Paper #3695-2019

SAS® Analytics for Characterizing the Relationship between Teacher Judgment and Student Performance

Sareh Meshkinfam, NC State University; Julie S. Ivy, NC State University;
Amy C. Reamer, NC State University

Abstract

A teacher's assessment or judgment of an individual student's performance can impact other educators' expectations of that student's ability as well as the student's future academic placement. Exploring the relationship between teacher judgment and student performance in primary education is critical, as early barriers can evolve into significant academic hurdles for individual students at the middle school and high school levels. Correlation analysis and regression models have been used to analyze the longitudinal and cross-sectional relationship between students' achievements and teacher judgment in reading and mathematics across grades and years, considering students' demographics. SAS® procedures (such as PROC CORR and PROC MIXED) were used to explore a large data set from the North Carolina Education Research Data Center (NCERDC), which includes 6,511,741 students in 3rd to 8th grades from 2006 to 2013. The data set includes information such as students' End-Of-Grade (EOG) test scores, demographic characteristics, and evidence of their academic performance in each grade and year. SAS provides an effective tool to explore this data set with accessible and easy-to-use analysis approaches. Results demonstrate moderate to high correlations, which are significantly higher for male and significantly lower for minority ethnic groups. The regression models reveal that students' gender, ethnicity, and previous grade performance significantly affect their EOG achievement score.

Introduction

Teacher judgment provides a measure of a teacher's perception about an individual student's academic ability, a student's anticipation of his or her own competence, and future academic placement (Südkamp, Kaiser, & Möller, 2012; Möller, Pohlmann, Köller, & Marsh, 2009; Rodriguez, 2004). Understanding the relationship between teacher judgment and actual student performance is critical. In general, teacher judgment is defined as a teacher's assessment of expected student academic performance. From the 1970s, research studies have used different statistical methods such as correlation, ANOVA, regression or hierarchical linear models to evaluate the strength of the relationship between teacher judgment and student performance and its dependency on student characteristics (e.g., gender, ethnicity, socioeconomic status, academic level, and parent education) (Coladarci, 1986; Beswick, Willms, & Sloat, 2005; Demaray & Elliott, 1998; Martínez, Borko, & Stecher, 2009; Valdez, 2013; Mowrey & Farran, 2016; Rausch, Karing, Dörfler, & Artelt, 2016). SAS has been used to study educational performance in multiple domains. Many institutes who use SAS for storing and/or analyzing the educational data, like North Carolina Educational Research Data Center (NCERDC) is an example of such an institute. SAS has also been used to develop and operationalize the Education Value-Added Assessment System (EVAAS) which is an educational resource for educators and policymakers for tracking student performance, their test scores, and their expected future performance over time. EVVAS offers tools for educators to improve student learning in North Carolina. The SAS

environment facilitates more reliable and precise estimates for larger number of students (SAS, 2014).

In this research, we utilize data from NCERDC, which includes information on more than 500,000 students per year in 3rd to 8th grades from 1993 to 2015. This study explores the relationship between student performance on standardized End-Of-Grade (EOG) tests and teacher judgment in reading comprehension (referred to hereafter as "reading") and mathematics longitudinally for 3rd to 8th grade from 2006 to 2013. Statistical analyses on correlations and their significance have been implemented to track the relationship over grade levels and years and based on student's gender and ethnicity. Regression and hierarchical models are developed to identify the relationship between student EOG test performance at each grade level in reading and mathematics with current and previous year teacher judgments, historical EOG test performance and student demographics. These analyses are designed to answer the following research questions: 1) What is the relationship between student EOG test performance and teacher judgment? 2) What is the influence of student demographics, grade level and school on this relationship over time?

Methods

Data overview

The NCERDC data set includes information on the state's public schools, teachers, and students collected by the North Carolina Department of Public Instruction (NC DPI) from 1993 to 2015. The NCERDC concatenates school-level data from NCDPI for each test and each grade, to include the student's test records, the usage of any testing modifications, the student's class participation, use of technology, the student's demographic data such as sex, exceptionality and ethnicity status, and the teacher's assessment of student expected performance. The state-mandated, curriculum-based North Carolina EOG tests are multiple choice tests in reading and mathematics that evaluate student's performance at the end of an academic year to see if he or she meets grade-level expectations. Initial raw test scores are scaled and converted to annually defined achievement levels I-IV (I and II defined as "non-proficient", and III and IV as "proficient") (Carolina, Assessment Brief: Understanding End-of-Grade Testing - Achievement Levels, 1999; Carolina, Achievement Level Descriptors for the EOG Mathematics Tests, 2006). Teachers are asked to predict each student's achievement level score for the academic year (a rating of I, II, III or IV), and those are recorded in the NCERDC data set.

To extract information for the analyses from NCERDC data set, SAS DATA steps are used. These DATA steps help to filter unwanted data, to eliminate missing observations in the data, and to merge fields and data sets (in combination with other procedures like PROC SORT) for the purposes of this study. To ensure that the sample size does not influence the significance of the results, SAS PROC SURVEYSELECT is combined with DATA steps to create 1% and 10% samples from each grade level and year data set. SAS PROC MEANS and PROC FREQ provide the statistical information about the distribution of the data set, such as the percentage of students that identify as a particular gender or ethnicity. Table 1 and Table 2 summarize the distribution of students and the proportion of data used in each grade level and year. The variable names provided for ethnicity were defined by NCERDC.

Statistical Analysis

In this study, various statistical methods are employed to analyze the relationship between students' EOG achievement scores in mathematics and reading, with their corresponding teacher judgment assessment from 2006-2013. To assess the strength of this relationship, Pearson correlation coefficients are calculated via PROC CORR for each grade level and year,

and also based on a student's gender and ethnicity. Hypothesis tests evaluate the significance of the differences among correlation, and calculated effect sizes assist in monitoring the influence of the sample size on the significance. SAS PROC EXPORT and ODS destinations are used to create Excel and pdf files from which we extract output values for simple calculations and statistical analyses. Hypothesis tests on the significance of the differences between correlations (i.e., reading vs. mathematics, male vs. female, ethnicity groups, grade levels and years) are used via Z-statistics on Fisher transformations of correlation (Zaiontz, 2017). To account for possible influence size of the population can have on the results, the similar analyses are done on 1% sample[1], which statistically has a similar distribution with population it is selected from.

