A National Analysis of Music Coursetaking, Social-Emotional Learning, and Academic
Achievement Using Propensity Scores

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy
in the Graduate School of The Ohio State University

By

Brian P. Shaw, MME

Graduate Program in Music

The Ohio State University

2020

Dissertation Committee:

Daryl W. Kinney, Advisor

Eugenia Costa-Giomi

Jan Edwards

Ann A. O'Connell

Copyright by

Brian P. Shaw

2020

Abstract

The purpose of this study was to examine the association between curricular high school music participation, academic achievement, and social-emotional learning. A large body of extant research has consistently demonstrated higher academic achievement among students who take elective music courses. Prior research has also shown that students who enroll and persist in American elective music courses have several important differences from students who do not. These differences have led researchers to wonder if music students' better observed performance on measures such as grade point average and standardized tests is a function of selection bias rather than a result of music instruction. Studies using techniques to control for selection bias have found that it significantly or entirely explained music students' higher scores. While the literature is replete with research on the association between music and academic achievement, this study is one of the first to investigate the relationship between music and social-emotional learning, one of the few studies using nationally representative data, and one of the few using techniques to support causal inference.

The data source for this investigation was the High School Longitudinal Study of 2009, a dataset assembled by the National Center for Education Statistics in the Institute of Education Sciences at the U.S. Department of Education. The dataset contained detailed transcript data, which permitted a thorough examination of which students enrolled in music courses, and for how long. The dataset also contained a large and heterogeneous body of covariates, which I used to estimate propensity scores that served to control for overt selection bias. The outcomes of

interest in this study were college entrance examination mathematics and reading scores, a proprietary test of mathematics reasoning ability, and scales for school motivation, mathematics self-efficacy, and science self-efficacy. I classified students' music involvement either as "short-term" (1-2 years) or "sustained" (3-4 years) enrollment in choral music, instrumental music, or general music. I addressed missing data through multiple imputation using chained equations. The analysis involved a "doubly robust" approach combining propensity score weighting and nested ordinary least squares (OLS; i.e., "linear") regression.

Results of the study were mixed. Preliminary unadjusted t tests revealed many significant differences between the choral and instrumental students and the control group, but the propensity score weighting moderated almost all of these effects to nonsignificance. The only unambiguously positive finding was that instrumental music students had somewhat higher reading scores. As there is a dearth of nationally representative research on ostensible social-emotional benefits of music participation, it is noteworthy that this analysis found that music students did not generally have higher *ceteris paribus* scores on those variables. Yet, subpopulation effects emerged for certain categories of music students based on factors such as race, sex, and prior school achievement. While the lack of widespread main effects in this study coheres with prior research, the results for certain subpopulations suggest intriguing future directions for research on music's benefits for students.

Dedication

This dissertation is dedicated to my teacher, hero, and partner, Dr. Julia Shaw.

Acknowledgements

Dissertations are commonly described as solitary ventures, but this belies the collective action necessary to produce them. This study has been enabled and shaped by many people. I owe an immeasurable debt of gratitude to my advisor Daryl Kinney, who patiently answered questions and offered prompt and insightful advice on matters big and small. Daryl, thank you for believing in me and for teaching me so much. My trajectory as a researcher was also molded by Ann O’Connell in the OSU College of Education and Human Ecology. She has generously served as my teacher, mentor, late-night-email-question-answerer, and co-presenter, for which I will always be appreciative. I am further grateful to Jan Edwards and Eugenia Costa-Giomi for agreeing to serve on my committee and for being outstanding mentors, even when they had better things to do. I have also been fortunate to learn from many other colleagues and professors at Ohio State, a place that will always be a part of my story. Several statisticians offered me technical assistance when I got stuck, even though they have never met me: Patricia Berglund from the University of Michigan and Beth Ann Griffin from the RAND Corporation were extremely generous with their time helping me use IVEware and twang, and Miguel Dorta from Stata Corporation helped me to finally get the standard deviations right. Thank you one and all.

I must also express my appreciation for my family. Since long before I started a PhD, my mother Dorothy Hastings, my father Ronald Shaw, my stepmother Elaine Shaw, and my father-in-law Dennis Dolken have helped and encouraged me in countless ways. My brilliant, loving, hilarious, musical, and energetic children Reid and Mary Claire have also supported the writing of this “distertation.” They are both voracious readers and I hope it meets their standards. Finally, if there is to be an “acknowledgements” section, it must include an acknowledgement

that, without Julia Shaw's success and support, there is no scenario in which I would be finishing a doctorate at Ohio State. J, even though of course it's nice to have an award-winning researcher and writer in the house, it's even better to have a soul mate. Thank you for putting up with me.

Vita

1998Diploma, Huron High School, Ann Arbor, MI

2002B.M.E., Indiana University

2002–2008Music Teacher, McCutcheon High School,
Lafayette, IN

2009M.M.E., The University of Illinois

2010Certificate in Educational Administration, DePaul
University

2008–2014Department Chair and Music Teacher, Glenbard
East High School, Lombard, IL

2014–2019Lecturer in Music Education, The Ohio State
University

Publications

Shaw, B. P. (2018). *Music assessment for better ensembles*. Oxford University Press.

Fields of Study

Major Field: Music

Graduate Interdisciplinary Specialization: Quantitative Research Methods

Table of Contents

ABSTRACT	i
DEDICATION	iii
ACKNOWLEDGEMENTS	iv
VITA.....	vi
LIST OF TABLES	xviii
LIST OF FIGURES	xxvi
CHAPTER 1: INTRODUCTION.....	1
Background	1
Need for the Study	3
Purpose and Research Questions	4
Methodological Overview.....	4
Data Source	4
Sample	5
Data Analysis.....	5

Definitions.....	6
Limitations	8
Chapter Organization	12
 CHAPTER 2: REVIEW OF LITERATURE	 13
Overview	13
Academic Benefits of Musical Involvement.....	14
Early Studies.....	16
James Catterall.....	17
REAP	19
Recent Studies	20
Studies Using Causal Techniques	21
Summary.....	22
Review of Social-Emotional Constructs	23
Self-Efficacy	24
Mindset	26
Belongingness.....	27
Social-Emotional Learning.....	28
Summary.....	30
Social-Emotional Benefits of Musical Involvement.....	31
Social and Emotional Outcomes Associated with Music	31
Social and Emotional Factors Involved in Music Success and Persistence	32
Summary.....	33

Mechanisms by Which Extramusical Benefits Accrue.....	33
Transfer of Learning.....	33
Transfer From Musical Involvement.....	35
Summary.....	37
Challenges in Researching Extramusical Benefits.....	38
Methodological Limitations in Extant Literature	38
Challenges With Measurement.....	38
Other Sources of Unobserved Confounding.....	40
Summary	41
CHAPTER 3: REVIEW OF PROPENSITY SCORE LITERATURE	44
Overview	44
Causal Inference.....	45
Foundations	45
The Counterfactual Tradition	46
Selection Bias and Ignorability	47
Experimental and Observational Studies.....	49
Propensity Score Foundations.....	54
The Theory of Matching.....	54
Defining the Propensity Score.....	55
The Stable Unit-Treatment Value Assumption	57
Propensity Score Estimation	57
Covariate Selection.....	57

Estimation Using Logit Models.....	59
Estimation Using Nonparametric Methods	60
CART and Ensemble Methods	61
Generalized Boosted Modeling	62
Covariate Balance	63
The Region of Common Support	64
Analytic Approaches Using Estimated Propensity Scores	65
Matching	65
Subclassification	67
Weighting	68
Weighting With Complex Samples	72
Sensitivity Analysis	73
Summary	75
 CHAPTER 4: RESEARCH METHODS AND DESIGN	 77
Overview	77
Data	77
Data Source	77
Overview	77
Waves of the HSLS:09	78
Foci of the HSLS:09	78
Sampling Procedure.....	79
High School Transcripts	80

Restricted-Use Data.....	80
Sample Restrictions	80
Data Preparation.....	81
High School Transcripts	82
Missing Data.....	86
Overview	86
Brief Review of Missing Data Methods	87
Missing Data Strategy in This Study.....	88
Analytic Strategy.....	90
Attenuating Selection Bias	90
Plausibility of Propensity Score Assumptions	91
Strong Ignorability.....	91
Stable Unit-Treatment Value Assumption	94
Consequences of Violated Assumptions	94
Propensity Score Estimation.....	95
Measures.....	96
Academic Outcomes.....	96
Social-Emotional Outcomes	96
Analysis Using Propensity Score Weighting	99
Regression Models	100
Procedures for Incorporating the Complex Sampling Design.....	101

CHAPTER 5: RESULTS	103
Propensity Score Estimation and Diagnostics	103
Propensity Score Results for Choral Music.....	104
Propensity Score Results for Instrumental Music	108
Propensity Score Results for General Music	112
Descriptive Statistics.....	116
Regression Predictors	118
Choral Music	118
Instrumental Music	120
General Music	122
Sustained Involvement	122
Comparisons Between Groups	124
Outcomes	130
Choral Music	130
Instrumental Music	131
General Music	132
Sustained Involvement	133
Comparisons Between Groups	134
Social-Emotional Variables Measured Only in Ninth Grade	138
Choral Music	139
Instrumental Music	140
General Music	140
Sustained Enrollment.....	141

Comparisons Between Groups	142
Regression Results for Academic Outcomes	144
Choral Music	145
Main Effects	145
Achievement in Eighth Grade	146
Sex	148
Race	150
Instrumental Music	152
Main Effects	153
Achievement in Eighth Grade	156
Sex	157
Race	161
General Music	161
Main Effects	161
Achievement in Eighth Grade	161
Sex	162
Race	163
Regression Results for Social-Emotional Outcomes	163
Choral Music	163
Main Effects	163
Achievement in Eighth Grade	164
Sex	165
Race	166

Instrumental Music	168
Main Effects	169
Achievement in Eighth Grade	172
Sex	173
Race	177
General Music	177
Regression Results for Base-Year Social-Emotional Variables	177
Music Groups Compared With Control Groups.....	177
Choral Music	179
Instrumental Music	180
Comparisons Between Ensemble Music Students According to Duration of	
Enrollment	180
Summary.....	181
CHAPTER 6: DISCUSSION	183
Review of Purpose	183
Review of Methodology.....	183
Summary of Findings.....	184
Research Questions 1 and 3: Academic Outcomes	184
College Examination Mathematics Scores	184
College Examination Reading Scores	186
Mathematics Reasoning.....	187
Summary.....	188

Research Questions 2 and 3: Social-Emotional Outcomes	189
School Motivation	190
Mathematics Self-Efficacy	190
Science Self-Efficacy	191
Summary.....	191
Discussion	192
Implications for Music Education Advocacy	194
Implications for Future Research on Extramusical Benefits	195
Unobserved Within-Person Variables	196
Unobserved Features of Individual Schools' Music Instruction	198
Unobserved Music Study	199
Emphasis on Subpopulations.....	200
Propensity Score Adjustment and Regression Adjustment	205
Null Hypothesis Significance Testing	208
Specificity and Breadth of Dependent Variables	209
The Unanswered Question	210
Summary.....	211
Conclusions	212
REFERENCES	214
APPENDIX A: VARIABLES USED FOR PROPENSITY SCORE ESTIMATION	259
APPENDIX B: MISSINGNESS RATES.....	267

APPENDIX C: STATA PROGRAMS FOR PROCESSING HSLs:09 TRANSCRIPT

DATA	273
qrec	273
x12	274
cors	275
fix12	276
swap	277
orphan	280
noswitch	281
wordlist	281
cleanup	282
sturedo	283
sked	283

APPENDIX D: STATA PROGRAMS FOR CALCULATING THE EFFECT SIZE..... 284

getsvysd	284
getsvyg	285

APPENDIX E: REGRESSION TABLES 286

Regression Tables for Academic Outcomes	286
Choral Music	286
Instrumental Music	292
General Music	298

Regression Tables for Social-Emotional Outcomes	301
Choral Music	301
Instrumental Music	307
General Music	313
Regression Tables for Base-Year Social-Emotional Variables	316

