2016 NCTM RESEARCH CONFERENCE
April 11-13 • San Francisco

Paper Title: Exploring Effects of Content Organization on Specific Algebraic Concepts: A Propensity Score Analysis

Author(s): Derrick Saddler

Session Title: Exploring Effects of Content Organization on Specific Algebraic Concepts

Session Type: Brief Research Report

Presentation Date: April 12, 2016

Presentation Location: San Francisco, California

Exploring Effects of Content Organization on Specific Algebraic Concepts: A Propensity Score Analysis

The Common Core State Standards for Mathematics (CCSSM) delineates the specific content students should learn while they are in high school (National Governors Association Center for Best Practices [NGA Center] and the Council of Chief State School Officers [CCSSO], 2010a, 2010b). As educational leaders consider how to implement the CCSSM, an important consideration is how to organize the high school mathematics program into courses that provide a strong foundation for success at the post-secondary level. To assist in addressing this need, a group of mathematics experts from different levels of academia and workforce representatives were convened to develop model course pathways in mathematics based on the CCSSM. One of the model course pathways is the U.S. traditional high school mathematics sequence that includes Algebra I-Geometry-Algebra II. Researchers refer to this pathway as a *subject-specific* approach (Grouws, Tarr, Chávez, Sears, Soria & Taylan, 2013; Tarr, Grouws, Chávez, & Soria, 2013) because the primary content in each course in the sequence is related directly to the name of each course. Another model course pathway suggested in the CCSSM is Mathematics I-II-III. In the U.S., this course pathway is typically offered in curricula materials developed in response to standards-based reform documents (Senk & Thompson, 2003). Researchers refer to this pathway as an *integrated mathematics* approach because multiple mathematical strands are integrated in each course (Grouws et al., 2013; Tarr et al., 2013). The content in each of the integrated mathematics courses includes number, algebra, geometry, probability and statistics.

The high school portion of the CCSSM can be implemented using either a subject specific or integrated course pathway. The decision regarding which approach to use is made by state or local education agencies. However, no information is provided as part of the CCSSM document about the advantages or disadvantages of either approach. Furthermore, little research provides comparison of the mathematics performance of high school students who learn from subject-specific and integrated course pathways (NMAP, 2008). The most recent studies that relate high school mathematics content organization to students' achievement is the Comparing Options in School Mathematics: Investigating Curricula (COSMIC) project.

In the COSMIC project, researchers conducted three studies in which they investigated the effects of content organization, prior achievement, and curriculum implementation on high school students' mathematics learning (Grouws et al., 2013; Tarr et al., 2013; Chavez et al., 2013). The researchers used multiple measures of learning to explore students' achievement. The measures included a state-mandated 8th grade test, a nationally standardized assessment known as the Iowa Test of Educational Development (ITED), and two project developed tests referred to as the Problem Solving and Reasoning Test (PSRT) and the Test of Common Objectives (TOCO). The state mandated 8th grade test was used as a measure of prior achievement. The ITED assessments were 40-question multiple-choice tests that assess students' computational and problem-solving skills in a number of mathematical contexts. The PSRT assessments consist of topics deemed appropriate based on content analyses and feedback from external reviewers. The tests were designed to assess nontrivial mathematical reasoning and problem-solving skills that focused on aspects of algebra, geometry, and statistics. The TOCO assessments consist mainly of constructed-response items that assess concepts and skills common to the respective implemented curricula materials.

To examine students' performance, in each study the researchers fit each outcome measure to a three-level hierarchical linear model (HLM). A hierarchical linear model was used to take into account the structure of the data in which students are nested in classrooms and classrooms are nested in schools. In the first study, Grouws et al. (2013) examined the performance of students from 10 schools. Findings reveal students who enrolled in Integrated Mathematics I (n=1256) performed statistically better on the ITED, PSRT, and TOCO than students who enrolled in Algebra I (n=1365) with small effect sizes of 0.166, 0.453, and 0.308, respectively. In the second study, Tarr et al. (2013) examined the performance of students from 11 schools. Findings reveal students who enrolled in Integrated Mathematics II (n=1171) performed statistically better on the ITED than students who enrolled in Geometry (n=2087) with a small effect size of 0.294. However, no statistical mean differences in performance on the PRST and TOCO were found between the two groups. In the third study, Chavez et al. (2013) examined the performance of students from 10 schools. In this study, the researchers only reported results on the ITED and TOCO mathematics assessments. Findings reveal students who enrolled in Integrated Mathematics III (n=892) performed statistically better on the TOCO than students who enrolled in Algebra II (n=1350) with a small effect size of 0.33. However, no statistical mean differences in performance on the ITED were found between the two groups.

