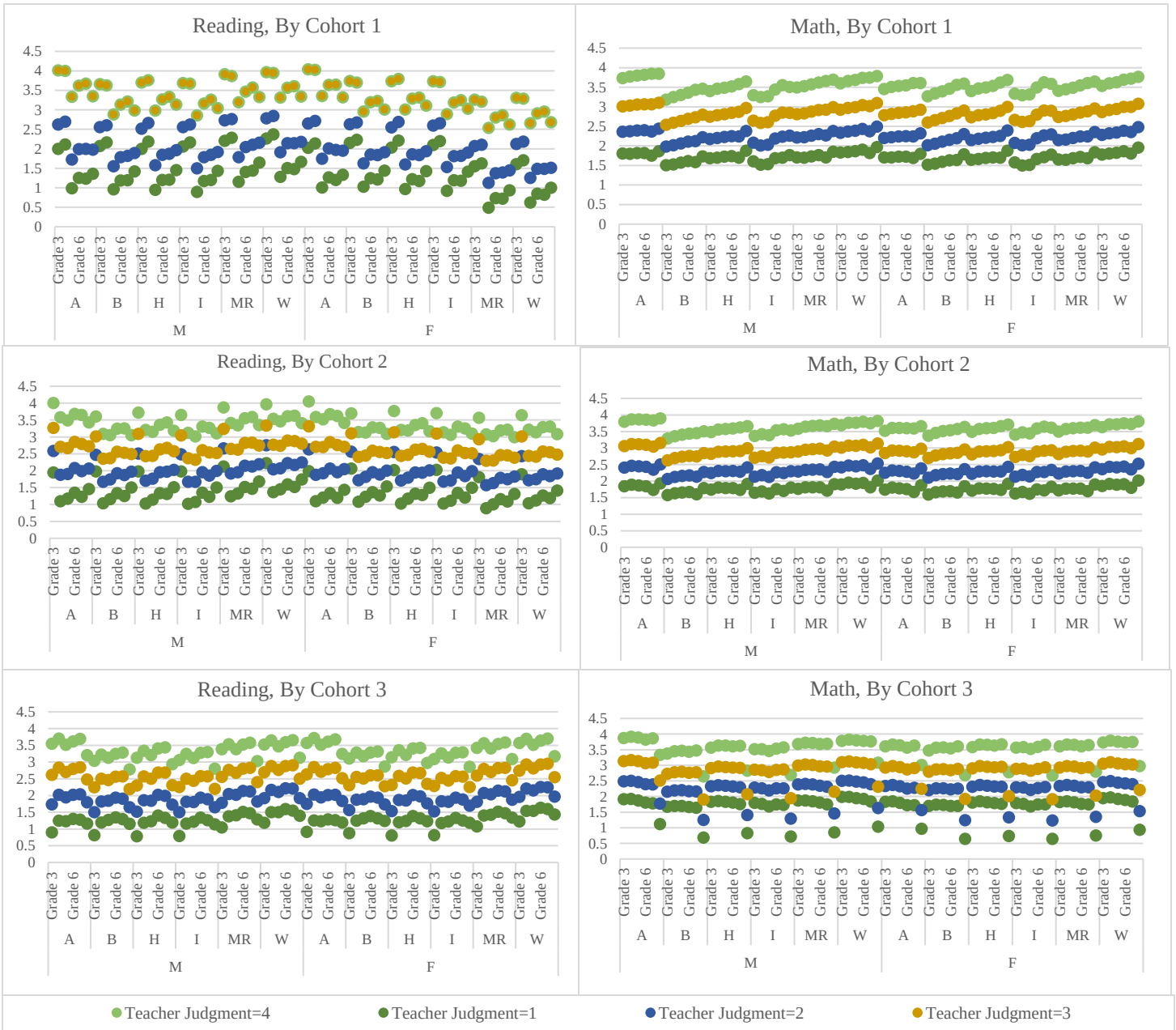

Figure A.1. Predicted students' achievement score, using the regression model when teacher judgment is level 1, 2, 3, and 4 in mathematics and reading for each cohort based on students' gender, and ethnicity, and grade level.

Sareh Meshkinfam

Sareh Meshkinfam is a Ph.D. student in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. She earned her undergraduate and master degrees from Sharif University of Technology, Iran, where she focused on production planning and published multiple papers in this regard. At North Carolina State University, she was awarded the prestigious Fitts Fellowship and is working as a research assistant in the healthcare system group under the guidance of Prof. Ivy. Her current research involves analyzing K-12 students' performance using statistical and optimization models.

Julie S. Ivy

Dr. Julie Ivy is a Professor at North Carolina State University in the Edward P. Fitts Department of Industrial and Systems Engineering and a Fitts Faculty Fellow in Health Systems Engineering. She previously spent several years on the faculty of the Stephen M. Ross School of Business at the University of Michigan. Dr. Ivy is actively involved in INFORMS, she is the current president of the Minority Issues Forum and is a past president of the Health Applications Section of INFORMS. She has co-authored more than forty journal articles, working papers, and conference proceedings.

Amy C. Reamer

Dr. Amy Reamer is the director of the North Carolina State Engineering 2+2 Transfer Program at UNC Wilmington and an adjunct teaching assistant professor in the NC State Edward P. Fitts Department of Industrial and Systems Engineering. She earned a Bachelor of Science degree in Industrial Engineering, a Master of Industrial Engineering and a Ph.D. in Industrial and Systems Engineering from NC State University. Previously, she worked as a cost engineer at Lenovo in Research Triangle Park. Currently, she continues her dissertation research in the application of Markovian models to the K-12 education system.