List of Tables

Table 4.1. Initial Course Classification Information	83
Table 4.2. Dependent Measures	97
Table 5.1. Sample Sizes.....	118
Table 5.2. Regression Predictors for 1-2 Years of Choral Music	119
Table 5.3. Regression Predictors for 3-4 Years of Choral Music	120
Table 5.4. Regression Predictors for 1-2 Years of Instrumental Music	121
Table 5.5. Regression Predictors for 3-4 Years of Instrumental Music	121
Table 5.6. Regression Predictors for 1-2 Years of General Music.....	122
Table 5.7. Choral Music Predictor Comparison.....	123
Table 5.8. Instrumental Music Predictor Comparison.....	124
Table 5.9. Outcomes for 1-2 Years of Choral Music	130
Table 5.10. Outcomes for 3-4 Years of Choral Music	131
Table 5.11. Outcomes for 1-2 Years of Instrumental Music	132
Table 5.12. Outcomes for 3-4 Years of Instrumental Music	132
Table 5.13. Outcomes for 1-2 Years of General Music.....	133
Table 5.14. Choral Music Outcome Comparison.....	134
Table 5.15. Instrumental Music Outcome Comparison.....	134
Table 5.16. Belongingness and Engagement for 1-2 Years of Choral Music	139

Table 5.17. Belongingness and Engagement for 3-4 Years of Choral Music	139
Table 5.18. Belongingness and Engagement for 1-2 Years of Instrumental Music	140
Table 5.19. Belongingness and Engagement for 3-4 Years of Instrumental Music	140
Table 5.20. Belongingness and Engagement for 1-2 Years of General Music	141
Table 5.21. Choral Music Belongingness and Engagement Comparison	141
Table 5.22. Instrumental Music Belongingness and Engagement Comparison	142
Table 5.23. Academic Conditional Mean Comparisons for 1-2 Years of Choral Music	145
Table 5.24. Academic Conditional Mean Comparisons for 3-4 Years of Choral Music	146
Table 5.25. Academic Conditional Mean Comparisons for 1-2 Years of Instrumental Music	152
Table 5.26. Academic Conditional Mean Comparisons for 3-4 Years of Instrumental Music	153
Table 5.27. Academic Conditional Mean Comparisons for 1-2 Years of General Music	161
Table 5.28. Social-Emotional Conditional Mean Comparisons for 1-2 Years of Choral Music	164
Table 5.29. Social-Emotional Conditional Mean Comparisons for 3-4 Years of Choral Music	164
Table 5.30. Social-Emotional Conditional Mean Comparisons for 1-2 Years of Instrumental Music	168
Table 5.31. Social-Emotional Conditional Mean Comparisons for 3-4 Years of Instrumental Music	169
Table 5.32. Social-Emotional Conditional Mean Comparisons for 1-2 Years of General Music	177

Table 5.33. Unadjusted Regression Results Summary: 9 th Grade School Belongingness and Engagement	178
Table 5.34. Social-Emotional Baseline Conditional Mean Comparisons Between Students Enrolled for 3-4 Years and 1-2 Years.....	181
Table 6.1. Propensity Score Adjusted Results Summary: College Entrance Examination Mathematics Score	185
Table 6.2. Propensity Score Adjusted Results Summary: College Entrance Examination Reading Score.....	187
Table 6.3. Propensity Score Adjusted Results Summary: Mathematics Reasoning Score	188
Table 6.4. Propensity Score Adjusted Results Summary: School Motivation	189
Table 6.5. Propensity Score Adjusted Results Summary: Mathematics Self-Efficacy	190
Table 6.6. Propensity Score Adjusted Results Summary: Science Self-Efficacy	191
Table 6.7. Propensity Score Adjusted Subpopulation Summary: Academic Outcomes by Eighth Grade Achievement	202
Table 6.8. Propensity Score Adjusted Subpopulation Summary: Academic Outcomes by Sex	202
Table 6.9. Propensity Score Adjusted Subpopulation Summary: Academic Outcomes by Race	203
Table 6.10. Propensity Score Adjusted Subpopulation Summary: Social-Emotional Outcomes by Eighth Grade Achievement	204
Table 6.11. Propensity Score Adjusted Subpopulation Summary: Social-Emotional Outcomes by Sex	205

Table 6.12. Propensity Score Adjusted Subpopulation Summary: Social-Emotional	
Outcomes by Race	205
Table 6.13. Selected Main Effects With and Without Regression and Propensity Score	
Adjustment	207
Table 6.14. Subpopulation Effect Differences Between Unadjusted and Adjusted	
Regression Results.....	208
Table A.1. Variables Used for Propensity Score Estimation: Demographics	259
Table A.2. Variables Used for Propensity Score Estimation: Socioeconomic Status	260
Table A.3. Variables Used for Propensity Score Estimation: Time Use.....	261
Table A.4. Variables Used for Propensity Score Estimation: Activities.....	262
Table A.5. Variables Used for Propensity Score Estimation: School Attitudes	263
Table A.6. Variables Used for Propensity Score Estimation: School Behaviors	264
Table A.7. Variables Used for Propensity Score Estimation: School and Postsecondary	
Expectations	264
Table A.8. Variables Used for Propensity Score Estimation: Prior Academic	
Achievement.....	265
Table A.9. Variables Used for Propensity Score Estimation: School-Level Covariates.....	265
Table B.1. Missingness Rates: Student-Level Covariates	267
Table B.2. Missingness Rates: School-Level Covariates	272
Table E.1. Regression Results for College Examination Mathematics Scores for 1-2	
Years of Choral Music.....	286
Table E.2. Regression Results for College Examination Mathematics Scores for 3-4	
Years of Choral Music.....	287

Table E.3. Regression Results for College Examination Reading Scores for 1-2 Years of	
Choral Music	288
Table E.4. Regression Results for College Examination Reading Scores for 3-4 Years of	
Choral Music	289
Table E.5. Regression Results for Mathematics Reasoning Scores for 1-2 Years of Choral	
Music	290
Table E.6. Regression Results for Mathematics Reasoning Scores for 3-4 Years of Choral	
Music	291
Table E.7. Regression Results for College Examination Mathematics Scores for 1-2	
Years of Instrumental Music	292
Table E.8. Regression Results for College Examination Mathematics Scores for 3-4	
Years of Instrumental Music	293
Table E.9. Regression Results for College Examination Reading Scores for 1-2 Years of	
Instrumental Music	294
Table E.10. Regression Results for College Examination Reading Scores for 3-4 Years of	
Instrumental Music	295
Table E.11. Regression Results for Mathematics Reasoning Scores for 1-2 Years of	
Instrumental Music	296
Table E.12. Regression Results for Mathematics Reasoning Scores for 3-4 Years of	
Instrumental Music	297
Table E.13. Regression Results for College Examination Mathematics Scores for 1-2	
Years of General Music	298

Table E.14. Regression Results for College Examination Reading Scores for 1-2 Years of General Music	299
Table E.15. Regression Results for Mathematics Reasoning Scores for 1-2 Years of General Music	300
Table E.16. Regression Results for School Motivation for 1-2 Years of Choral Music	301
Table E.17. Regression Results for School Motivation for 3-4 Years of Choral Music	302
Table E.18. Regression Results for Mathematics Self-Efficacy for 1-2 Years of Choral Music	303
Table E.19. Regression Results for Mathematics Self-Efficacy for 3-4 Years of Choral Music	304
Table E.20. Regression Results for Science Self-Efficacy for 1-2 Years of Choral Music	305
Table E.21. Regression Results for Science Self-Efficacy for 3-4 Years of Choral Music	306
Table E.22. Regression Results for School Motivation for 1-2 Years of Instrumental Music	307
Table E.23. Regression Results for School Motivation for 3-4 Years of Instrumental Music	308
Table E.24. Regression Results for Mathematics Self-Efficacy for 1-2 Years of Instrumental Music	309
Table E.25. Regression Results for Mathematics Self-Efficacy for 3-4 Years of Instrumental Music	310
Table E.26. Regression Results for Science Self-Efficacy for 1-2 Years of Instrumental Music	311

Table E.27. Regression Results for Science Self-Efficacy for 3-4 Years of Instrumental Music	312
Table E.28. Regression Results for School Motivation for 1-2 Years of General Music	313
Table E.29. Regression Results for Mathematics Self-Efficacy for 1-2 Years of General Music	314
Table E.30. Regression Results for Science Self-Efficacy for 1-2 Years of General Music	315
Table E.31. Regression Results for 9 th Grade School Belongingness for 1-2 Years of Choral Music	316
Table E.32. Regression Results for 9 th Grade School Engagement for 1-2 Years of Choral Music	317
Table E.33. Regression Results for 9 th Grade School Belongingness for 3-4 Years of Choral Music	318
Table E.34. Regression Results for 9 th Grade School Engagement for 3-4 Years of Choral Music	319
Table E.35. Regression Results for 9 th Grade School Belongingness for 1-2 Years of Instrumental Music	320
Table E.36. Regression Results for 9 th Grade School Engagement for 1-2 Years of Instrumental Music	321
Table E.37. Regression Results for 9 th Grade School Belongingness for 3-4 Years of Instrumental Music	322
Table E.38. Regression Results for 9 th Grade School Engagement for 3-4 Years of Instrumental Music	323

Table E.39. Regression Results for 9 th Grade School Belongingness for 1-2 Years of	
General Music	324
Table E.40. Regression Results for 9 th Grade School Engagement for 1-2 Years of	
General Music	325
Table E.41. Regression Results for 9 th Grade School Engagement for 3-4 and 1-2 Years	
of Choral Music	326
Table E.42. Regression Results for 9 th Grade School Belongingness for 3-4 and 1-2 Years	
of Choral Music	327
Table E.43. Regression Results for 9 th Grade School Engagement for 3-4 and 1-2 Years	
of Instrumental Music	328
Table E.44. Regression Results for 9 th Grade School Belongingness for 3-4 and 1-2 Years	
of Instrumental Music	329

List of Figures

Figure 5.1. Propensity Score Overlap Plot for 1-2 Years of Choral Music.....	105
Figure 5.2. Absolute Standardized Mean Difference (d_x) Plot for 1-2 Years of Choral Music	106
Figure 5.3. Propensity Score Overlap Plot for 3-4 Years of Choral Music.....	107
Figure 5.4. Absolute Standardized Mean Difference (d_x) Plot for 3-4 Years of Choral Music	108
Figure 5.5. Propensity Score Overlap Plot for 1-2 Years of Instrumental Music	109
Figure 5.6. Absolute Standardized Mean Difference (d_x) Plot for 1-2 Years of Instrumental Music	110
Figure 5.7. Propensity Score Overlap Plot for 3-4 Years of Instrumental Music	111
Figure 5.8. Absolute Standardized Mean Difference (d_x) Plot for 3-4 Years of Instrumental Music	112
Figure 5.9. Propensity Score Overlap Plot for 1-2 Years of General Music.....	113
Figure 5.10. Absolute Standardized Mean Difference (d_x) Plot for 1-2 Years of General Music	114
Figure 5.11. Propensity Score Overlap Plot for 3-4 Years of General Music.....	115
Figure 5.12. Absolute Standardized Mean Difference (d_x) Plot for 3-4 Years of General Music	116

Figure 5.13. Sample Comparison: Proportion of Students Identified as White	126
Figure 5.14. Sample Comparison: Proportion of Students Identified as Male.....	126
Figure 5.15. Sample Comparison: Mean Socioeconomic Status	127
Figure 5.16. Sample Comparison: Mean 9 th Grade Mathematics Reasoning Score	127
Figure 5.17. Sample Comparison: Mean 9 th Grade Mathematics Self-Efficacy Score	128
Figure 5.18. Sample Comparison: Proportion Receiving Low Grade in 8 th Grade Mathematics	128
Figure 5.19. Sample Comparison: Mean 9 th Grade Science Self-Efficacy Score	129
Figure 5.20. Sample Comparison: Proportion Receiving Low Grade in 8 th Grade Science	129
Figure 5.21. Sample Comparison: Unadjusted College Examination Mathematics Score	135
Figure 5.22. Sample Comparison: Unadjusted College Examination Reading Score	136
Figure 5.23. Sample Comparison: Unadjusted 11 th Grade Mathematics Reasoning Score	136
Figure 5.24. Sample Comparison: Unadjusted 11 th Grade School Motivation Score	137
Figure 5.25. Sample Comparison: Unadjusted 11 th Grade Mathematics Self-Efficacy Score	137
Figure 5.26. Sample Comparison: Unadjusted 11 th Grade Science Self-Efficacy Score	138
Figure 5.27. Sample Comparison: Unadjusted 9 th Grade School Belongingness Score	143
Figure 5.28. Sample Comparison: Unadjusted 9 th Grade School Engagement Score.....	143
Figure 5.29. Predicted College Entrance Examination Mathematics Scores by 8 th Grade Mathematics Achievement for 1-2 Years of Choral Music.....	147
Figure 5.30. Predicted College Entrance Examination Mathematics Scores by 8 th Grade Mathematics Achievement for 3-4 Years of Choral Music.....	147