Collectively, findings from the COSMIC project reveal content organization and prior achievement were key factors in improving students' learning in the first three years of high school. However, the findings based on the different outcomes measures are not consistent throughout the three studies. For example, on the ITED assessments, the researchers found statistical mean differences in students' performance in the first and second year studies in favor of students enrolled in the integrated courses (Grouws et al., 2013; Tarr et al., 2013). However,

in the third year study no statistical mean differences in performance on the ITED were found between students enrolled in the two curricular approaches (Chavez et al., 2013). The findings based on the ITED assessments suggest differences exist in students' learning during the first two years of high school, but the differences disappear by the time the students complete the eleventh grade. On the project developed tests, and more specifically the TOCO assessments, the researchers found statistical differences in the first and third year studies in favor of students enrolled in the integrated courses (Grouws et al., 2013; Chavez et al., 2013). However, in the second year study no statistical differences in performance on the TOCO were found between students enrolled in the two curricular approaches (Tarr et al., 2013). The findings suggest statistical differences exist in students' learning during the first and third years of high school, but no statistical differences exist in students' learning during the second year of high school. More specifically, the findings suggest no statistical differences exist in students' performance on common geometry concepts, but differences exist in students' performance on common algebraic concepts in favor of students who enroll in integrated courses.

There are some methodological shortcoming associated with the COSMIC project. In particular, the researchers of the project conducted each of their studies on a single academic year. However, an examination of students' learning over multiple years can add new insights related to the influence of the entire curriculum sequence. Another methodological shortcoming is results from the COSMIC projects are based solely on the aggregate measures. Findings based solely on aggregate measures can suggest whether one curriculum is better than another in improving students' learning, but they do not reveal more nuanced differences such as specific content in which students benefited. The COSMIC researchers could have identified more detailed differences in students' performance if they performed an analysis of specific test items

or concepts (Cai et al., 2011; Huntley, Rasmussen, Villarubi, Sangtong, & Fey, 2000). In sum, a longitudinal study that incorporates a pretest-posttest design in which specific test items are examined can be used to investigate students' learning growth over the course of the first three years of high school.

Purpose & Research Question

The purpose of this study was to statistically compare mathematics outcomes of high school students who learned from subject-specific course pathways (i.e., Algebra I-Geometry-Algebra II) to a comparable group of students who learned from an integrated course pathway (i.e., Mathematics I-II-III). The question this study investigated is "How does the algebraic performance of high school students enrolled in integrated course pathways relate to the algebraic performance of high school students enrolled in subject-specific course pathways?"

Data Source: High School Longitudinal Study

The data from a large scale observational study conducted by the National Center for Education Statistics (NCES) known as the High School Longitudinal Study of 2009 (HSLS) was used in this study. The HSLS collected data on the high school and postsecondary experiences of a nationally representative sample of high school students beginning with their ninth grade year. The target population for the HSLS included all ninth-grade students who attended public and private schools in the United States. The HSLS is a complex sample survey that includes 21,444 students who were selected from 944 schools. The study participants were administered a survey instrument and assessment at the beginning of their ninth grade year and at the end of their eleventh grade year. The content of the student questionnaire included demographic information, such as race, gender, socioeconomic status, and students' high school mathematics

courses. Data from the HSLS student questionnaire were the primary data used in this study. Therefore, the unit of analysis for this study is the student.