Total Number									
Year / Grade	2006	2007	2008	2009	2010	2011	2012	2013	Grade Total
3	104,808	112,280	135,755	168,638	169,835	163,799	160,699	106,518	1,122,332 (0.736) [a]
4	102,831	108,971	112,250	159,688	159,679	160,581	155,666	114,669	1,074,335 (0.79)
5	103,615	107,263	132,679	160,177	160,409	160,549	161,244	114,435	1,100,371 (0.746)
6	106,772	108,594	108,508	154,715	152,570	154,070	156,317	116,314	1,057,860 (0.8)
7	106,774	110,995	109,920	157,845	157,840	158,474	159,886	115,381	1,077,115 (0.771)
8	107,968	110,348	133,736	158,065	151,732	151,945	152,990	112,944	1,079,728 (0.762)
Year Total	632,768 (0.982)	658,451 (0.953)	732,848 (0.869)	959,128 (0.68)	952,065 (0.69)	949,418 (0.7)	946,802 (0.703)	680,261 (0.957)	**6,511,741 (0.794)**

[a] Values in parentheses demonstrate the proportion of the data which is used after eliminating missing elements in each grade level and year.

Table 1. Number of students in each grade and year for NCERDC (Data from 2006 - 2013)

Gender		**Ethnicity**					
Male	Female	White	Asian	Black	Hispanic	American Indian	Multi-Racial
51.26%	48.74%	51.25%	2.21%	28.92%	12.38%	1.62%	3.62%

Table 2. Distribution of students' demographics for NCERDC (Data from 2006 – 2013)

To analyze how student EOG test scores are affected by teacher judgment and student's demographics, regression and hierarchical linear models are employed. We selected from PROC GLM, PROC GLMSELECT, and PROC MIXED in SAS to determine the best method for

[1] PROC SURVEYSELECT were used to randomly pick 1% of each data set from 3rd till 8th grade in 2006 to 2013, separately.

modeling the data, considering the structure of the data set. PROC MIXED allowed us to easily and efficiently apply 2-level hierarchical linear models (HLM) and multiple regression models, proving to be the most advantageous for our use. Thus, to remain consistent in comparing results from different models (Multiple regression and Hierarchical), PROC MIXED has been used for all modeling calculations. For each modeling approach, a separate model is defined for each grade level of 3rd to 8th in reading and mathematics. In both sets of models, the teacher judgment score of the corresponding grade and subject, and previous grades' teacher judgment and historical EOG achievement level scores in both reading and mathematics are considered. Moreover, categorical variables representing gender, ethnicity, corresponding year for each specific grade (grade-year), cohort and school district (represented the Local Education Agency [LEA]) are included. All possible interactions between these variables are also included in the models.

A Multiple Regression Model (Model1) considering the aforementioned predicting variables was applied first. Next, Two-level HLMs are being used because student data from the NCERDC can be considered at two levels within the system hierarchy: students within schools and students by school. Level-1 outcome (related to students) can be examined as a function of level-2 (related to schools) predictor variables. The school identifier is defined by combining the school code and associated LEA to have a unique element pointing to a specific school. Modeling starts by fitting the unconditional model to examine variation of students' achievement level scores across schools. Then, the model has been updated to examine the effects of school level (level-2) and student level (level-1) predictors, by adding related variables in one model together.

Next, we considered the school identifier as a predictor variable in the models instead of as a grouping factor. Thus, the similar modeling structure for variables are used to design new multiple regression models (Model2) with an extra school identifier variable. The general Model2 formulation is presented as follows where a_{igjk} and p_{igjk} are EOG achievement level scores, and teacher judgment score of kth student in subject i, for school j and grade level g, respectively. Similarly, y_g, I_g and l_g specify the year a student studies in grade level g and the school identifier and the LEA code for gth grade, respectively. Students' gender, ethnicity and cohort are represented by s, e and c, respectively, γ is coefficient multiplier, and r_{ijk} is random error.

$$
\begin{aligned}
a_{igjk} &= \gamma_{i00} + \gamma_{i1} p_{igjk} + \gamma_{i2} s + \gamma_{i3} e + \gamma_{i4} c + \gamma_{i5} y_g + \gamma_{i6} l_g + + \gamma_{i7} I_g \\
&+ \sum_i \sum_{k' \in All\ grade < k} \gamma_{ik'0} a_{ijk'} + \sum_i \sum_{k' \in All\ grade < k} \gamma_{ik'0} p_{ijk'} \\
&+ \sum_i \sum_{k' \in All\ grade < k} \gamma_{ik'0} a_{ijk'} * y_{k'} + \sum_i \sum_{k' \in All\ grade < k} \gamma_{ik'0} p_{ijk'} * y_{k'} \\
&+ \sum_i \sum_{k' \in All\ grade < k} \gamma_{ik'0} s * y_{k'} + \sum_i \sum_{k' \in All\ grade < k} \gamma_{ik'0} e * y_{k'} + \gamma_{il} e * s \\
&+ \sum_{k' \in All\ grade \le k} \gamma_{ik'} y_{k'} + \gamma_{in0} p_{i \cdot k} * y_k + r_{ijk}
\end{aligned}
$$

In each model, significant factors can be identified by analyzing the related p-values for t-tests in the solution for fixed effects of each variable. Since the large sample size can affect the factors' significances, the Model2 structure has been implemented on a 10% sample[2] of the data set. By considering stepwise selection, PROC GLMSELECT is used to check whether

[2] For generating 10% sample via PROC SERVEYSELECT, initially a random sample of 10% from each 3rd grade data in 2006, 2007, and 2008 were selected separately, and then these students have been tracked during their trajectory to 8th grade in 2011, 2012, 2013 using DATA steps.

each predictor variable is significant enough to keep that in the model or if it can be removed from both the entire data set and the 10% sample. Finally, Model3 is created by only considering selected effects from PROC GLMSELECT of the three cohorts in PROC MIXED. Criterion based likelihood of Akaike information criterion (AIC), and Bayesian information criterion (BIC), are considered for selecting the model, where the model with the lower value will be selected. Finally, we compared the execution time and fit statistics among the different models to see which formulation provided a better prediction for the NCERDC data set.