Figure 5.31. Predicted College Entrance Examination Mathematics Scores by Sex for 3-4	
Years of Choral Music.....	148
Figure 5.32. Predicted College Entrance Examination Reading Scores by Sex for 3-4	
Years of Choral Music.....	149
Figure 5.33. Predicted Mathematics Reasoning Scores by Sex for 3-4 Years of Choral	
Music	149
Figure 5.34. Predicted Mathematics Reasoning Scores by Race for 1-2 Years of Choral	
Music	150
Figure 5.35. Predicted Mathematics Reasoning Scores by Race for 3-4 Years of Choral	
Music	151
Figure 5.36. Predicted College Entrance Examination Reading Scores by Race for 3-4	
Years of Choral Music.....	151
Figure 5.37. Predicted College Entrance Examination Reading Scores for 1-2 Years of	
Instrumental Music	154
Figure 5.38. Predicted College Entrance Examination Reading Scores for 3-4 Years of	
Instrumental Music	154
Figure 5.39. Predicted Mathematics Reasoning Scores for 1-2 Years of Instrumental	
Music	155
Figure 5.40. Predicted Mathematics Reasoning Scores for 3-4 Years of Instrumental	
Music	155
Figure 5.41. Predicted College Entrance Examination Mathematics Scores by 8 th Grade	
Mathematics Achievement for 1-2 Years of Instrumental Music	156

Figure 5.42. Predicted College Entrance Examination Reading Scores by 8 th Grade	
Mathematics Achievement for 3-4 Years of Instrumental Music	157
Figure 5.43. Predicted College Entrance Examination Mathematics Scores by Sex for 1-2	
Years of Instrumental Music	158
Figure 5.44. Predicted College Entrance Examination Mathematics Scores by Sex for 3-4	
Years of Instrumental Music	158
Figure 5.45. Predicted College Entrance Examination Reading Scores by Sex for 1-2	
Years of Instrumental Music	159
Figure 5.46. Predicted College Entrance Examination Reading Scores by Sex for 3-4	
Years of Instrumental Music	159
Figure 5.47. Predicted 11 th Grade Mathematics Reasoning Scores by Sex for 1-2 Years of	
Instrumental Music	160
Figure 5.48. Predicted 11 th Grade Mathematics Reasoning Scores by Sex for 3-4 Years of	
Instrumental Music	160
Figure 5.49. Predicted 11 th Grade Mathematics Reasoning Scores by 8 th Grade	
Mathematics Achievement for 1-2 Years of General Music	162
Figure 5.50. Predicted College Entrance Examination Reading Scores by Sex for 1-2	
Years of General Music	163
Figure 5.51. Predicted 11 th Grade Mathematics Self-Efficacy by 8 th Grade Mathematics	
Achievement for 3-4 Years of Choral Music	165
Figure 5.52. Predicted 11 th Grade Science Self-Efficacy by 8 th Grade Mathematics	
Achievement for 3-4 Years of Choral Music	166

Figure 5.53. Predicted 11 th Grade Science Self-Efficacy by Race for 1-2 Years of Choral	
Music	167
Figure 5.54. Predicted 11 th Grade Science Self-Efficacy by Race for 3-4 Years of Choral	
Music	167
Figure 5.55. Predicted 11 th Grade School Motivation for 1-2 Years of Instrumental Music	170
Figure 5.56. Predicted 11 th Grade School Motivation for 3-4 Years of Instrumental Music	170
Figure 5.57. Predicted 11 th Grade Science Self-Efficacy for 1-2 Years of Instrumental	
Music	171
Figure 5.58. Predicted 11 th Grade Science Self-Efficacy for 3-4 Years of Instrumental	
Music	171
Figure 5.59. Predicted 11 th Grade Mathematics Self-Efficacy by 8 th Grade Mathematics	
Achievement for 3-4 Years of Instrumental Music	172
Figure 5.60. Predicted 11 th Grade Science Self-Efficacy by 8 th Grade Science	
Achievement for 3-4 Years of Instrumental Music	173
Figure 5.61. Predicted 11 th Grade School Motivation by Sex for 1-2 Years of	
Instrumental Music	174
Figure 5.62. Predicted 11 th Grade School Motivation by Sex for 3-4 Years of	
Instrumental Music	174
Figure 5.63. Predicted 11 th Grade Mathematics Self-Efficacy by Sex for 1-2 Years of	
Instrumental Music	175
Figure 5.64. Predicted 11 th Grade Mathematics Self-Efficacy by Sex for 3-4 Years of	
Instrumental Music	175

Figure 5.65. Predicted 11th Grade Science Self-Efficacy by Sex for 1-2 Years of

Instrumental Music 176

Figure 5.66. Predicted 11th Grade Science Self-Efficacy by Sex for 3-4 Years of

Instrumental Music 176

Chapter 1: Introduction

Background

The relationships between music, emotion, and cognition have been debated and studied since Plato and Aristotle (Regelski, 1998; Schoen-Nazzaro, 1978). These debates have intensified since the “Mozart effect” (Rauscher et al., 1993, 1995) received a great deal of attention in mainstream American culture. Even though the “Mozart effect” studies were criticized for a lack of external validity and failure to replicate (e.g., Chabris, 1999; Jenkins, 2001), researchers continue to link musical engagement and subsequently enhanced cognitive and spatial abilities (e.g., Kraus & Strait, 2015; Moreno et al., 2011; Schellenberg, 2004; Tierney & Kraus, 2013). Such links may mean that music instruction supports cognition necessary for academic achievement. However, there is currently no scientific consensus on the existence of enduring cognitive benefits associated with music performance or listening.

While investigations of music’s cognitive benefits are relevant and ongoing, they do not directly address the question of whether school music improves achievement in mathematics, science, or other subjects. There are few rigorous large-scale studies of this topic in the literature. Secondary music students have been found to be higher-achieving than their peers who were not enrolled in music (e.g., Fitzpatrick, 2006; Gouzouasis et al., 2007; Guhn et al., 2020; Kinney, 2008; Vaughn & Winner, 2000). However, music students, particularly instrumental music students, are also less likely to come from underserved racial and socioeconomic communities (Elpus & Abril, 2011, 2019), which raises the specter of selection bias in these studies.

Multilevel analyses of the NELS:88 (Miksza, 2007; Southgate & Roscigno, 2009) and ELS:2002 (Miksza, 2010) found that music students had higher test scores after including socioeconomic status and race as predictors in regression models, although they did not control for confounding due to unobserved factors such as students' personality characteristics or parents' involvement with music. In Elpus' (2013b) analysis of the ELS:2002, music students' SAT score advantage was moderated to nonsignificance after controlling for within-person fixed effects. Elpus observed, "Further research using U.S. Department of Education transcript data would help the field" (p. 17).

In addition to traditional academic outcomes, schooling in the U.S. is increasingly focusing on a set of nonacademic abilities including self-efficacy, self-regulation, and resilience that has collectively come to be popularly known as social-emotional learning (SEL). Noncognitive abilities under the SEL umbrella have been shown to predict improved academic performance (e.g., Pajares, 2003; Taylor et al., 2017; Zimmerman et al., 1992). Research is also beginning to connect school music instruction and improved SEL abilities (Hallam, 2010; Winner et al., 2013). However, transfer of abilities cultivated in a music class to other situations is not straightforward (Bransford & Schwartz, 1999; Burton et al., 2000). For example, Foster and Jenkins (2017) compared students who did and did not take private music lessons and found that selection entirely explained music students' higher performance on social-emotional measures. Although this finding conforms with Elpus' (2013b) results, more research is needed in this area. No extant large-scale research tests the hypotheses that school music builds social-emotional abilities that are transferable to other academic endeavors, or that SEL mediates the relationship between music and academic achievement.

Need for the Study

Numerous authors (e.g., Brenner, 2010; Demorest & Morrison, 2000; Eisner, 1999a, 2002; Reimer, 1999; C. West & Clauhs, 2015) warn of risks posed by disingenuous or ill-informed efforts to advocate for music education based on its purported extramusical benefits. Nevertheless, arts advocates and music teachers continue to rely on ostensible extramusical benefits to justify and promote music programs. Even professional organizations such as the National Association for Music Education (2017) and Americans for the Arts (Cohen, 2017) prominently tout links between music and academic achievement on their websites. If rigorous research can establish a sound empirical basis upon which to base these advocacy efforts, it would clearly be an asset to the profession and to students who would benefit from programs designed to leverage music's cognitive or social-emotional benefits. Conversely, if no such link exists, dissemination of nationally representative research to accompany the findings of Elpus (2013b) and Winner and Cooper (2000) may also encourage arts advocates to focus their efforts elsewhere.

Beyond its implications for educators and policymakers, this research has the potential to contribute to the methodological literature pertaining to propensity score analysis in education research. Even though propensity score analysis is a natural fit for many educational research applications involving nonequivalent comparison groups, there is not a large literature base surrounding their application to national databases such as those provided by the National Center for Education Statistics (NCES). I searched Google Scholar for articles citing the recent literature on the use of propensity scores with complex samples (Dugoff et al., 2014; Lenis et al., 2017; Ridgeway et al., 2015). The published studies incorporating these techniques were overwhelmingly in health sciences. Only five (Ansari & Pianta, 2018; Hogrebe & Strietholt,

2016; Newman et al., 2016; Wagner et al., 2016, 2017) were in education journals. None of the education studies used adaptive nonparametric propensity score estimation, which deserves consideration by educational researchers.

Purpose and Research Questions

The purpose of this study was to examine the association between high school music participation, academic achievement, and social-emotional learning, using the nationally representative data in the High School Longitudinal Study of 2009, abbreviated HSLS:09 (Duprey et al., 2018; Ingels et al., 2011, 2013, 2015). Three research questions guided the investigation.

1. Is there an association between high school music coursetaking and academic achievement, as measured by test scores?
2. Is there an association between high school music coursetaking and social-emotional outcomes such as self-efficacy, belongingness, or motivation?
3. Do these relationships differ by type of music studied or the duration of study?

Methodological Overview

Data Source

The primary data source for this study was the HSLS:09, which was ideal because it integrated detailed coursetaking data with multiple academic and social-emotional measures. The HSLS:09 was a large undertaking by the U.S. Department of Education's National Center for Educational Statistics, involving a multistage stratified sample of over 25,000 high school students from nearly 1,000 schools across America who entered ninth grade in the fall of 2009. The sampling procedure was designed to be representative of the population of American high school students. Although they require additional steps in data analysis (Heeringa et al., 2017;

Stapleton & Thomas, 2008), nationally representative datasets alleviate concerns about unobserved bias and lack of generalizability when samples are derived locally. In addition to the HSLS:09's individual- and school-level data, the final dataset also incorporated school-level information from NCES's Common Core of Data (Chen et al., 2012).

Sample

The analytic sample for this study included only students who attended the same “regular” public school for four years and who had complete transcripts in the database. The final sample size (rounded in accordance with NCES guidelines) was $N=8,800$ cases representing a population of over 2 million students. I handled missing data with multiple imputation using fully conditional specification in IVEware (Raghunathan et al., 2016) to generate complete cases for analysis and avoid bias resulting from listwise deletion and less rigorous missing data methods (J. W. Graham, 2009; Schafer & Graham, 2002). Transcripts were initially coded using the School Courses for the Exchange of Data system, or SCED (Bradby et al., 2007). I undertook substantial data preparation in order to ensure that music courses were accurately classified. For simplicity and accuracy, I grouped the long list of music courses that students had taken into three types: instrumental music, vocal music, and general music. For each type of music, I created a “short-term” group and a “sustained” group according to whether the students enrolled in that type of music for 1-2 years or 3-4 years. I also created a control group of students who did not enroll in music courses in high school.