Sample groups

The variable examined in this study is high school mathematics course pathways, specifically, the subject-specific pathway (Algebra I-Geometry-Algebra II) and the integrated pathway (i.e. Mathematics I-II-III). The selected cases from the HSLS represent two well-defined groups of high school students enrolled in the same course pathway and same school for the first three years of high school. The HSLS public data set only inquires about courses students are enrolled in during their ninth grade year and their eleventh grade year. As a result, there were no data indicating what course the students were enrolled in during their tenth grade year. Therefore, the course pathways were inferred based on students' ninth and eleventh grade mathematics course enrollment.

One group represents high school students who learned mathematics from a subject-specific course pathway (Algebra I-Geometry-Algebra II) and the other group represents high school students who learned mathematics from an integrated course pathway (Integrated Mathematics I-II-III). To identify students in the subject-specific group, cases from the HSLS dataset that indicated students enrolled in Algebra I in the ninth grade and Algebra II in the eleventh grade, and not in Integrated Mathematics I in the ninth grade and Integrated Mathematics III in the eleventh grade were selected. This yielded a sample with 4956 students who learned from a subject-specific course pathway. Similarly, to identify students in the integrated group, cases that indicated students enrolled in Integrated Mathematics I in the ninth grade and Integrated Mathematics III in the eleventh grade and not in Algebra I in the ninth

grade and Algebra II in the eleventh grade were selected. This yielded a sample with 73 students who learned from an integrated course pathway.

Outcome Measures

The HSLS administered a mathematics assessment to the participants at the beginning of their ninth grade year and again at the end of their eleventh grade year. The 40-item assessment measured students' performance on algebraic skills, reasoning and problem solving. The assessment included a mixture of ninth and eleventh grade items in both stages of the test (Ingels et al., 2011). The HSLS outcome measures include an item response theory (IRT) based estimate of the score for each participant on the full set of items. In addition, the HSLS includes sets of clustered-items that represent a broad spectrum of algebraic concepts. Each set of clustered-items represents four questions from the assessment and relates to specific content. The sets of clustered-items measured student proficiency with algebraic concepts and represent the outcome variables for the study (Ingels et al., 2010). The following proficiency levels represents the specific algebraic concept.

- Proficiency 1: Evaluate simple algebraic expressions and translate between verbal and symbolic representations of expressions,
- Proficiency 2: Solve proportional situation word problems, find the percent of a number, and identify equivalent algebraic expressions for multiplicative situations,
- Proficiency 3: Link equivalent tabular and symbolic representations of linear equations, identify equivalent lines and find the sum of variable expressions,
- Proficiency 4: Solve systems of equations algebraically and graphically and characterize lines represented by a system of linear equations, and
- Proficiency 5: Find and use slopes and intercepts of lines, and use functional notation.

The five levels are hierarchical in the sense that mastery of a higher level typically implies proficiency at the lower levels (Ingels et al., 2011). The IRT-estimated reliability of the HSLS test is 0.92 after sample weights are applied. This 0.92 reliability applies to all scale scores derived from the IRT estimation including the probability of proficiency scores.

Methods

Design

This quantitative study employs a nonequivalent comparison group design. The design is similar to a true experiment because subjects in each group took a pretest and a posttest. However, unlike a true experiment, subjects in the nonequivalent comparison group design were not randomly assigned to treatment and control groups. Consequently, the main threat to the internal validity of a nonequivalent comparison group design is the possibility that group differences on the outcome variables will be a result of preexisting group differences rather than to a treatment effect, or selection bias (Gall, Gall, & Borg, 2007). The main problem causing selection bias in non-randomized control trials is nonequivalence of treatment and control groups. If differences between students who enroll in subject-specific and integrated course pathways can be eliminated, then presumably the threat of selection bias will be eliminated. To reduce the threat of selection bias due to non-random assignment of students, a propensity score matching procedure will be employed

Propensity Score Matching Procedure

The goal of the propensity score matching procedure is to match, as closely as possible, each student who learned from an integrated course pathway with a student who learned from a subject-specific course pathway. The first step in the propensity score matching procedure is to create a logistic regression model to calculate the probability students are placed in their

respective course pathway. Included in the logistic regression model are covariates related to students' prior achievement, gender, race, socioeconomic status, and a student longitudinal analytic weight. The inclusion of these variables was based on their use in previous research studies (Cai et al., 2011; Cai et al., 2013; Chavez et al., 2013; Grouws et al., 2013; Tarr et al., 2013; DuGoff, Schuler, & Stuart, 2014).