Results

Different subsets of the NCERDC data set are used for different statistical analysis: (i) population data of 3rd to 8th grade from 2006 to 2013 and (ii) three longitudinal cohorts of students from 3rd to 8th grade as cohort 1 from 2006 to 2011, cohort 2 from 2007 to 2012 and cohort 3 from 2008 to 2013. For these analyses, SAS 9.4 has been used on a 64-bit operating system Intel(R) Core(TM) i7-2600 CPU @ 3.40 GHz and 16.0 GB RAM via windows 10.

The results of the correlation and hypothesis analyses for EOG test and teacher judgment scores for each grade level and year and for each subset of the NCERDC data can be found in Meshkinfam et al. (2019a) and Meshkinfam et al. (2019b), respectively (Meshkinfam, Ivy, & Reamer, 2019a; Meshkinfam, Ivy, & Reamer, 2019b).

For example, Figure 1 and Figure 2 show the evolution of the correlation between students' EOG achievement level scores and teacher judgment in each grade level for each academic year, and cohort, respectively. The Pearson correlation coefficient for the relationship between teacher judgment and student EOG test scores in each grade and year for the population (i.e., all students), in general, are significant (different from 0) and range from 0.58 to 0.71. The correlation is significantly stronger for mathematics than reading across all grade levels and years. Correlation coefficients decrease from 3rd to 8th grade in reading, although they are relatively constant in mathematics. Within each year from 2006-2013, the correlation decreases for higher grade levels over time; there is a larger correlation for more recent years. Similarly, the correlation coefficients in Figure 2 demonstrate an increasing trend by grade (except from 7th to 8th grade), that is not detectable for reading comprehension. Generally, it can be noticed that correlation values are higher for cohort 3 (the most recent cohort), and lower for cohort 1 in most of the grade levels for both mathematics and reading.

To model the relationship between student's EOG test scores and their corresponding teacher judgment scores, historical performance, gender, ethnicity, and grade level-year, multiple regression and multilevel hierarchical linear models have been used. To explore the effect of student's EOG test performance and teacher judgment in previous grade levels on the student's EOG score in the current year, only samples of students in the three cohorts are used. For each cohort, only students who spent exactly one year in each grade level (not retained or skipped) are considered. For this purpose, students are tracked across grade levels and years via their unique master identification number. To consider students' school effect, since each student may have been in at most six different schools during his/her studies from 3rd to 8th grade, only the school identifier of the corresponding year is used to group students for each grade level in the model as we model the student performance in the current grade.

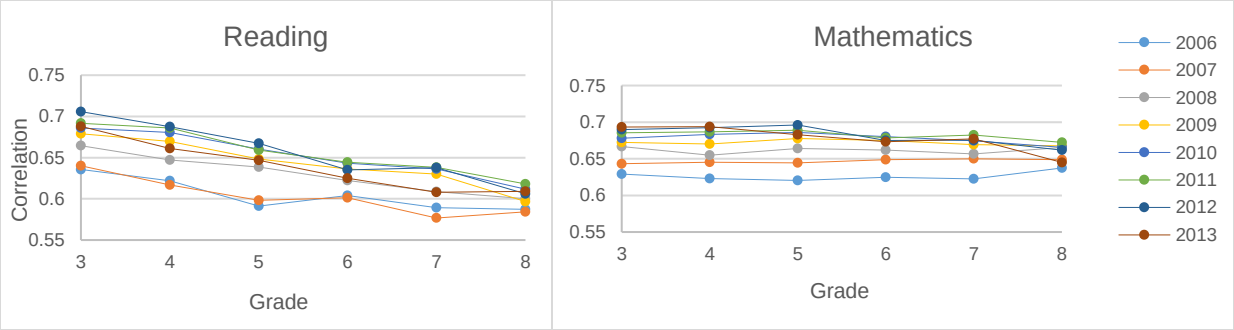

Figure 1. Correlation between EOG achievement score and teacher judgment in mathematics and reading comprehension by grade level over time (Meshkinfam, Ivy, & Reamer, 2019b)

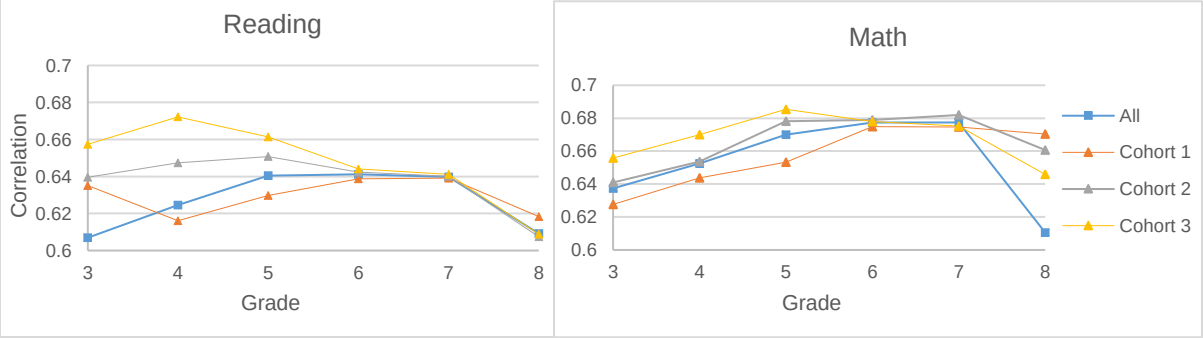

Figure 2. Correlation between EOG achievement score and teacher judgment in mathematics and reading comprehension by grade level for cohorts (Meshkinfam, Ivy, & Reamer, 2019a)