Data Analysis

I used propensity scores (Guo & Fraser, 2015; Rosenbaum, 2010; Rubin, 2006) to control for the documented selection biases associated with enrollment in school music. The broad range of covariates in the HSLS:09, including eighth-grade pretest information, served to attenuate the

possibility of omitted variable bias (Hallberg et al., 2018; Steiner et al., 2010; Stuart, 2010). I estimated the propensity scores using boosted regression with the R package `twang` (Ridgeway et al., 2017). I then used the propensity scores in a series of weighted OLS regression models to estimate the effects of music participation on several outcomes, conditionally independent from selection bias (Harder et al., 2010; Ridgeway & McCaffrey, 2007).

The outcomes I modeled were SAT and ACT scores, a test of mathematics reasoning, mathematics and science self-efficacy, and school motivation. I also conducted a limited analysis of students' 9th grade school engagement and belongingness. The estimands of interest in this analysis were average treatment effects on the treated groups of student musicians, or ATTs (Morgan & Winship, 2007). I followed current recommendations for propensity score analysis with multistage samples (e.g., Dugoff et al., 2014; Lenis et al., 2017; Ridgeway et al., 2015). Since music students exhibit differential performance across musical fields (Kinney, 2008, 2018), I repeated the analysis for each type of music, using only non-music students as the control reservoir for each. As individual coefficients are difficult to interpret in complex models, the core analytic lens was the propensity-score-adjusted predicted score comparison between the non-music control group and each group of music students.

Definitions

For the purpose of clarity, I will use the following terms and definitions consistently throughout this study:

Estimand. An estimand is a quantity of interest in a causal study adhering to the counterfactual tradition. Studies typically estimate one of three estimands. Average treatment effect (ATE) is the estimated treatment effect averaged across all cases. The average treatment effect on the treated (ATT) is the estimated treatment effect averaged across the treatment cases

only, and the average treatment effect on the untreated or control (ATC) is the estimated (counterfactual) treatment effect that would have accrued to control participants had they instead been in the treatment. In this study, the estimand of interest was the ATT, or the estimated average treatment effects of music participation for those classified as musicians in the study.

NCES dataset. Under the auspices of the U.S. Department of Education’s Institute for Education Sciences (IES), the National Center for Education Statistics (NCES) collects and analyzes data on a variety of educational topics. Most importantly for this study, NCES has conducted a series of large-scale longitudinal data collection projects utilizing multistage stratified sampling designs to yield estimates that are representative of the entire population of American students. These surveys, referred to in this study as “NCES datasets,” are the National Education Longitudinal Study of 1988 (NELS:88), Education Longitudinal Study of 2002 (ELS:2002), the Early Childhood Longitudinal Study—Kindergarten Class of 1998-99 (ECLS-K), and most recently the High School Longitudinal Study of 2009 (HSLS:09).

Participant. For the purposes of this study, a student who participated in music is defined as a student who completed at least one term of instruction with a passing grade in course classified as a music course in the School Courses for the Exchange of Data system (Bradby et al., 2007) according to the student’s coded transcript.

Non-participant. A non-participant in this study is a student who, for whatever reason, did not have at least one completed term of musical study in their coded transcript as defined above. This operationalization is deliberately not intended to characterize students as not being musicians, or not being musical. All children are musicians if they are musically engaged in ways that are meaningful to them.

Socioeconomic status. Socioeconomic status (SES) is a composite representation of one's relative position in society, operationalized as a combination of household income, educational attainment, and occupational prestige. Children's SES is generally assumed to be equal to their parents'.

Treatment. For the purposes of this study, the "treatment" was the set of experiences that occurred as a direct or indirect result of enrollment in one or more semesters of a high school music course. A "treatment group" is a particular set of high school music participants and a "control group" is a set of comparable nonparticipants (i.e., students who are similar on observable characteristics and did not enroll in any music courses in high school).

Treatment effect. A treatment effect is the estimated within-person difference in a given outcome that can be attributed to the "treatment" of participating in high school music relative to the counterfactual state of nonparticipation.

Limitations

Propensity score models remove selection bias from analyses involving nonequivalent comparison groups, conditional on the assumption that the observed covariates provide sufficient information to predict treatment selection. Even though the HSLs:09 provides a wide range of individual- and school-level covariates that may be related to selection into music programs (see Hallberg et al., 2018), the potential exists for this study to suffer from unobserved confounding. Music students are, on the whole, systematically different from non-music students (Elpus & Abril, 2011, 2019; Kinney, 2010, 2018). Likewise, music programs are not identical across schools (Bates, 2012; Costa-Giomi, 2008; Costa-Giomi & Chappell, 2007; Fitzpatrick, 2011; Miksza, 2013; Miksza & Gault, 2014). Even though researchers have examined the results of the complex phenomena contributing to individual students' selection into and persistence through

elective music programs (McPherson & Hendricks, 2010; Schmidt, 2007; Sichivitsa, 2007), it is not sufficiently understood to rule out the possibility that unobserved variables related to selection or persistence biased the results.

One important confounder is unobserved music participation. Curricular music participation in each semester of high school is observed, but middle-school music participation was not included in the dataset. A large majority of school music students begin music study in middle school and are continuously enroll thereafter. It is therefore reasonable to assume that students with a high propensity to enroll in high school music, but who did not, were also more likely to have enrolled in music previously than their control-group peers with lower propensity.

Certainly, the dataset contains numerous participants who took music courses before high school, potentially for three or more years, and ceased taking music prior to data collection. If a substantial portion of music's nonmusical benefits accrue during the first years of study (Costa-Giomi, 1999), it is possible that these students could upwardly bias the "non-musician" group's performance even though they may well be considered to be musicians in different contexts. This phenomenon is certainly plausible, particularly for instrumental music programs that often emphasize development of technique and music literacy in middle school, and tend to emphasize performance in high school. As the proportion of high school instrumental music students who did not enroll in music programs in middle school is typically small, those results are less likely to be affected. These estimates may also be confounded by students who were "non-participants" in formal school music, but engaged in limited or persistent musical study or activity that is not reflected in the dataset.

While there are advantages to using an NCES dataset such as the HSLs:09, the fact that it is not designed for music research is also a limitation. Certain variables that would be central to

this study are unmeasured. For example, if the theory is that sustained positive engagement with music peers cultivates belongingness in music class, which then transfers to a general feeling of belongingness in school, a measure of music belongingness would be desirable. Furthermore, the outcomes that do exist in the dataset are imperfect. Standardized tests have been criticized for being too narrow in scope and having cultural and gender bias (Apple, 1979; Baker & Johnston, 2010; Gipps, 1994, 1995, 1999; Kane, 2013; Wiliam, 2000; Willie, 2001). Certain putative benefits of arts instruction, many of them dispositional in nature, and many of them regarded by practitioners as among the most important, are not represented in the HSLs:09 because they are not easily quantifiable (see Chapman, 1982, 2001; K. F. McCarthy, 2004; Stevenson & Deasy, 2005). Furthermore, this study relies on the validity and psychometric properties of several measures that I did not create.

The use of transcript data rather than direct observation to operationalize “music enrollment” also creates limitations. The exclusion of students who participated in more than one type of music, while necessary to ensure unconfoundedness, implies that some students—likely among the more musically motivated and successful—were not represented in any of the music “treatment” groups. This could have biased the music students’ scores, presumably downward, although hopefully the small number of students who participated in multiple types of music means that any such bias was minimal. The exclusion from the sample of students who changed schools over the course of the study, and those who did not take college entrance examinations, is also a potential source of bias. This bias, while a limitation, would only affect the results insofar as student transience was independent of the other predictor variables used to generate the propensity scores. The sample exclusion of students who did not take a college entrance examination is also a limitation. Like the potential bias from transient students, it is reasonable

(or, at least, plausible) that the factors leading students to take or not take a college entrance examination are adequately explained by the other predictors, but this is also unknown.

Relatedly, the unknown reasons behind individual students' enrollment decisions may also have affected the results of this study. The reduction of music participants into two groups: one for 1-2 years of study, and the other for 3-4 years of study, belies the complexity and heterogeneity involved with the reasons that students choose their high school classes. On the whole, students who only participated for 1-2 years presumably ceased their musical participation because they were less attached to, successful in, or motivated by their school's music programs than students who enrolled for 3-4 years. However, this is neither supported or refuted by the data. Just as some students probably participated halfheartedly for one year in order to fulfill a graduation requirement, others may have reluctantly dropped out of music in order to pursue an overload of college prep coursework or to take only courses with honors-weighted grades in pursuit of an impressive class rank. It is also possible that differences between students' experiences of middle school music and high school music were related to which students continued and which did not.

Even though NCES uses rigorous design and sustained effort to ensure that its sampling and replication procedures provide reliable national estimates, sampling variability is an ever-present threat (Lohr, 1999). In particular, since the HSLS:09 was not a music-focused investigation, no efforts were made to ensure that the sampled schools' music programs or students were nationally representative. In addition, music instruction cannot be assumed to have been implemented with consistent fidelity across the sampled schools, or assumed to remain consistent regardless of which and how many students were enrolled, or to resonate equally with each of the students in the sample. Prior research (Johnson & Memmott, 2006) has suggested

that differences between music programs may result in differential accrual of putative extramusical benefits.

Chapter Organization

The following chapters detail the study's background, methodology, findings, and implications. Chapter 2 contains a review of the literature on nonmusical outcomes associated with music instruction and the challenges faced by researchers investigating them. Chapter 3 reviews propensity score analysis, the central methodological technique used in this study, while Chapter 4 outlines the specific methods used for data analysis. Results of the analyses are presented in Chapter 5, with discussion, implications, and conclusions in Chapter 6.

Chapter 2: Review of Literature

Overview

The literature on the associations between music, various types of cognition, and resultant connections to schooling is vast and varied, but has not provided conclusive answers to core questions. Researchers have attempted to quantify the effects of musical engagement in many of its forms, including listening, playing, studying in private lessons, learning in elementary school classes, and participating as a member of a school ensemble.¹ The outcomes examined in these studies may be broadly classified into three categories: cognitive and neurological, academic, and social-emotional. Each category of literature has methodological questions that have not been definitively answered.

Widely considered to be the inventor of modern mathematics, Pythagoras and his intellectual forebears in Babylon and Egypt drew heavily on musical ratios for inspiration, and music was commonly regarded as an integral element of the mind's functions (Pont, 2004; see also Schoen-Nazzaro, 1978). More recently, links between musical engagement, brain activity and development, and cognition have been closely studied since the widely known "Mozart effect" studies (Rauscher et al., 1993, 1995) became a national media sensation. These studies

¹ Notably, creative engagement with music (e.g., composition and improvisation) is underrepresented among the dimensions of musical involvement researchers have examined, even though creative production and expression are at the center of the theater/drama and visual art literature.

found that listening to ten minutes of a particular Mozart piano sonata (the sonata for two pianos in D, K. 488) resulted in undergraduate students performing better on a spatial reasoning subset of a standard IQ test than their peers who heard a piece by Philip Glass or silence. Subsequent scholarship (e.g., Chabris, 1999; Jenkins, 2001; Thompson et al., 2001) has questioned whether the “Mozart effect” is consequential or even real, with Frances Rauscher herself expressing “horror” at the resulting sensational media accounts and ensuing creation of a “Mozart effect industry” (2002, p. 267). Nevertheless, the notion that “music makes you smarter” remains indelible in the American popular consciousness—particularly among music educators (Vitale, 2011).

While the cognitive and neurological literature is beyond the scope of this investigation, several excellent reviews (Bugaj & Brenner, 2011; Costa-Giomi, 2012, 2015; Hallam, 2010; Schellenberg & Weiss, 2013) are available. This chapter is focused on studies of the academic and social-emotional benefits of school music education and associated theoretical and methodological topics. The review is organized into four sections. The first provides a review of studies connecting musical involvement to academic achievement. The second section offers an overview of several relevant constructs from educational and social psychology, and the third examines the links between these outcomes and participation in school music. The final two sections address the question of transfer and methodological issues in the study of extramusical benefits.