The calculated probability from the logistic regression model is called a propensity score, which takes a value between 0 and 1. The second step is to use a caliper to create a 1:1 match of students in the groups. The pairs are selected at random from subject-specific and integrated course pathway students whose difference in propensity scores is less than 0.1 of each other. The caliper matching procedure matches each student with a given propensity score who learned from an integrated course pathway with a student who has a nearly identical propensity score, but enrolled in a subject-specific course pathway. The students in the integrated course pathway function like a treatment group. The students in the subject-specific course pathway function like a control group. The subject-specific student in each pair provides a "counterfactual" estimate of what the outcome for the integrated student would have been if that student had learned from a subject-specific pathway.

Statistical Analysis

To examine whether differences exist in students' outcome measures, the mean performance gain of students who enrolled in a subject-specific course pathway was statistically compared to the mean performance gain of students who enrolled in integrated course pathways. Mean gain scores for each student were determined by calculating the difference between their posttest and pretest scores. After gain scores were calculated for each student, the difference in the gains of students in each matched pair was calculated. Finally, the SAS statistical software

package (SAS 9.4) was used to perform correlated means *t*-test to determine if the mean difference in gain scores on the overall assessment and proficiency levels were statistically different from zero. In addition, effect sizes (i.e. Cohen's *d*) are calculated by dividing the difference in mean gains score by the pooled standard deviation of the mean gain scores.

Results

Descriptive Statistics

In the propensity score matching procedure numerous logistic regression models were specified with a goal of identifying the best model with the best balance on all observed variables. Although no rule exists for how close to zero will achieve adequate balance, researchers suggest balance is achieved when the index is very close to zero for each of the pretest covariates and also for the propensity score itself (Steiner et al., 2010). Steiner and colleagues suggest acceptable balance is achieved when Cohen's *d* for all continuous measures was $d < 0.20$, and the odds ratio for all categorical variables was between 0.80 and 1.25. The model with the best balance included variables related only to students' gender, race, socioeconomic status, weights, and prior achievement. The descriptive statistics of all the covariates included in the logistic regression model before and after matching are presented in Tables 1 - 3. Table 4 presents the mean and standard deviation of students' performance on the pretest and posttest measures, and the mean gain scores of the groups.

Table 1
Students' Demographics Before and After Matching (Percentage of Sample in Parentheses)

Characteristics	Before Matching			After Matching		
	Subject-specific (n=4956)	Integrated (n=73)	Effect size (*OR*)	Subject-specific (n=73)	Integrated (n=73)	Effect size (*OR)*
Male	2450 (0.49)	38 (0.52)	0.94	36 (0.49)	38 (0.52)	0.94
Female	2506 (0.51)	35 (0.48)	1.05	37 (0.51)	35 (0.48)	1.05
Black	526 (0.11)	10 (0.14)	0.78	9 (0.12)	10 (0.14)	0.90
White	2893 (0.58)	43 (0.59)	0.99	43 (0.59)	43 (0.59)	1.00
Hispanic	748 (0.15)	9 (0.12)	1.26	8 (0.11)	9 (0.12)	0.91
Other Races	789 (0.16)	11 (0.15)	1.06	13 (0.18)	11 (0.15)	1.18

Note. Other races represent students classified as Asian, Native Hawaiian, American Indian, and Multi-racial.

Table 2

Students' Mean Socioeconomic Status, Weight, and Propensity Score Before and After Matching (Standard Deviation in Parentheses)

	Before Matching			After Matching		
	Subject-specific (n=4956)	Integrated (n=73)	Effect size(*d*)	Subject-specific (n=73)	Integrated (n=73)	Effect size (*d)*
SES	0.11 (0.72)	-.067 (0.69)	0.24	-0.097 (0.71)	-.067 (0.70)	0.04
Weight	209.64 (295.7)	170.48 (156.4)	0.16	189.85 (180.8)	170.48 (156.4)	0.11
Propensity	0.01(0.01)	0.02(0.01)	0.52	0.021 (.01)	0.021 (0.01)	0.00