Despite the benefits of using HLM, the processing time is increased (more than 3 hours for each model) in comparison with the multiple regression models 1 to 3 (less than five minutes each). The estimates of the predictor variables are nearly the same across all models, with similar levels of significance. The values of the Fit statistics (either AIC, BIC or -2log likelihood) are smaller for multiple regression Model2 than for the HLM, which shows that these models are a better fit. Table 3 demonstrates the corresponding values for the -2log likelihood in each model. The models are able to predict the achievement scores better for the higher grade levels in both reading and mathematics as the fit statistics become smaller in those models. Although fit statistics for Model3 are larger than related values in HLM and Model2, these models show significant effects that are important for this analysis. Based on these results, school identifier is an important factor that can affect students' performance, along with the students' cohort, their previous grades' scores and teacher judgments interaction with year and ethnicity. IN addition, PROC GLMSELECT applied to the 10% sample point found has similar importance for school identifier effects, which suggests that population size has not influenced these results.

The statistical results of implementing Model2 from 3rd to 8th grades in mathematics and reading using the three cohorts of students are shown in Table 4. We use the results for the unconditional models and the models with predictors to assess how much of the variability can be explained by these models. The covariance parameter estimates for the residual in the unconditional models ($\hat{\sigma}_0^2$), and final models ($\hat{\sigma}_{final}^2$) are shown in Table 4. Adding the predictors to the original unconditional models explained over 60% of the explainable variation (when this proportion can be calculated via $\Pr op. = \frac{\hat{\sigma}_0^2 - \hat{\sigma}_{Final}^2}{\hat{\sigma}_0^2} * 100$) for each grade

level in reading and mathematics (ranged from 50.21% to 75.67%), for this purpose the proportion is similar to R^2.

	Reading				Mathematics			
Grade	**HLM[a]**	**Model1[b]**	**Model2[c]**	**Model3[d]**	**HLM**	**Model1**	**Model2**	**Model3**
3	631187.4	647458.4	630263.9	636524.1	568042.4	591267.9	568242.6	574778.9
4	501008.5	506836.3	498542.3	505556.1	438234.9	452144.9	437916.4	445096.6
5	449426.3	453931.5	446648.9	453840.2	368184.5	381786.8	367001.1	374551.0
6	419523.0	422071.6	417891.4	422402.7	336207.9	347827.7	335894.1	340584.7
7	409297.1	411009.8	408018.9	412520.4	325955.7	333981.6	325639.3	330408.7
8	336255.8	337916.2	335066.1	339858.3	341229.6	352848.1	343187.6	348147.6

[a] Hierarchical models (HLM) considered the school identifier variable for grouping students and creating the hierarchical level.

[b] Multiple regression models (Model1) applied the regression without grouping students and considering any hierarchy level.

[c] Multiple regression models (Model2) implemented the school identifier variable as categorical predictor variables instead to evaluate the degree and significance of schools in a different way for prediction of achievement scores.

[d] Multiple regression models (Model3) applying the Model2 only on the most significant factor from stepwise selection.

Table 3. Comparison Fit statistics in Hierarchical and multiple regression models in reading and mathematics for each grade level

	Reading			Math		
Grade	$\hat{\sigma}_0^2$	$\hat{\sigma}_{final}^2$	*Prop.*	$\hat{\sigma}_0^2$	$\hat{\sigma}_{final}^2$	*Prop.*
3	0.9511	0.4169	56.17	0.6903	0.3437	50.21
4	0.9625	0.3155	67.22	0.7243	0.257	64.52
5	0.8625	0.2932	66.01	0.6846	0.22	67.86
6	0.9424	0.2952	68.68	0.6932	0.2149	69.00
7	1.0297	0.3064	70.24	0.725	0.2194	69.74
8	0.7567	0.2414	68.1	1.0268	0.2498	75.67

Table 4. Comparison of covariance, to see the explainable proportion by implementing multiple regression models with school identifier as predictor variable in reading and mathematics for each grade level

Due to the size of the Model2 results, only the most significant results are summarized. In all models the intercepts are insignificant. In each grade level model for mathematics and reading, teacher judgment for the current grade level is a significant factor where the absolute value for the estimates decreases over grade from 3rd to 8th. Similarly, students' previous grade EOG scores and historical teacher judgments are highly significant. The

coefficient estimates become smaller for the oldest information, and for the scores not associated with the subject that the model is predicting. These estimates are decreasing for higher grade level models in which more factors are considered. Students' gender is more significant than their ethnicity (p-value of less than 0.001 in most case for gender in comparison with p-values around 0.05 for ethnicity), while the estimate for students' ethnicity is bigger. The grade-year and the cohort students belong are significant for predicting EOG score, where the older cohort has a larger effect on the score. The previous grade-year are also significant. Specifically, the recent grade-year have bigger estimate values. Moreover, the interaction between students' current and previous grade scores are highly significant (p-value<0.001). After applying Model2 on 10% sample, we found that teacher judgment for current grade, previous grade EOG score, and teacher judgments (specifically the most recent one which is in the same subject as the EOG score being predicted) along with year and cohort factors are the most significant effects. These results are similar to the PROC GLMSELECT results using three cohorts.