Academic Benefits of Musical Involvement

Although the literature contains many studies attempting to determine whether school music predicts improved academic achievement, definitive conclusions can seldom be drawn. Methodological decisions driven by convenience have led to many studies being less than

convincing, with inquiry that is limited in scope. Many studies involve a single school district or an otherwise nonrepresentative sample, limiting their generalizability. Even within school contexts, because elementary school music is often offered uniformly throughout a school or district, its effects are difficult to isolate without rigorous experimental or quasi-experimental procedures (see Murnane & Willett, 2011; Shadish et al., 2002). As a result, the vast majority of studies in school settings focus on students in secondary ensemble classes. While this approach delivers a ready-made comparison group, it narrows the operational definition of music and the developmental stages at which students are observed. Early-childhood and elementary-school music, musical creativity, arts integration and multi-arts programs, secondary general music, and extracurricular musical engagement are rarely considered (cf. Kinney & Forsythe, 2005; Rabkin, 2004).

Furthermore, the nature of easily available data frequently does not allow for the use of techniques for causal inference. Students who do and do not enroll in music classes have been shown to be systematically different in a host of respects. Demographic differences are well established in the literature (Elpus & Abril, 2011, 2019; Kinney, 2010, 2018) but other important differences have also been found, such as personality (Corrigall et al., 2013; Corrigall & Schellenberg, 2015; Hille & Schupp, 2015), and direct parental involvement (McPherson, 2009; Pitts et al., 2000). These differences, known as selection bias (Hernán et al., 2004; Shadish et al., 2002), mean that additional methodological steps are necessary in order to determine the degree to which the apparent effects of music enrollment are actually the result of music enrollment and

not simply the result of pre-treatment² differences between groups. Selection bias, and ways in which researchers can address it, are described in more detail in Chapter 3.

Early Studies

Some of the earliest published studies comparing the academic achievement of music and non-music students showed no significant differences between the two groups (Groff, 1963; Kvet, 1985; Robitaille & O’Neal, 1981; Wallick, 1998). While these investigations are commendable for beginning the systematic study of music students’ academic achievement, methodological limitations lessen their current scientific utility. This line of inquiry was used by music educators to advocate for pullout music instruction by showing that pullouts did not hurt academic achievement, rather than attempting to prove that they boosted it (Gillespie, 1992; Hash, 2004).

Young (1971) and Klinedinst (1991, 1992) analyzed student records and concluded that factors such as scholastic aptitude, academic achievement, and socioeconomic status were more important predictors of enrollment and retention in instrumental music than musical aptitude. This finding was later replicated by Hash (2011). While both of these studies examined single school districts, and were not well controlled, their findings would be consistently replicated in later empirical work and foretold a central methodological challenge in this line of research. The fact that American elective music students are systematically different from their peers who do not enroll in music courses continues to bedevil attempts by researchers, policymakers, and the public to understand the causal effects of music participation.

² Consistent with the methodological literature, I use the term “pre-treatment” to refer to participant characteristics or measures prior to enrollment in the “treatment” of a school music program.

James Catterall

James S. Catterall, the late chair of the graduate education department at the University of California, Los Angeles, wrote extensively on the benefits of arts education. While Catterall was not himself a music educator by trade, his wife was a studio artist, and he undertook several research projects comparing students who did and did not participate in school-based arts programs using NCES datasets. In the monograph *Involvement in the Arts and Success in Secondary School* (1997), Catterall used data on course enrollment, extracurricular participation, and leadership positions from the National Longitudinal Study of 1988 (NELS:88) to create a composite index of arts involvement. He divided students into a “high-arts” group at the top quintile of the arts involvement index, and a corresponding “low-arts” group of students in the bottom quintile. The “high-arts” group had higher standardized test scores in both 8th and 10th grade, and had lower rates of school dropout.

In his first seminal piece (Catterall et al., 1999), the introductory chapter in the widely cited report *Champions of Change* (Fiske, 1999), Catterall again reported findings from the NELS:88. Catterall attempted to control for selection into music with a composite socioeconomic status variable, creating a “low SES” group of students in the bottom quintile of SES, and a “high SES” group of students in the uppermost quintile. Catterall found that American 8th grade band and orchestra students in the “low SES” group achieved high mathematics test scores at twice the rate of their peers who did not study instrumental music. Band and orchestra students in 12th grade continued to score higher on math tests. Later works replicated these results. In his monograph *Doing Well and Doing Good By Doing Art* (2009), Catterall observed that the arts were not only associated with increased academic performance, but also higher rates of college attendance and involvements like community service, using the

NELS:88. In an analysis of Early Childhood Longitudinal Study and National Longitudinal Survey of Youth data from NCES, Catterall, Dumais, and Hampden-Thompson (2012) found that low-SES students who studied the arts were 19 percentage points more likely than their non-arts peers to desire a bachelor's degree and 12 points more likely to complete one.

Catterall's work, while scholarly in nature, straddled the line between research and advocacy and mostly appeared in books published by arts advocacy organizations rather than peer-reviewed journals. While it feels good for arts supporters to read his studies, there are reasons to be somewhat skeptical about his results and conclusions. Catterall tended to report only positive findings, leaving readers to wonder about unreported null or negative results. Catterall did not use causal methods or any individual-level analyses, only comparisons in the aggregate such as grouping by SES. The direction of causality is unclear, and there are many possible third variables explaining the association between the arts and academic achievement, including energy level and motivation (Eisner, 1999b). High-SES students may also have engaged in more unobserved arts activities outside of school, further confounding Catterall's analyses (Southgate & Roscigno, 2009). Moreover, with a normally distributed variable such as the SES index, the top and bottom quantiles (the "tails" of the distribution) are by definition skewed and are likely to contain considerable within-group heterogeneity.

Catterall himself was aware of the nonequivalence (selection) issue, writing that "A substantial portion of the academic advantages associated with the arts should be attributed to just who has access and encouragement for the arts" (1997, p. 6, see also 1999). Catterall was an enthusiastic and unapologetic advocate and a pioneer in the analysis of nationally representative data to study the outcomes of students who participated in school-based arts programs.

REAP

In 2000, the Journal of Aesthetic Education devoted an entire special issue to extramusical benefits of the arts. Labeled the Reviewing Education in the Arts Project (REAP), it was curated by invited guest editors Ellen Winner and Lois Hetland from Harvard's Project Zero. Among the studies in the REAP special issue were meta-analyses of literature on the relationships between music and reading (Butzlaff, 2000), music and mathematics (Vaughn, 2000), music and SAT scores (Vaughn & Winner, 2000), and the arts and academic achievement in general, which carried the provocative title *Mute Those Claims: No Evidence (Yet) for a Causal Link Between Arts Study and Academic Achievement* (Winner & Cooper, 2000).

The major finding was that, in each subject area, the extant literature at the time was plagued by methodological deficiencies and mixed findings, and that few causal conclusions could be drawn about the arts' extramusical benefits. REAP is an important milestone in scholarly examination of extramusical outcomes. Coming shortly after the "Mozart Effect" craze, REAP was a much-needed reminder that systematic inquiry did not yet support—and, twenty years later, still does not support—many frequently-made claims about the benefits of music instruction. Still, as an oft-cited milestone, it framed potential extramusical benefits entirely through standardized tests in reading and math, excluding humanities, scientific reasoning, and any other pursuits (Chapman, 2001). As such, REAP represented a missed opportunity to emphasize viewing the benefits of arts instruction through a wider lens. The authors themselves recognized this limitation, writing that they "urge researchers to move beyond measuring the effects of the arts in terms of scores on paper and pencil tests, to assess how the arts affect learning in areas that are more difficult to measure, but that may well be more important" (Winner & Hetland, 2000, p. 7). Despite this sentiment, the articles in their curated

collection did not implement this recommendation, foreshadowing a major limitation of subsequent scholarship.

Recent Studies

Since the publication of REAP, study of the standardized test scores of music students has continued. Fitzpatrick (2006) analyzed data from the urban Columbus City Schools in Ohio. Similar to some of Catterall's work, Fitzpatrick analyzed aggregated rather than individual achievement data. She noted that instrumental music education students had significantly higher initial test performance, and that instrumental music students receiving free or reduced-price school lunches (FRL) scored below their musician peers who did not participate in the school lunch program.³ Fitzpatrick also observed that the test-score gap between instrumental music students who did and did not receive FRL closed over the period of music enrollment. Also using data from a large metropolitan school district in Ohio, Kinney (2008) observed significant differences between musicians and nonmusicians as well as students enrolled or not enrolled in FRL. For both 6th-grade and 8th-grade students, there was not a significant interaction between FRL status and music; i.e., participation in music did not provide additional advantages to students beyond their FRL status.

Other researchers have also recently found similar results with local datasets, albeit with small effect sizes (Cabanac et al., 2013; Gouzouasis et al., 2007; Helmrich, 2010; Johnson &

³ Researchers have frequently used FRL programs as a proxy measure of socioeconomic status in datasets that do not contain specific variables pertaining to SES (i.e., income, educational attainment, and occupational prestige). While FRL data is likely to be readily available from school districts, which may be hesitant to collect too much personal information from families, scholars (e.g., Harwell & LeBeau, 2010) have questioned the validity of this approach. Still, at times, it is the only information available.

Memmott, 2006; Wetter et al., 2009). In a longitudinal growth curve analysis of individual data from the nationally representative NELS:88, Miksza (2007) largely replicated Kinney's and Fitzpatrick's results. Miksza speculated that the narrowing of test-score gaps between instrumental music students and their nonmusical peers may have been the result of a ceiling effect, since the instrumental students were already a high-achieving population. In a similar study using the ELS:2002 (Miksza, 2010), music students' academic achievement remained significantly higher even after accounting for a range of controls, a finding echoed by sociologists Southgate and Roscigno in their (2009) analysis of the ECLS-K and NELS:88. Most recently, Guhn et al. (2020) showed a range of positive effects with a sample of over 100,000 Canadian music students after controlling for prior academic achievement, sex, "cultural background," and neighborhood socioeconomic status in regression analyses.

Studies Using Causal Techniques

The strongest foundation for establishing causality is a randomized experiment, which "breaks the link" between selection into treatment and characteristics of control and comparison groups (Ho et al., 2007). Costa-Giomi (1999, 2004) conducted a randomized experiment with an initial sample of 117 fourth-grade children. Students in the treatment group were provided with a free piano and weekly lessons with experienced teachers. Although descriptive analysis revealed higher math computation scores among the children in the treatment group, few comparisons achieved statistical significance. Only 80 students completed all of the follow-up testing, however, and some tests may have achieved significance with increased power. Children who received the piano lesson condition scored higher on general and spatial cognitive tests than the control group for the first two years of the study, but the control group's scores increased more rapidly in the third year and the piano lesson group's score advantage disappeared. Costa-Giomi

hypothesized that this may have been due to a decrease in motivation among some piano lesson participants “after the initial enthusiasm disappeared and progress in learning the piano required more effort and intense involvement” (1999, p. 208).

Studies of nonmusical outcomes using techniques for causal inference with secondary data have been slow to arrive in the literature. One notable exception is Elpus’ (2013b) analysis of the Education Longitudinal Study of 2002 (ELS:2002) using fixed-effects regression. Fixed-effects models permit stronger causal inferences through the use of differencing with longitudinal data, effectively using the individual as her or his own control (Allison, 2009). Elpus found that music students’ college admissions test scores were higher than their non-music peers. However, after controlling for between-person differences, the effects of music participation were moderated to nonsignificance. Elpus commented that his study “demonstrates that controlling for demographic factors is ‘not enough’ to reduce omitted-variables bias in comparisons of music and non-music students” (p. 15). Similar to Miksza’s and Southgate and Roscigno’s results, Yang’s (2015) study using fixed-effects regression with the German Socio-Economic Panel found that selection moderated, but did not completely attenuate, music students’ higher educational attainment.

Summary

Study after study has found an association between music engagement or participation and various academic and cognitive outcomes. Still, the research is lacking conclusive evidence that the association is causal rather than correlational, and there have also been a large number of studies that have found no linkage between music and these outcomes with even a simple set of controls. There is a need for further study that incorporates methods to reduce or control for selection bias.