Table 3

Students' Means 9th Grade Performance Before and After Matching (Standard Deviation in Parentheses)

	Before Matching			After Matching		
Score	Subject-specific (4956)	Integrated (n=73)	Effect size (*d*)	Subject-specific (n=73)	Integrated (n=73)	Effect size (*d)*
Overall	39.42 (9.03)	41.93 (10.76)	0.25	42.01 (10.17)	41.93 (10.76)	0.01
Proficiency 1	0.92 (0.19)	0.91 (0.23)	-0.04	0.92 (0.23)	0.91 (0.23)	0.04
Proficiency 2	0.64 (0.30)	0.70 (0.32)	0.19	0.71 (0.30)	0.70 (0.32)	0.03
Proficiency 3	0.41 (0.31)	0.51 (0.35)	0.31	0.51 (0.34)	0.51 (0.35)	0.01
Proficiency 4	0.14 (0.15)	0.21 (0.23)	0.33	0.20 (0.20)	0.21 (0.23)	0.03
Proficiency 5	0.07 (0.04)	0.09 (0.07)	0.34	0.08 (0.06)	0.09 (0.07)	0.10

Table 4

Statistics of Students' Mean Performance (Standard Deviations in Parentheses)

	Integrated			Subject-Specific		
	Pretest	Posttest	Gain	Pretest	Posttest	Gain
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Overall	41.9 (10.7)	68.1 (17.70)	26.2 (11.9)	42.0 (10.2)	66.9 (15.4)	24.9 (11.1)
Proficiency 1	0.91 (0.22)	0.94 (0.18)	0.03 (0.19)	0.92 (0.23)	0.95 (0.17)	0.04 (0.24)
Proficiency 2	0.70 (0.32)	0.81 (0.31)	0.11 (0.25)	0.71 (0.3)	0.84 (0.26)	0.12 (0.25)
Proficiency 3	0.51 (0.35)	0.73 (0.35)	0.22 (0.30)	0.51 (0.34)	0.73 (0.31)	0.22 (0.29)
Proficiency 4	0.21 (0.23)	0.35 (0.34)	0.14 (0.25)	0.20 (0.20)	0.29 (0.3)	0.08 (0.23)
Proficiency 5	0.09 (0.08)	0.21 (0.30)	0.12 (0.26)	0.08 (0.06)	0.16 (0.26)	0.07 (0.24)

Upon completion of numerous iterations of the propensity score matching procedure, and based on an *a priori* power analysis (effect size = .50, alpha level = .05, power = .99), a sufficient sample size of 146 students or 73 matched pairs was yielded. Collectively, the effect sizes of the differences in students' prior achievement, socioeconomic status, propensity score, and weighting after matching were trivial, which suggests the groups were balanced on all covariates included in the propensity score model. Therefore, and theoretically, statistical differences in outcomes measures between students who learned from subject-specific and integrated course pathways represent an unbiased estimate of the population mean difference (Graham, 2010; Rosenbaum & Rubin, 1983).

Inferential Statistics

In this section, results of the correlated means *t*-test and calculated effect sizes (*d*) based on statistics presented in Tables 4 are presented. On the overall assessment which included 72 questions, students who enrolled in integrated courses had slightly greater gains (M =26.19, SD =11.88) than students who enrolled in subject-specific courses (M=24.9, SD =11.13). However, the results of the *t*-test ($M = 1.29, SD = 16.10, t(200) = 0.13, p = .89$) reveal no statistically significant difference in gain scores. The calculated effect size value is $d = .11$.

On items related to algebraic expressions (i.e. proficiency 1) on the algebra assessment, students who enrolled in subject-specific courses had slightly greater gains (M=0.04, SD =0.24) than students who enrolled in integrated courses (M =0.03, SD =0.19). However, the results of the *t*-test ($M = -0.01, SD = 0.29, t(200) = -0.36, p = 0.72$) reveal no statistically significant difference in gain scores. The calculated effect size value is $d = .03$.