The residuals can be calculated based on Model2 to assess the relationship between students' actual EOG achievement level score with the predicted EOG score based on teacher judgment, students' demographics and historical academic performance in Model2. For this purpose, ODS OUTPUT in combination with DATA steps to filter the data are implemented. To analyze the models for specific group of students, PROC SGPLOT is applied to visualize the residuals for each gender-ethnicity group like Female-Asian, Male-Asian and so on. Different colors have been used to differentiate the residuals based on student's actual EOG score. PROC SORT is used to order the residuals based on their actual EOG levels to identify patterns in plots. Figure 3 to Figure 5 show the residuals in mathematics and reading based on students' gender, ethnicity, and grade level for each cohort, as a function of the actual EOG test score where levels 1 through 4 are color-coded. The density of residual plot for each gender-ethnicity combination is associated with the population size for that group. Moreover, it seems like there are more extreme residual points for White and Black students (either positive or negative). The height of each plot corresponds to their variability, which is bigger for White, Black and Hispanic students, which have larger populations. The residuals ranged from -2 to +2, and they are larger (smaller) for higher (lower) achievement scores - generally more negative for a lower achievement scores, and more positive for higher ones. As the results are sorted based on the students actual EOG score, in each plot we can see almost the same color for each specific level. However in the 8^{th} grade mathematics in cohort 1 and cohort 2, which have different colors because no Female Asian students achieved EOG score of 1 or 2. There is a pattern in residuals for EOG levels in all other grade and cohort models which suggests there may be other factors that was not considered in the models, that can help us to identify this over/underprediction by EOG scores. It can be noticed that range of the residuals are increasing and becoming more positive for lower EOG scores (1, 2 and even 3), to have a more balanced residual plots around 0 along 3^{rd} to 8^{th} grade in each cohort for mathematics and reading. Specifically, it can be seen that these residuals are negative in 3^{rd} grade reading for cohort 1 and 2. While the residuals become more spread out for lower EOG scores in reading from cohort 1 to cohort 3, the widths of their parts in plots seem to be narrower in mathematics across cohorts.

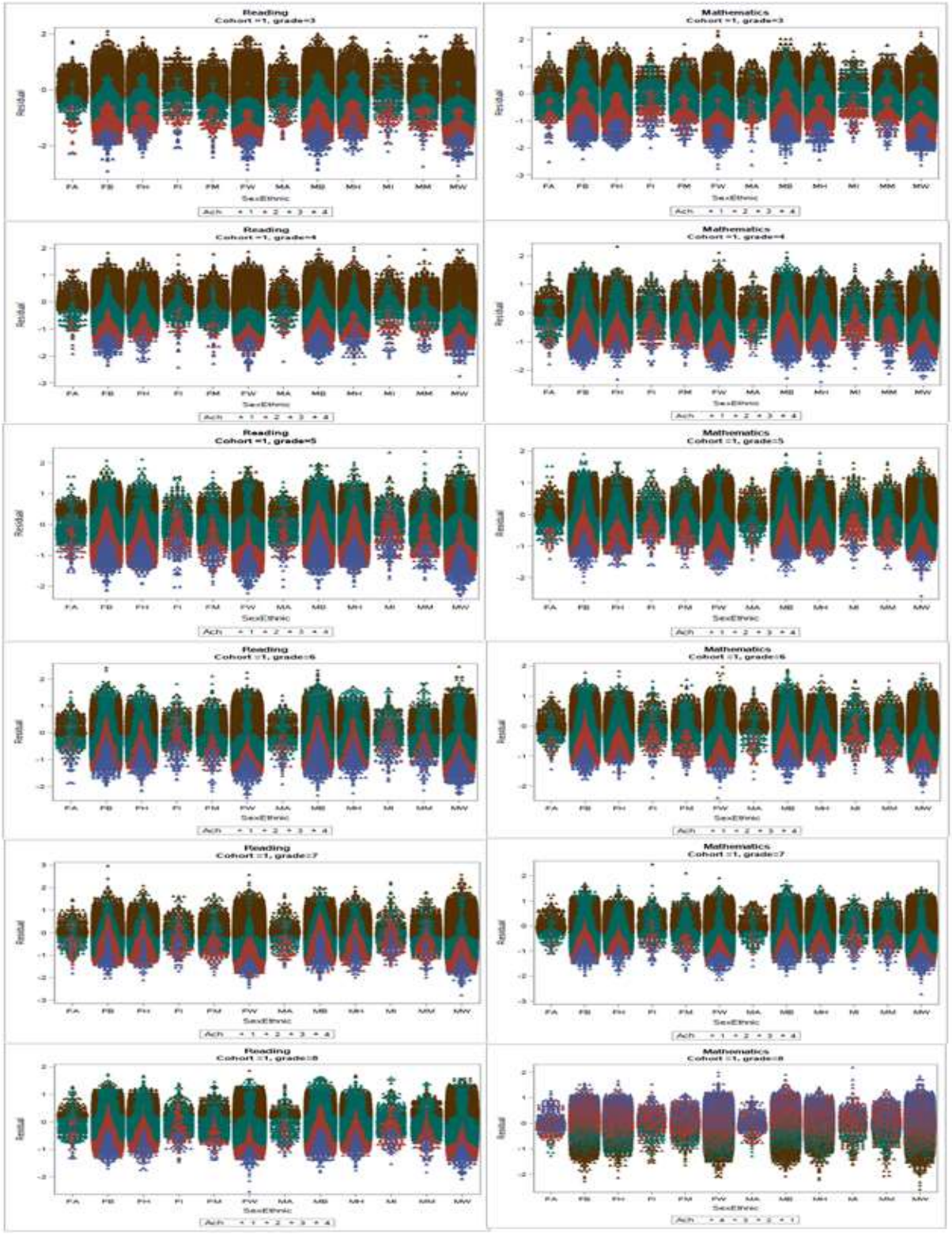

Figure 3. Residual of comparing Predicted students' achievement score, using the multiple regression model (Model2) with student EOG score when the EOG test score is level 1, 2, 3, and 4 in mathematics and reading for cohort1 based on students' gender, and ethnicity, and grade level