Review of Social-Emotional Constructs

In schools, both teaching and learning are inherently social processes (Gehlbach, 2010; Goodenow, 1992; Zins, Weissberg, et al., 2004). Students' academic abilities are only one factor contributing to their performance. Ability, self-efficacy, formal and informal feedback, motivation, and performance are constantly interacting (Schunk, 1991). As Stanford professor Albert Bandura (1986, p. 390) observed, "People often do not behave optimally even though they know full well what to do. This is because self-referent thought mediates the relationship between knowledge and action." Accordingly, psychological aspects of teaching and learning deserve primary consideration from educators and educational researchers:

It may even be reasonably argued that teachers should pay as much attention to students' self-beliefs as to actual competence, for it is the beliefs that more accurately predict students' motivation, behavior, and future choices...Low self-efficacy, not lack of capability or skill, can be responsible for maladaptive academic behaviors, avoidance of courses and careers, and diminishing school interest and achievement. (Pajares & Schunk, 2002, p. 18)

Psychology is the study of human thought processes and behavior, and educational psychology is the application of these pursuits in the context of teaching and learning (Tuckman & Monetti, 2010). Educational psychology is a web of similar and overlapping subdisciplines, each with its own literature base. While their effects were certainly felt by participants in this investigation, constructs including self-concept (Eccles, 2009; Huang, 2011; Markus & Nurius, 1986; Marsh et al., 2018), self-regulation (Zimmerman, 1989, 2008), self-esteem (Leary, 1999; Leary et al., 1998), and other theories of motivation, such as the 2x2 achievement goal framework (e.g., Cury et al., 2006; Elliot & McGregor, 2001), self-determination theory (Deci et

al., 2001), and expectancy-value theory (e.g., Wigfield & Eccles, 2000) were not central to the analysis and are not reviewed here. Dependent measures selected by the designers of the HSLS:09 included self-efficacy, belongingness, and school motivation. This review addresses the psychological outcomes modeled in this study: self-efficacy, the closely related concept of mindset, and school belongingness. It also addresses social-emotional learning, the way that these outcomes are often grouped in current educational discourse.

Self-Efficacy

A foundational concept for the intersection of educational and social psychology is Bandura's (1986) social-cognitive theory, which he explained as people being motivated by the intersection of several forces:

In the social cognitive view people are neither driven by inner forces nor automatically shaped and controlled by external stimuli. Rather, human functioning is explained in terms of a model of triadic reciprocity in which behavior, cognitive and other personal factors, and environmental events all operate as interacting determinants of each other.

(p. 18)

In two seminal books, Bandura (1986, 1997) codified the concept of self-efficacy as a central component of his social cognitive theory. According to Bandura, "Perceived self-efficacy refers to beliefs in one's capabilities to organize and execute the courses of action required to produce given attainments" (1997, p. 3). Since people are most likely to undertake a task if they believe that they have the capacity to succeed, self-efficacy is a critical element of motivation inside and outside of school (Bandura, 1997; Schunk, 1991). Low self-efficacy leads students to avoid activities and give up in the face of obstacles, while high self-efficacy begets the kind of

goal-setting and sustained engagement that leads to long-term skill growth (Bandura, 1986, 1997; Zimmerman, 1989, 2000).

There is also a literature base connecting self-efficacy to academic achievement (e.g., Pajares, 1996; Schunk, 1991). Undergraduate students' math problem solving was better predicted by self-efficacy than factors such as math self-concept, high school math level, and college credits earned (Pajares & Miller, 1994). Using path analysis, Zimmerman, Bandura, and Martinez-Pons (1992) demonstrated that high school students' self-efficacy beliefs had direct effects on academic achievement, and also indirect effects mediated by goal setting. In a follow-up study, Zimmerman and Bandura (1994) found similar results with students in a college writing course. The social-cognitive factors in these studies accounted for 31% and 35% of the variance in the dependent variable, respectively. Contextualizing these results, Bandura explained, "self-regulatory efficacy contributes to a strong sense of academic self-efficacy, which, in turn, is positively related to subsequent academic achievement both directly and through its impact on self-set goals" (2012, p. 26), which Usher and Pajares (2008a) termed "self-efficacy for self-regulated learning." In one meta-analysis of longitudinal studies of the effect of self-beliefs on achievement, the 95% confidence interval for overall effect size was smaller at .07–.09 (Valentine et al., 2004).

The ubiquity of self-efficacy in educational discourse has led to frequent misunderstandings about its conceptual boundaries. Self-efficacy is the belief in the possibility of success with a particular task, not a global judgment about oneself, as in self-concept or self-esteem (Pajares, 2003; Schunk, 1991; Zimmerman, 2000), nor is it a nonspecific feeling of "confidence" (Pajares, 1996; Pajares & Miller, 1994). Another common misconception is that there can be some kind of "vague global measure" (Bandura, 1997, p. 38) of self-efficacy. Even

though someone's self-*concept* may include beliefs of competence with sports in general, self-*efficacy* beliefs about making a free throw are only tangentially related to beliefs about fielding a fly ball: efficacy is not "a fixed ability that one does or does not have in one's behavioral repertoire" (Bandura, 1997, p. 36). This greatly complicates both measurement of self-efficacy and interventions intended to improve it (Bandura, 2012; Usher & Pajares, 2008b). Finally, self-efficacy alone does not produce results; competence is also necessary (Bandura, 1997; Schunk, 1991). Still, there is compelling evidence that competence is not the only factor in achieving desired outcomes, and educators ignore self-efficacy at their peril.

Mindset

Bandura (1986) further noted that self-efficacy influences students' attributions following success and failure. Psychologist Carol Dweck merged the concept of self-efficacy with attribution theory and Rotter's (1966) locus of control theory in her work on students' theories of intelligence, which she later popularized as student "mindset" (Dweck, 2006). Dweck hypothesized that students holding an "entity theory" of intelligence (a "fixed mindset") believe that their abilities are inborn or otherwise attributable to factors beyond their control. In contrast, students holding an "incremental theory" of intelligence (a "growth mindset") believe that, regardless of one's initial abilities, they can be improved over time with effort. Clearly, the growth mindset is the belief system most conducive to school success. When faced with challenges, students with a fixed mindset give up (the "helpless" response), become argumentative, or change the subject (Dweck et al., 1978), while students with a growth mindset redoubled their efforts (the "mastery-oriented" response). Dweck observed, "Although children displaying the different patterns do not differ in intellectual ability, these patterns can have profound effects on cognitive performance" (Dweck, 1986, p. 1040). Although there is not a

strict dichotomy between the two mindsets, student beliefs often fall into one category or the other in a given setting (Dweck, 2006).

Dweck also hypothesized that students' achievements follow their mindsets: "These goals appear to fall into two classes: (a) *learning goals*, in which individuals seek to increase their competence, to understand or master something new, and (b) *performance goals*, in which individuals seek to gain favorable judgments of their competence or avoid negative judgments of their competence" (Dweck, 1986, p. 1040). Dweck and Leggett noted further, "In challenging achievement situations, helpless [fixed-mindset] children might be pursuing the performance goal of *proving* their ability, whereas the mastery-oriented [growth-mindset] children might be pursuing the learning goal of *improving* their ability" (Ames, 1992; Dweck & Leggett, 1988, p. 259; Wolters et al., 1996).

Even though they are distinct constructs, self-efficacy and mindset are closely related and have similar effects on student outcomes. Mindset and associated goal orientation have been demonstrated to be related to student interest and motivation (Dweck, 2002) and ultimately academic achievement (Dweck, 1986, 2006; Dweck & Leggett, 1988). Researchers and clinicians in the field of educational psychology often emphasize Bandura's view of self-efficacy, while practicing teachers and administrators are more likely to be familiar with Dweck's work. Both conceptions of students' academic self-beliefs reinforce the central role of psychology in the learning process.

Belongingness

Belongingness refers to "students' sense of being accepted, valued, included, and encouraged by others (teacher and peers) in the academic classroom setting and of feeling oneself to be an important part of the life and activity of the class" (Goodenow, 1993, p. 25). In

their seminal article, Baumeister and Leary described belongingness as the notion “that human beings have a pervasive drive to form and maintain at least a minimum quantity of lasting, positive, and significant interpersonal relationships” (1995, p. 497). Baumeister and Leary identified frequent interaction and consistent caring as two dimensions of belonging, while Lee and Robbins (1995) operationalized belonging as social connectedness and social assurance, or one’s reliance on others. Belongingness has also been described in the literature as relatedness, connectedness, social support, and feeling a sense of community (Furrer & Skinner, 2003; Osterman, 2000). In any conceptualization, the “need to belong” (Baumeister & Leary, 1995) may be particularly acute during adolescence (Goodenow, 1993).

While the personal benefits of belonging are clear, its documented academic effects are less so. Physical and emotional wellbeing are enhanced by social attachment (Baumeister & Leary, 1995). Belongingness is associated with positive attitudes toward school, school work, and teachers (Osterman, 2000). It may also provide the social support necessary to help students navigate challenges (Furrer & Skinner, 2003) and improve students’ sense of control or autonomy (Ryan et al., 1994), both of which are central features of Dweck’s work highlighted in the previous section of this review. Each of these may contribute to the most consistent finding about school belongingness, that it is associated with heightened academic motivation (Furrer & Skinner, 2003; Goodenow, 1993; Osterman, 2000; Ryan et al., 1994). Still, research is inconclusive on whether belongingness causally affects academic achievement (Gillen-O’Neel & Fuligni, 2013).

Social-Emotional Learning

A persistent issue in applications of social and educational psychology in classrooms is that the social psychology research base has not received mainstream attention from practicing

educators (Aronson, 2002). Indeed, given the central importance of students' belief patterns in education, it is puzzling why they are not treated with more urgency in schools. Many of the traditions in American schooling were codified during the behaviorist era of psychology, when the influence of Ivan Pavlov, B. F. Skinner, and G. Stanley Hall was at its apex (Pajares & Schunk, 2002). This legacy of behaviorism may be one reason that schools have been slow to focus on students' thoughts, motivations, and reactions to their environments (see also Kohn, 2013). Furthermore, in the current accountability era, easily measurable educational goals are prized above all else (Figlio & Ladd, 2015). Given that social-emotional abilities do not lend themselves well to valid measurement using standardized tests, and given that they are rarely publicly reported as school accountability data, it is no surprise that they have been underemphasized in schools relative to their potential to improve students' educational and life outcomes (Zins, Bloodworth, et al., 2004).

Recently, there has been a surge of interest in these and other nonacademic aspects of schooling, which have been collectively termed social-emotional learning (SEL). As it has become more widely known, SEL's conceptual boundaries have become less clear. While it has roots in educational psychology, "SEL" has broadened to become a somewhat omnibus term encompassing everything from growth mindset to stress management to empathy to teen pregnancy and drug use (Walberg et al., 2004). A leading voice on SEL has been the Center for Academic, Social, and Emotional Learning (CASEL) at the University of Illinois at Chicago, which has distilled five SEL competencies: self-awareness, social awareness, responsible decision-making, self-management, and relationship skills (2017). Although these concepts may each appear to represent distinct conceptual territory, the theory is that addressing noncognitive factors comprehensively is the best approach in the long run as they are often closely related and

“one-off” interventions are generally not effective (Zins, Bloodworth, et al., 2004; see also M. T. Greenberg et al., 2003).

Greenberg and associates at CASEL (2003) noted that a frequent reason for poor school performance is the presence of negative behaviors and poor coping strategies among students, arguing that programs aimed at positive youth development would improve educational outcomes. This conceptualization of SEL as primarily the prevention of risk behaviors, while legitimate, has led some practitioners to emphasize youth development programs at the expense of goals drawn from SEL’s roots in educational psychology, such as self-efficacy. Still, scholars affiliated with CASEL have provided empirical evidence validating their approach. In a meta-analysis encompassing 213 studies that involved 270,034 students, Durlak, Weissberg, Dymnicki, Taylor, and Schellinger (2011) found that school-based SEL interventions were frequently effective at improving students’ SEL competencies, with a median effect size of Hedges’ $g = .3$. The authors found a similar effect on academic achievement as measured by tests ($g = .27$) and grades ($g = .33$) among schools implementing SEL interventions, relative to controls. Findings of a follow-up meta-analysis (Taylor et al., 2017) were consistent with the earlier results, with mean effect sizes for various outcomes reviewed ranging from $g = .13$ to $.33$.

Summary

Educational psychology is central to learning and instruction across subject areas. Students’ self-cognitions are an invisible but powerful force that educators are increasingly leveraging to help students succeed in school and other pursuits. Research has demonstrated that improving students’ self-efficacy, goal orientation, and other social-emotional competencies can have beneficial effects on important academic outcomes.