On items related to multiplicative and proportional thinking (i.e. proficiency 2) on the algebra assessment, students who enrolled in subject-specific courses had slightly greater gains (M=0.12, SD =0.25) than students who enrolled in integrated courses (M =0.11, SD =0.25). However, the results of the *t*-test ($M = -0.01, SD = 0.31, t(200) = .03, p = 0.97$) reveal no statistically significant difference in gain scores. The calculated effect size value is *d* = .05.

On items related to linear equivalents (i.e. proficiency 3) on the algebra assessment, the gains of students who enrolled in subject-specific courses is similar to the gains (M=0.22, SD = 0.29) of students who enrolled in integrated courses (M =0.22, SD =0.30). The results of the *t*-test ($M = 0.00, SD = 0.39, t(200) = 0.43, p = 0.67$) reveal no statistically significant difference in gain scores. The calculated effect size value is *d* = .003.

On items related to systems of equations (i.e. proficiency 4) on the algebra assessment, students who enrolled in integrated courses had greater gains (M =0.14, SD = 0.25) than students who enrolled in subject-specific courses (M=0.08, SD =0.23). However, the results of the *t*-test ($M = 0.05, SD = 0.35, t(200) = -0.05, p = 0.96$) reveal no statistically significant difference in gain scores. The calculated effect size value is *d* = .22.

On items related to linear functions (i.e. proficiency 5) on the algebra assessment, students who enrolled in integrated courses had greater gains (M =0.12, SD =0.26) than students who enrolled in subject-specific courses (M=0.07, SD =0.24). However, the results of the *t*-test ($M = 0.05, SD = 0.33, t(200) = -0.43, p = 0.66$) reveal no statistically significant difference in gain scores. The calculated effect size value is *d* = .19.

Conclusion

The question this study investigated is "How does the algebraic performance of high school students enrolled in integrated course pathways relate to the algebraic performance of

high school students enrolled in subject-specific course pathways?" Collectively, on the overall algebra assessment and specific content items, the results of the correlated *t*-tests reveal no statistically significant differences exist in the algebraic performance gains between the high school students who learned mathematics from integrated course pathways and the high school students who learned from subject-specific course pathways. In addition, the calculated effects sizes (*d*) suggest content organization has low practical significance on students' algebraic proficiency. Consistent with expectations of the National Mathematics Advisory Panel (NMAP, 2008), the results of this study suggest high school students can perform comparably through algebraic content regardless of whether the students enroll in a subject-specific or integrated course pathway.

The non-statistically significant results from this study contradict results from the COSMIC project that suggest high school students who study from an integrated curriculum are advantaged over students who study from subject-specific curricula (Chavez et al., 2013; Grouws et al., 2013; Tarr et al., 2013) during the first three years of high school. The differences in methodological approaches and findings between this study and findings from the COSMIC project suggest a need for more research in this area. New knowledge related to effects of content organization on students' achievement can assist in making decisions related to the type of course pathway to implement into high school mathematics programs. In particular, similar to this study, data from the COSMIC project can be further examined to identify more detailed differences in students' performance on specific concepts or test items. An analysis of items from assessment used in the COMIC project may identify specific content that attributed to statistical differences in the projects results.

Significance

A significant aspect of this exploratory study is it demonstrates a propensity score matching procedure that can be used in studies of curricular effectiveness to reduce the threat of selection bias when random assignment is not possible. Selection bias is a problem for mathematics education researchers who seek to make inferences about the effects of different instructional methods or curricular approaches in mathematics on student outcomes (Graham, 2010). For example, the students who participated in the COSMIC project were not randomly assigned to course pathways. Instead, they were observed in their natural occurrence. Grouws et al. (2013) and Tarr et al. (2013) used logistic regression to generate a propensity score for each student to investigate the threat of selection bias. The propensity scores revealed Hispanic and African American students were more likely to be assigned to the subject-specific course pathway. As a result, differences in outcomes can be attributed to these preexisting differences. In particular, Grouws et al. (2013) and Tarr et al. (2013) found that Hispanic and African American students performed statistically lower than White students on all measures. Because minority students were more likely to be assigned to the subject-specific course pathway, the results of studies from the COSMIC project (Grouws et al., 2013; Tarr et al., 2013) could have been negatively biased, causing students enrolled in the integrated course pathway to appear to perform better on some of the outcome measures.