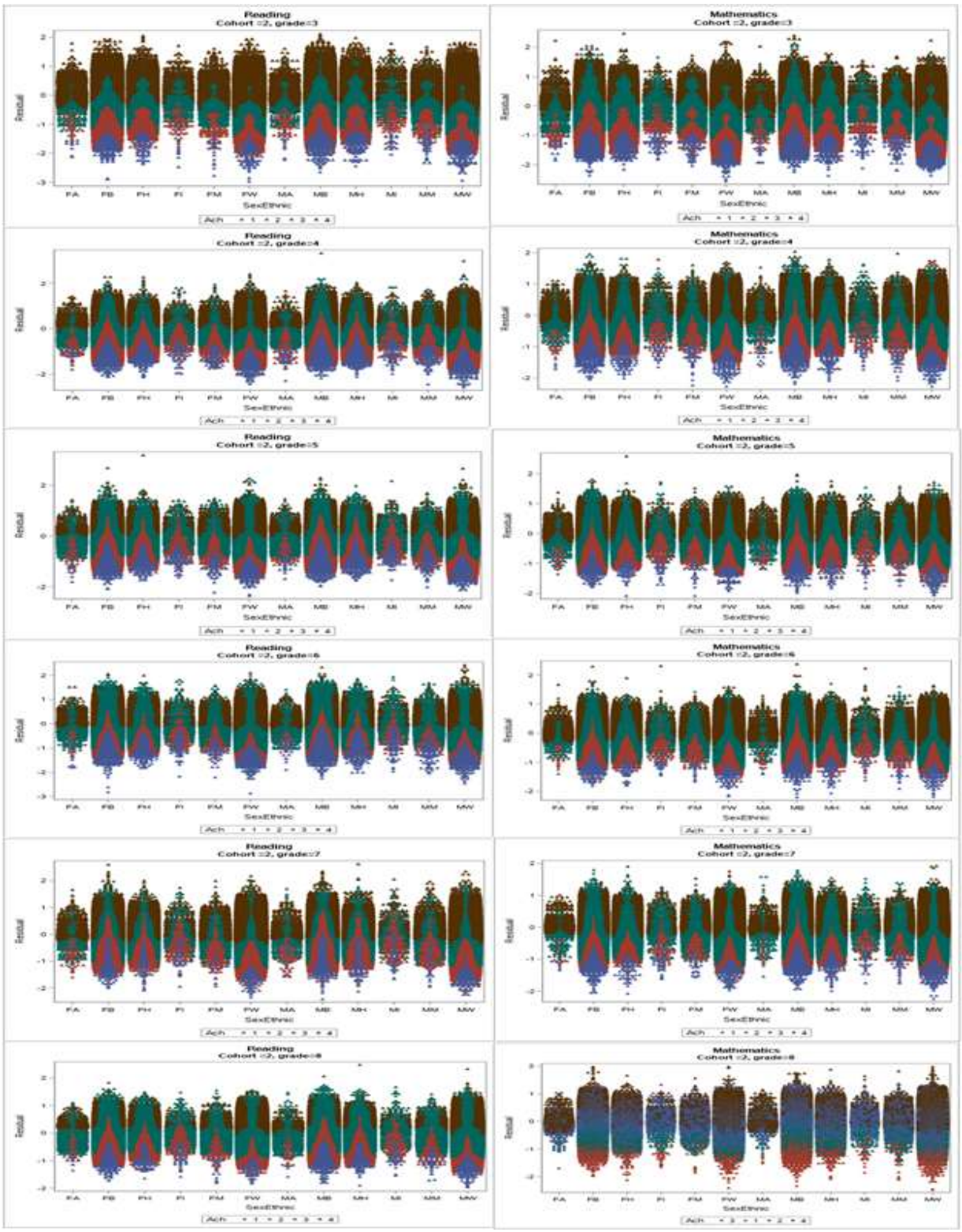

Figure 4. Residual of comparing Predicted students' achievement score, using the multiple regression model (Model2) with student EOG score when the EOG test score is level 1, 2, 3, and 4 in mathematics and reading for cohort2 based on students' gender, and ethnicity, and grade level