Social-Emotional Benefits of Musical Involvement

Teachers can improve students' SEL abilities through targeted programming, but it can also be infused into routine instruction (Zins, Bloodworth, et al., 2004). Fortunately, music classrooms are natural laboratories for social-emotional learning (Edgar, 2017; Jacobi, 2012). Teachers tend to have long-term relationships with students, music classrooms tend to be positive social environments, and music learning is a natural laboratory for educational psychology. Yet, with the exception of a few early studies (e.g., M. Greenberg, 1970; Michel, 1971; Michel & Farrell, 1973; Scott, 1992; and see Hanshumaker, 1980 for an early review), research into the social-emotional benefits of musical involvement was scarce until around the turn of the last century and remains understudied relative to cognitive and academic benefits (Hallam, 2010). There is now an emerging literature surrounding the intersections between musical involvement and educational psychology. Some studies have examined the social and psychological benefits of music participation, while others have treated these issues as predicting enrollment, success, and persistence in music programs.

Social and Emotional Outcomes Associated with Music

In *Champions of Change*, Burton, Horowitz, and Abeles (1999, 2000) noted that students engaged in various arts learning environments experienced improved self-concept, risk-taking, imagination, and sense of social support. In subsequent years, studies found music study to be associated with increased self-esteem (Costa-Giomi, 2004; Dagaz, 2012; Devroop, 2012), self-efficacy (Holochwest et al., 2016; Stevenson & Deasy, 2005), and academic self-concept (Degé et al., 2014; Degé & Schwarzer, 2018; Shin, 2011). Research has also established preliminary connections between music participation and a heightened sense of belongingness and social support (Adderley et al., 2003; Bartolome, 2013; Dagaz, 2012; Kokotsaki & Hallam, 2007;

Miksza, 2010) as well as school engagement and commitment (Dagaz, 2012; Miksza, 2010; Stevenson & Deasy, 2005). Analysis of a national dataset using propensity score matching also found high school music participation to be associated with a reduction in risk behaviors including substance abuse and suspension (Elpus, 2013a). In a (2006) survey conducted by Harris Interactive on behalf of MENC, more than half of school principals reported that they believe that music programs have benefits for school motivation, engagement, and graduation, and a subsequent survey revealed that parents and teachers believe that choral singing has social-emotional benefits as well (Chorus America, 2009).

There have also been a number of studies finding null effects. Higher music self-esteem attributable to music instruction did not translate into higher general self-esteem (Portowitz et al., 2009). School-based music interventions in Australia modestly improved participants' self-esteem, but the effect of the music interventions was no better than other programming (Rickard et al., 2012, 2013). Similarly, Shields' (2001) musical intervention raised participants' musical self-concept, but that this had limited effects on global self-concept. In a large-scale analysis of the Panel Study of Income Dynamics ($N = 18,000$) using propensity score weighting to control for selection bias, participation in extracurricular music lessons did not have a statistically significant effect on later outcomes (Foster & Jenkins, 2017). The authors did find tentative support for the notion that music lesson participation builds social-emotional abilities such as self-esteem, self-efficacy, and academic self-concept, but cautioned that further research was warranted.

Social and Emotional Factors Involved in Music Success and Persistence

Success in music programs is inextricably linked to self-efficacy (McCormick & McPherson, 2007). Practice quality, an important factor in the success of music students and

particularly instrumental music students, is also related to motivation, self-regulation, and personality (Miksza, 2009, 2011; Varela et al., 2016). Motivation, value for music, competence, and perceived competence are also related to persistence in school music (Hallam, 2006; McPherson & Hendricks, 2010; Schmidt, 2007; Sichivitsa, 2007), as “higher levels of motivation lead to music achievement via practice and perceived competence” (MacIntyre et al., 2012, p. 138). It may be that sustained musical engagement enhances these abilities. It may also be true that, similar to music’s ostensible academic and cognitive benefits, that these factors predict persistence in music rather than growing as a result of it. There is also a third possibility, that music students have certain pre-existing cognitive or attitudinal advantages *and* that music helps to further develop those capacities (Degé et al., 2014; Foster & Jenkins, 2017).

Summary

Research into music’s social-emotional benefits is hampered by many of the same issues as research into academic benefits. Still, it is less studied relative to academic and neurological benefits and remains a promising area for scholars to investigate.

Mechanisms by Which Extramusical Benefits Accrue

Transfer of Learning

Transfer is the *sine qua non* of schooling: “Subject disciplines are included within the curriculum in the belief that the knowledge they provide has a relevance to wider human competencies” (Brown, 2001, p. 86). Transfer is also at the heart of any theory of music’s extramusical benefits. However, transfer of learning to new situations is not as neat and tidy as it may first appear. At best, the study of transfer suffers from fuzzy definitions: lateral and vertical, near and far, specific and nonspecific (Royer et al., 2005). At worst, it happens far less often than educators would like to admit, with some authors (e.g., Detterman, 1993) making the case that

the very idea of transfer is overhyped and rarely occurs. Transfer may simply be the application of the same underlying cognitive processes in both situations, similar to the selection bias problem in the study of extramusical benefits.

The publication of Detterman's (1993) *Transfer on Trial* predictably led to rejoinders and ultimately a reexamination of what does and does not constitute transfer, and when and how it occurs. Bransford and Schwartz's (1999) seminal article articulated a distinction between "sequestered problem solving" (SPS), a narrow conceptualization of transfer, and a broader "preparation for future learning" (PFL). Classical research methods, particularly quantitative ones, may steer researchers away from studying PFL, as it is often temporally distal and difficult to measure. Indeed, the kind of transfer to adult life that supposedly underpins our educational system—whether from the study of music, mathematics, health, or civics—is nearly impossible to assess (Barnett & Ceci, 2002). It is possible that previous studies and even theories of transfer had conceived it too narrowly to observe its effects, definitionally excluding or statistically controlling for what could have been promising findings (Schwartz et al., 2005).

Beyond the distinctions between SPS and PFL, an additional dimension of transfer is how similar or different the novel situation is to the original content. A classic distinction, which is more a continuum than a dichotomy, is "near" transfer vs. "far" transfer. Barnett and Ceci (2002) criticized the "near vs. far" question as overly simplistic, noting that transferability involves factors beyond merely the knowledge domain, including temporal relationship, social context, and modality of response. Whatever the definition, the literature is in agreement that near transfer occurs more reliably. This notion challenges music educators' beliefs that their efforts miraculously result in extramusical benefits and complicates the efforts of researchers to isolate the effects of transferred skills or dispositions from other factors.

Transfer From Musical Involvement

The critical matter of the mechanism through which music or arts instruction improves students' academic or social-emotional abilities remains unanswered. The majority of research studies linking music with extramusical outcomes do not even attempt to provide an explanation for the connection (Catterall, 2002; Miksza, 2010; Winner et al., 2013). Nevertheless, an understanding of these specific connections is vital (Royer et al., 2005). There are as many ways of teaching music as there are music teachers. Teaching can be designed to maximize opportunities for transfer (Bruner, 1977; Engle et al., 2012; Wiggins & McTighe, 2005), but in order to do so, music educators need to understand which aspects of their practice are beneficial for students.

It may be that music lessons prompt positive interactions between teacher and child, and between parent and child as the pupil demonstrates what she or he has accomplished (Costa-Giomi, 2004). It may also be that music improves intelligence or specific cognitive functions (e.g., Rauscher & Hinton, 2011; Schellenberg, 2004, 2006, 2011). One of the most commonly expressed views by scholars is that the social aspects of music participation may promote positive social-emotional attributes such as motivation, engagement, or belongingness, which could transfer to other school pursuits (e.g., Catterall, 2009; Degé et al., 2014; Elpus, 2013a; Winner et al., 2013; Winner & Hetland, 2000). This explanation, while plausible, is not yet supported by extant research.

It may also be the case that artmaking engenders dispositions or “habits of mind” that benefit all learners but are disproportionately used or cultivated in artistic pursuits (e.g., Baum et al., 1997; Brown, 2001; Eisner, 2002; Gardner, 1990; Hetland et al., 2013; Scripp, 2002; Stevenson & Deasy, 2005). Two more related dimensions of music study that are frequently

absent from other school subjects are meaning-making and feeling. Noted music education philosopher Bennett Reimer wrote, “Music, I suggest, is perhaps the most effective mode for cultivating, extending, and refining the felt undergoings that are the basis for human consciousness and cognition” (2003, p. 80; see also Dewey, 1934), and indeed it is possible that the richness of personal and collective experience gained through individual or collaborative music-making is a transferable positive good in the lives of students. Because the arts are more multilayered and less concrete than, say, mathematics, it follows that transfer might also be less concretely observable in the ways that classical educational measurement prizes (Brown, 2001).

These benefits, whether or not they are measurable or visibly transferable to other measurable pursuits, are inherently worth pursuing. This is especially the case in the current milieu of hyperfocus on reductive, quantifiable student outcomes. As Alfie Kohn quipped, “We miss the forest while counting the trees” (2012, p. 36). Indeed, if arts education advocates focus only on measurable outcomes, the most important—and germane—benefits of school music participation may be overlooked entirely (Brown, 2001; Burton et al., 2000; Chapman, 2001; Eisner, 1998; Hetland et al., 2013; Kohn, 2015; Reimer, 1999).

Perhaps the most compelling research-supported theory to date is the peer group hypothesis (e.g., Catterall, 2009; Miksza, 2010). Music students have been repeatedly shown to be high academic achievers with college ambitions, cultural capital, strong social-emotional skills, and low incidence of at-risk behaviors. Music programs usually bring the same group of students together year after year, and often involve shared and salient activities such as performances, trips, and social events. It is plausible that the most direct way that music promotes higher achievement or social-emotional abilities is through sustained engagement with a positive peer group. On one hand, this can be seen as a research-backed correlate of music

participation. On the other, if the desirable characteristics of the music peer group are largely a reflection of racial and socioeconomic privilege, it may also be seen as a form of social exclusion and reproduction:

Prior cognitive outcomes and performance predict arts participation and subsequent outcomes. The strength of these associations is troubling. Arts participation may be yet another advantage enjoyed by [privileged] children. If arts participation produces better outcomes, it builds on the child's prior success. In this way, arts education may reinforce inequality generated by child- and family-level advantages. (Foster & Jenkins, 2017, p. 434)

In light of research showing that band and orchestra students tend to have both higher socioeconomic status and better school outcomes than choir students (Elpus, 2013b; Kinney, 2008, 2018), it is worth considering whether selection into vocal or instrumental music is more a function of students' personal preferences, schools' systematic bias against marginalized social groups (see also Bourdieu & Passeron, 1990; Roscigno & Ainsworth-Darnell, 1999), or an interaction of these forces.

Summary

This section of the review highlights the fact that research into extramusical benefits via transfer from musical engagement will continue to be unconvincing unless and until empirically supported theories about the nature of the transfer emerge. The development and validation of any such theories will almost certainly require collaboration with specialists in fields outside of music education.

Challenges in Researching Extramusical Benefits

Methodological Limitations in Extant Literature

An inescapable conclusion noted earlier in this review and supported by authors of several reviews and meta-analyses (Elpus, 2011; Foster & Jenkins, 2017; See & Kokotsaki, 2016; Winner et al., 2013; Winner & Cooper, 2000; Winner & Hetland, 2000) is that a great deal of the research into the extramusical benefits of music instruction suffers from a lack of methodological rigor. Studies tend to either have small sample sizes, susceptibility to uncontrolled confounding, or both. See and Kokotsaki (2016, p. 251) lamented, “Despite the huge body of work on arts education, little can be concluded about the beneficial effects of arts on children’s learning and wider outcomes. This is largely due to weaknesses in the research, lack of replication and inconsistent findings across studies.” These limitations are understandable given the research resources afforded to the typical music researcher. A few obstacles are enumerated next. Others are articulated in Chapters 3 and 6.

Challenges With Measurement

Quantitative research is predicated upon accurate measurement of both the musical treatment and the outcome to which transfer supposedly occurs. The majority of the studies in the literature rely on available data as the dependent variable, rather than seeking out theoretically implied outcomes or potential transfer mediators. This is particularly true with psychological constructs such as self-efficacy and goal orientation, which require specific and sustained efforts to capture. Qualitative studies (e.g., Adderley et al., 2003; Burton et al., 1999, 2000; Dagaz, 2012; Pitts, 2007; Stevenson & Deasy, 2005) have revealed not only intriguing directions for future study of music transfer but also the limitations of quantitative methods in studying it as it is currently understood, particularly with administrative data.