Similar to the current study, the researchers could have investigated further and matched students based on their propensity scores to reduce the threat of selection bias. More specifically, if ancillary analysis were conducted using data from the COSMIC project, it would be interesting to determine whether the results would be consistent with the project's original findings if a propensity score matching method is employed.

References

Cai, J., Moyer, J., Wang, N., Hwang, S., Nie, B., & Garber, T. (2013). Mathematical problem posing as a measure of curricular effect on students' learning. *Educational Studies in Mathematics, 83*(1), 57–69.

Chávez, O., Papick, I., Ross, D. J., & Grouws, D. A. (2010). Developing fair tests for mathematics curriculum comparison studies: The role of content analyses. *Mathematics Education Research Journal, 23(4),* 397–416.

Cai, J., Wang, N., Moyer, J. C., Wang, C., & Nie, B. (2011). Longitudinal investigation of the curricular effect: An analysis of student learning outcomes from the the LieCal Project in the United States. *International Journal of Educational Research, 50,* 117-136.

Chavez, O., Tarr, J. E., Grouws, D. A., & Soria, V. M. (2013). Third-year high school mathematics curriculum: Effects of content organization and curriculum implementation. *International Journal of Science and Mathematics Education*. *(13)*1, 97-120.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

DuGoff, E. H., Schuler, M., & Stuart, E. A. (2014). Generalizing observational study results: Applying propensity score methods to complex surveys. *Health Services Research, 49*(1), 284-303.

Gall, M. D., Gall, J. P., & Borg, W. R. (2007). *Educational Research: An Introduction.* Pearson.

Graham, S. E. (2010). Using propensity scores to reduce selection bias in mathematics education research. *Journal for Research in Mathematics Education, 41*(2), 147-168.

Grouws, D. A., Tarr, J. E., Chavez, O., Sears, R., Soria, V., & Taylan, R. D. (2013). Curriculum and implementation effects on high school students' mathematics learning from curricula

representing subject-specific and integrated content organizations. *Journal for Research in Mathematics Education, 44*(2), 416-463.

Huntley, M. A., Rasmussen, C. L., Villarubi, R. S., Sangtong, J., & Fey, J. T. (2000). Effects of standards-based mathematics education: A study of the Core-Plus Mathematics Project algebra and functions strand. *Journal for Research in Mathematics Education,* 31(3), 328-361.

Huntley, M. A., & Terrell, M. (2014). One-step and multi-step linear equations: a content analysis of five textbook series. *ZDM - The International Journal on Mathematics Education, 46*(5), 751-766.

Ingels, S. J., Pratt, D. J., Herget, D. R., Burns, L. J., Dever, J. A., Ottem, R., . . . LoGerfo, L. (2011). *High School Longitudinal Study of 2009 (HSLS:09): Base-year data file documentation.* U.S. Department of Education; Institue of Education Sciences; National Center for Education Statistics.

Lanehart, R. E., Rodriguez, P., Kim, E. S., Ballara, A. P., Kromrey, J. D., & Lee, R. S. (2012). *Propensity score analysis and assessment of propensity score approaches using SAS procedures.* SAS Global Forum.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010a). *Common core state standards for mathematics*. Retrieved from http://www.corestandards.org

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010b). *Common core state standards for mathematics, Appendix A: Designing high school mathematics courses based on the common core state standards*.

Retrieved from http://www.corestandards.org/assets/CCSSI_Mathematics_Appendix_A.pdf

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine, 127*, 757-763.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological methods*, *15*(3), 250.

Thompson, D. R., & Senk, S. L. (Eds.) (2003). *Standards-based school mathematics curricula: What are they? What do students learn?* Mahwah, NJ: Lawrence Erlbaum.

Tarr, J. E., Grouws, D. A., Chávez, Ó., & Soria, V. M. (2013). The effects of content organization and curriculum implementation on students' mathematics learning in second-year high school courses. *Journal for Research in Mathematics Education, 44*(4), 683-729.