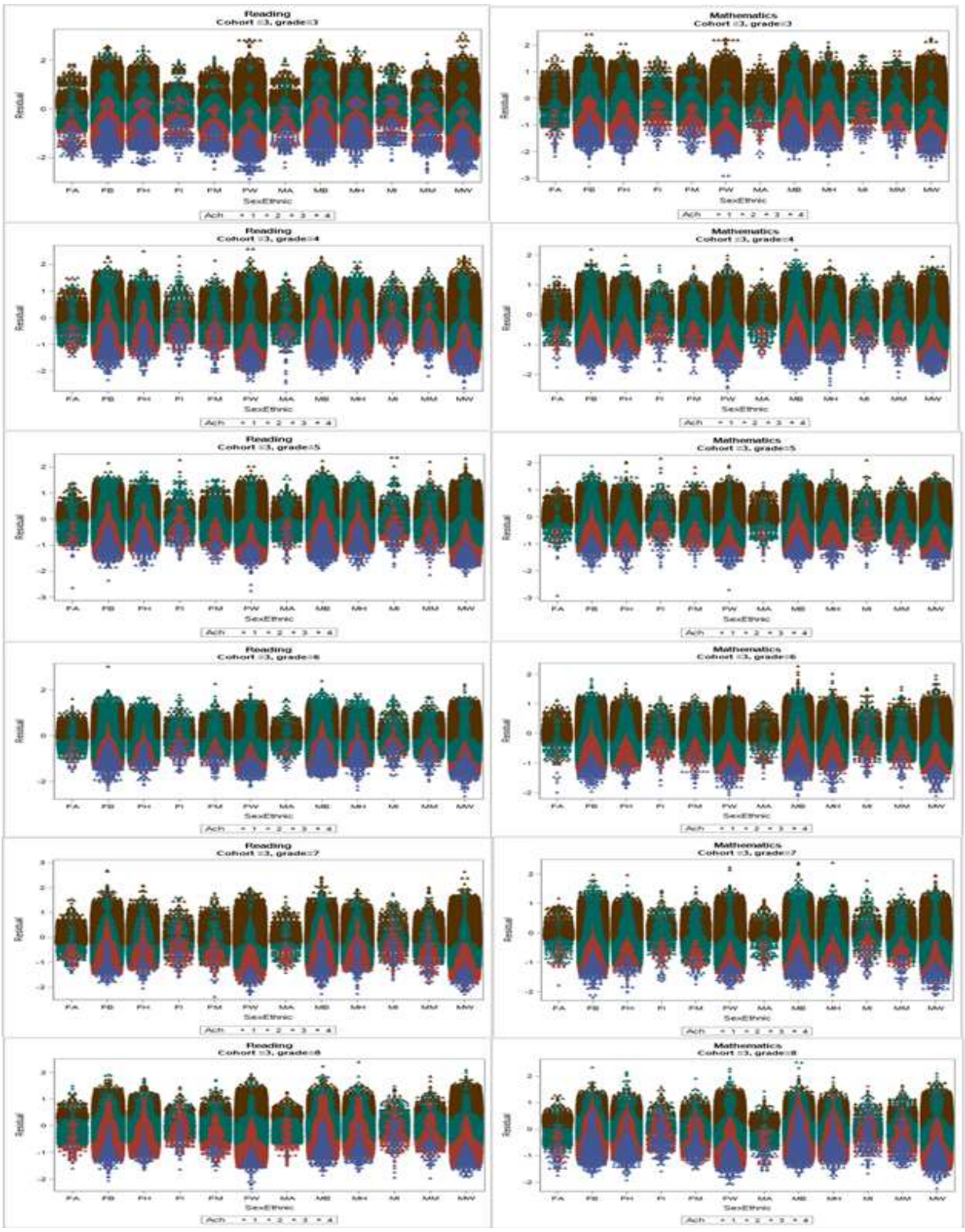

Figure 5. Residual of comparing Predicted students' achievement score, using the multiple regression model (Model2) with student EOG score when the EOG test score is level 1, 2, 3, and 4 in mathematics and reading for cohort3 based on students' gender, and ethnicity, and grade level

Conclusion

In this research, SAS provides an effective platform for the analysis of a large set of system-wide education data. Specifically, we explored the role of an educator's assessment on student performance and their future academic placement. Teacher judgment is an influencing factor on student's perception of his/her own academic abilities as can be seen by the literature and confirmed by our analyses. This research studies the relationship between a student's EOG test achievement level score and the corresponding teacher judgment score as a function of student's demographics. This data was sourced from the NCERDC and records performance in both reading and mathematics in 3rd to 8th grade from 2006 to 2013 in North Carolina. Prior work analyzed the relationship using correlation and hypothesis tests for different samples of data. In general, there is a positive and significant relationship between teacher judgment and student's EOG performance in mathematics and reading. The correlations generally higher in mathematics than reading, which is expected as reading can be assessed more subjectively than mathematics. For each grade level, correlations are increased for more recent time periods (from 2006 to 2013), which may be related to curriculum alignment with the EOG exams. For each year, higher grade levels demonstrate weaker correlations, particularly in reading which may reflect the differences in the level of student-teacher interaction in middle school as compared to elementary school. We modeled the relationship using multiple regression and hierarchical approaches via PROC GLM and PROC MIXED.

This paper is focused on comparing the HLMs with similar multiple regression models. School identifier is defined and is used to cluster students for both models. In 2-Level HLMs, the school identifier is used to identify within and between school effects. Similarly, in the multiple regression model (Model2), school identifier is considered as another predictor variable to monitor the effect of each specific school on the relationship and resulted models. Comparative results for fit statistics in both models demonstrate that Model2 provides a better fit for each grade level model than HLM in both reading and mathematics. The selected effects of Model3 for each grade level model demonstrates that school identifier is a highly influencing factor which cannot be ignored. Moreover, these results show the models provide a better estimation for higher grade levels as there is more information from the previous grades available to help predict students' EOG test performance. Even by considering all of these variables and their interactions, there is around 30% of variation that remains unexplained. This can be related to models' linear assumption or the absence of factors such as socioeconomic status, parental education level, or the parents' occupations however, those records are not available in all data sets and cannot be considered consistently for modeling.

Based on the Model2, in each grade level model, teacher judgment and students' gender are significant factors for predicting student EOG performance in mathematics and reading comprehension. The previous grade's EOG test scores and teacher judgment scores, especially the more recent ones, significantly affect EOG test performance in the current grade level in both mathematics and reading. In each subject of each grade level model, the effect of the previous grade's EOG performance in that subject is greater than teacher judgment for the same subject. Results also demonstrate that the year students attend a specific grade level is a significant factor in each grade level model, where the corresponding year has the higher influence and the effect is decreased for the previous year. Predicted student EOG performance is lower than actual student EOG score for proficient students (EOG achievement levels 3 or 4), but is higher for non-proficient students (achievement levels 1 or 2). This suggests that Model2 overestimates EOG performance for low-performing students, while it underestimates performance for high-performing ones. Moreover, analyses demonstrate the importance of school effect, and its

significance even for 10% sample. Besides, applying Model2 on the 10% sample shows that students' demographics are not as important as current grade teacher judgment and previous grade EOG score and teacher judgment and the cohort and the year these scores are belong.

The correlation and regression analyses in this study revealed student EOG performance is related to academic performance in previous grade levels as well as historical teacher judgements, and there is a discrepancy in this relationship by gender and ethnicity over time. It is important to understand the relationship between student performance and teacher judgement due to its effect on students' future academic placements.

This research used SAS to analyze the longitudinal relationship between students' demographics, teacher judgement and student EOG test performance in mathematics and reading comprehension over time using the NCERDC data set. SAS provides an easy and efficient environment to analyze this large data set and assess a variety of hypotheses and models on a desired sample of data. It enables us to efficiently manipulate with very large and bulky data set. SAS provides a safe and secure place to ensure a privacy of data. This makes SAS environment popular and efficient for educational analysis. EVAAS was also implemented in SAS to take advantage of its benefits like reliable and precise estimation for large data. Here, we used SAS for data visualization and color-coding, which helped us to identify patterns that prompted us to explore other factors that may help us to improve the models. This research provides an initial framework of comprehensive statistical evidence to detect student's learning trajectory where teacher judgement proves especially notable.

References

Beswick, J. F., Willms, J. D., & Sloat, E. A. (2005). A comparative study of teacher ratings of emergent literacy skills and student performance on a standardized measure. *Education, 126*, 116-137.

Carolina, P. S. (1999). *Assessment Brief: Understanding End-of-Grade Testing - Achievement Levels.* Retrieved November 1, 2018, from http://www.ncpublicschools.org/docs/accountability/testing/eog/asb_achlev.pdf

Carolina, P. S. (2006). *Achievement Level Descriptors for the EOG Mathematics Tests.* Retrieved November 1, 2018, from http://www.ncpublicschools.org/docs/accountability/testing/eog/eogmathgr3to5ald.pdf

Coladarci, T. (1986). Accuracy of Teacher Judgments of Student Responses to Standardized Test Items. *Journal of Educational Psychology, 78*(2), 141-146.

Demaray, M. K., & Elliott, S. N. (1998). Teachers' Judgments of Students' Academic Functioning: A Comparison of Actual and Predicted Performances. *School Psychology Quarterly, 13*(1), 8-24.

Gallant, D. J., & Moore, J. L. (2008). Ethnic-Based Equity in Teacher Judgment of Student Achievement on a Language and Literacy Curriculum-Embedded Performance Assessment for Children in Grade One. *Educational Foundations*, 63-77.

Hinnant, J. B., O'Brien, M., & Ghazarian, S. R. (2009). The longitudinal relations of teacher expectations to achievement in the early school year. *Journal of Educational Psychology, 101*, 662-670.

Hoge, R. D. (1983). Psychometric properties of teacher-judgment measures of pupil aptitudes, classroom behaviors, and achievement levels. *The Journal of Special Education, 17*, 401-429.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59*, 297-313.

Hoge, R., & Butcher, R. (1984). Analysis of teacher judgments of pupil achievement level. *Journal of Educational Psychology, 76*, 777-781.

Martínez, J. F., Borko, H., & Stecher, B. (2009). Classroom Assessment Practices, Teacher Judgments, and Student Achievement in Mathematics: Evidence from the ECLS. *Educational Assessment, 14*, 78-102.

Meisels, S. J., Bickel, D. D., Nicholson, J., Xue, Y., & Atkins-Burnett, S. (2001). Trusting teachers' judgments: A validity study of a curriculum-embedded performance assessment in Kindergarten-Grade 3. *American Educational Research Journal,, 38*(1), 73-95.

Meissel, K., Meyer, F., Yao, E. S., & Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments: Exploring student characteristics that influence teacher judgments of student ability. *Teaching and Teacher Education, 65*, 48-60.

Meshkinfam, S., Ivy, J., & Reamer, A. (2019a). Quantifying the Role of Sex and Ethnicity on the Relationship between Teacher Judgement and Student Performance on Standardized Exams in Mathematics and Reading. *American Society for Engineering Education (ASEE) Southeastern Section Conference*, (p. 16). Raleigh, NC. Retrieved from https://papers.asee-se.org/openconf/modules/request.php?module=oc_program&action=view.php&id=47&type=5&a=

Meshkinfam, S., Ivy, J., & Reamer, A. (2019b). Perception vs. Reality: Correlation Analysis, Teacher Judgment, and Student Performance. *American Educational Research Association (AERA) Annual Meeting*, (p. 20). Toronto, Canada.

Möller, J., Pohlmann, B., Köller, O., & Marsh, H. W. (2009). A metaanalytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research, 79*, 1129-1167.

Mowrey, S. C., & Farran, D. C. (2016). Performance and Preparation: Alignment between student achivement, teacher ratings, and parent perceptions in urban middle-grades mathematics classrooms. *Journal of Urban Learning Teaching and Research, 12*, 61-74.

Rausch, T., Karing, C., Dörfler, T., & Artelt, C. (2016). Personality similarity between teachers and their students influences teacher judgement of student achievement. *Educational Psychology, 36*(5), 863-878.

Ready, D., & Wright, D. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal, 48*(2), 335-360.

Rodriguez, M. C. (2004). The role of classroom assessment in student performance on TIMSS. *Applied Measurement in Education, 17*(1), 1-24.

SAS. (2014). *SAS® EVAAS® for K-12*. Retrieved from https://www.sas.com/content/dam/SAS/en_us/doc/factsheet/education-evaas-104850.pdf

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of Teachers' Judgments of Students' Academic Achievement: A Meta-Analysis. *Journal of Educational Psychology, 104*(3), 743-762.

Valdez, A. (2013). Teacher Judgment of Reading Achievement: Cross-Sectional and Longitudinal Perspective. *Journal of Education and Learning, 2*(4), 186-200.

Zaiontz, C. (2017). *Real Statistics Using Excel, Correlation testing via Fisher transformation.* Retrieved Nov 2017, from http://www.real-statistics.com/correlation/one-sample-hypothesis-testing-correlation/correlation-testing-via-fisher-transformation/

Contact Information

Your comments and questions are valued and encouraged. Contact the author at:

Sareh Meshkinfam
NC State University
smeshki@ncsu.edu