There is also the issue of treatment fidelity. Certainly, no one believes that music instruction is implemented uniformly across states or even within schools. Students' experiences in music classrooms are highly variable. Differences in music teachers' expertise, commitment, philosophy, curriculum, and instructional approaches are difficult to operationalize, let alone measure on a large scale, complicating efforts to ascertain an "average" treatment effect. Factors including instructional time and family and school resources also have an effect on students' musical experiences. For example, there was a significant amount of heterogeneity in teachers' implementation of even a single SEL intervention within music programs in the Philadelphia schools (Holochwost et al., 2016). Existing state and national datasets do not include information about these factors (Foster & Jenkins, 2017), and funding to conduct comprehensive research incorporating them is scarce.

While medical studies routinely track participants' compliance with, for example, routinely taking the medication under study, there is typically no information in music studies about students' "treatment compliance" in music. Students who appear comparable in a dataset may have vastly different—and unobserved—amounts of social engagement, practice time, musician identity, parental experience and encouragement, commitment, motivation, private instruction, extracurricular musical involvement, and myriad other constructs that are likely to both predict "treatment compliance" and affect whatever the transfer mechanism is hypothesized to be. These influences are consequential, with the ability to affect a study's conclusions. For example, in Costa-Giomi's (1999) study, attendance and practice time explained greater than 20% of the variance in students' resulting cognitive ability. When researchers speculate about transfer mechanisms, they assume that the mechanism works similarly for all students, which cannot be assumed to be the case in practice (Gardner, 1990).

Other Sources of Unobserved Confounding

In addition to complexities resulting from variation between treatments and treatment compliance, there are other sources of unobserved heterogeneity affecting studies of extramusical benefits that are frequently unaddressed in the literature. Two of these are psychological constructs: general intelligence or IQ, and personality. In one study linking music and academic achievement, IQ moderated all other predictors to nonsignificance (Babo, 2004). In another, music lessons and IQ were each positively associated with a measure of emotional comprehension, but music was nonsignificant when both music participation and IQ were included as predictors (Schellenberg & Mankarious, 2012). Personality characteristics, frequently measured by the Five-Factor Model popularly known as the “Big Five” dimensions of extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience (Digman, 1990; Goldberg, 1993; McCrae & John, 1992), have also been shown to be related to music persistence: “Openness-to-experience predicted duration of music training with demographics and cognitive ability held constant, but cognitive ability did not predict music training when demographic and personality variables were held constant” (Corrigall et al., 2013, p. 8; see also Catterall, 2009; Corrigall & Schellenberg, 2015). Information about constructs from cognitive or personality psychology is almost never collected by school districts or for general-purpose datasets such as those produced by NCES. IQ and the Five-Factor Model are not without their critics (e.g., Block, 1995), and they do not always result in expected gains in achievement (e.g., Blakey et al., 2020), but any study not accounting for students’ core cognitive traits is suffering to a certain extent from omitted variable bias.

In addition, unobserved family characteristics are likely playing a role in the observed association between music and other outcomes (Catterall, 2009). As elaborated earlier, parents

are an important (but difficult to quantify) influence on their students' success or failure with instrumental music. However, parental involvement specifically with music may not be the only force at work:

Additionally, the unobserved characteristics of parents could confound the observed relationship between a given investment and children's outcomes. For example, some parents may be "high-investing" regardless of how they perceive their child's ability. In this case, the relationship between exposure to investments, like music, and children's academic success would be confounded by the unobserved characteristics of their parents... These children would perform well at many tasks because their parents created a strong and perhaps diverse child investment portfolio. This makes it quite challenging to identify the unique causal effect of any one investment (i.e., arts education). (Foster & Jenkins, 2017, p. 411)

Summary

It is noteworthy that, twenty years after the REAP curators' admonition to "mute those claims," there remains no published research describing firm causal relationships between school music participation and academic achievement or SEL, and no momentum toward establishing any. There are legitimate reasons why the gold-standard randomized experiment has rarely been used, and in Costa-Giomi's study was used for extracurricular piano lessons rather than school ensembles. Experimental manipulation of students' elective choices is infeasible to achieve in practice; few students or parents would consent to a researcher prescribing their courses of study.

Even so, several techniques for causal inference in nonexperimental studies have not been used as often as they could have been. This may be because music researchers are unaware of them; Catterall himself wrote that "There are no ready ways to measure [unobservable family

effects] or to separate selection effects from the effects of actual participation” (Foster & Jenkins, 2017, p. 411). Perhaps this belies an unfamiliarity with propensity scores and similar techniques that would benefit this line of scholarship if they were to become more commonly used. Music education scholars may also be unable to secure the kinds of research support available to researchers in other disciplines. Still, as this area of research continues to interest the field, Winner and Hetland issued a “call for more rigorously designed, theory-driven, quasi-experimental research with better-conceived comparison groups so that we may draw firmer conclusions about how and when the arts transfer to other subjects” (2000, p. 7). This study attempts to meet the challenge of designing a comparable comparison group.

Quasi-experimental designs such as interrupted time series and regression discontinuity (Shadish et al., 2002), and the model-based instrumental variables approach (Murnane & Willett, 2011; Wooldridge, 2002), while rigorous, seem less feasible in most practical applications in this research context. It is therefore unsurprising that fixed-effects models have been among the profession’s first forays into causal methods. They are straightforward to estimate, and in most cases can be used with available data—an important concern to music researchers who, as this review demonstrates, are overwhelmingly using general-purpose datasets that were neither designed to answer music-specific research questions nor be well-suited for causal inference. The absence (excepting Elpus’ 2013a and Foster and Jenkins’ 2017 studies) of propensity score models from the literature, however, is perplexing, as they are also well suited to general-purpose datasets but offer much greater modeling flexibility. The major advantage of fixed-effects models over propensity score models is that while propensity score models can yield estimates unbiased by selection conditional on the observable variables in the dataset, estimates from fixed-effects models are unbiased by both observable and unobservable factors

(Wooldridge, 2002). Still, propensity scores enable a wider range of analyses and are estimable with a greater variety of data structures. Propensity score analysis and causal inference are detailed in the next chapter.

Chapter 3: Review of Propensity Score Literature

Overview

When determining the effect of an intervention or treatment on educational outcomes, isolating the effects of the treatment from potential confounders is critical. With sufficient sample sizes, randomization generates comparable groups that differ only to the extent that they received the treatment. Observational studies are undertaken when the researcher cannot randomize or otherwise assign participants to the various treatment states. Observational studies require additional design-based or model-based strategies to isolate the effects of the intervention. Sometimes, as in the case of “natural experiments” or other serendipitous events that approximate randomization, complicated statistical models are unnecessary as ostensibly comparable control and treatment groups⁴ arise on their own. However, in many situations involving retrospective analysis, it is necessary to use statistical techniques to create otherwise comparable groups for comparison. Propensity scores, the focus of this review, are one such technique.

⁴ While methodological traditions have various conventions for the naming of groups of participants in a study, I will use “treatment group” to refer to the population that received any intervention of interest and “control group” or “comparison group” to refer to the population(s) that either received no intervention or an alternate intervention.

Causal Inference

Foundations

The very notion of what constitutes a causal relationship is a matter of debate (see, for example, Cook & Steiner, 2010; Holland, 1986; Pearl, 2000; Rubin, 1978, 1986; Shadish, 2010; Shadish et al., 2002). Causal inference in statistics is built around the examination of the ostensible cause-and-effect relationship between known phenomena A and B , rather than an open-ended search for what the causes of B might be (Morgan & Winship, 2007). Correlation or covariance can establish an associational relationship (i.e., a joint probability) between A and B , but not necessarily a causal one. A could cause B , B could cause A , or a third variable could be affecting both A and B . Judea Pearl (2000) distinguished the causal relationship from a probabilistic relationship in terms of understanding a stable causal mechanism, stating that causal networks are built

on the assumption that each parent-child relationship in the network represents a stable and autonomous physical mechanism—in other words, that it is conceivable to change one such relationship without changing the others. . . A joint distribution tells us how probable events are and how probabilities would change with subsequent observations, but a causal model also tells us how these probabilities would change as a result of external interventions. (Pearl, 2000, p. 22)

The idea that understanding and isolating the mechanism through which A ostensibly causes B is widely regarded as necessary to support an inference of a causal effect. Furthermore, Sociologist Paul Lazarsfeld (1959) identified three criteria for a relationship to be causal⁵. First,

⁵ This has also been attributed to the 18th century philosopher John Stuart Mill (Shadish et al., 2002).

the cause must precede the effect (temporal ordering). Second, the two variables must be empirically related (covariance). Third, the observed empirical relationship cannot be explainable by any third variable. These three criteria were later adopted as part of the definition of internal validity by Cook and Campbell (1979) and Shadish, Cook and Campbell (2002) in their influential books on causality in the social sciences, conceptualizing threats to internal validity as factors other than the intervention that may have affected the outcomes. Pearl (1995, 2000) formalized the identification and isolation of causal mechanisms with the use of directed acyclic graphs and the Structural Causal Model. Taken together, these principles give theoretical justification to the axiomatic truth that correlation does not imply causation.

The Counterfactual Tradition

An approach to estimating causality that has been widely adopted in the social sciences is Rubin’s Causal Model (RCM), also known as the “potential outcomes” or counterfactual framework (Holland, 1986; Rubin, 1974, 1978). Rubin observed that a given individual never exposes both the control (Y^0) and treatment (Y^1) states simultaneously. Therefore, each individual i has one observed outcome, either Y_i^1 if the observed outcome results from the treatment condition, or Y_i^0 if the individual experienced the control condition⁶. Each individual also has one unobserved potential outcome: the counterfactual, or what *would* have happened had the individual been assigned to the other condition. According to Rubin, “The causal effect

⁶ There are many notational conventions for potential outcomes. For consistency, throughout this manuscript I will use Morgan and Winship’s (2007) notation of Y_i^1 representing the value of potential outcome Y for individual i in the treatment group and Y_i^0 representing the potential value of Y for an individual in the control group. I will use Z as a binary treatment indicator where $Z = 0$ is the control condition and $Z = 1$ is the treatment condition.

of one treatment relative to another for a particular experimental unit is the difference between the result if the unit had been exposed to the first treatment and the result if, instead, the unit had been exposed to the second treatment” (1978, pp. 34–35). Mathematically, the estimated individual-level causal effect δ_i is the “what-if difference” (Morgan & Winship, 2007): the difference between the observed outcome and the counterfactual (Equation 3.1).

$$\delta_i = Y_i^1 - Y_i^0 \quad (3.1)$$

Since one outcome is always the hypothetical counterfactual, it is impossible to observe both Y_i^0 and Y_i^1 simultaneously for any one individual, a challenge known as the “fundamental problem of causal inference” (Holland, 1986, p. 947). Therefore, causal effects using the potential outcomes framework must be estimated in the aggregate. The average causal effect is $E(\delta)$, the expected difference between an individual’s potential outcomes if she or he had or had not been assigned to the treatment.

Selection Bias and Ignorability

In what Rosenbaum (2010) labeled the “naïve model,” $E(\delta)$ is assumed to be an unbiased estimate of the average causal effect of the treatment. This is a “naïve” assumption since it only holds if Lazarsfeld’s third criterion of no spurious association is met. However, in nearly all observational studies, the naïve estimate of $E(\delta)$ suffers from selection bias: systematic differences between control and treatment participants that are related to the outcome Y (Hernán et al., 2004; B. Schneider et al., 2007). In order to rule out a spurious association between treatment assignment and the Y , treatment assignment must be statistically independent of the potential outcomes, as in Equation 3.2:

$$(Y^0, Y^1) \perp\!\!\!\perp Z \quad (3.2)$$

While Equation 3.2 implies that the forces resulting in assignment to either the control or treatment condition are uncorrelated with Y , it does not imply that the treatment is independent of Y —if this were the case, the analysis would be unnecessary. In nonrandomized studies, it almost always fallacious to assume that (3.2) is true. Thus, the standard of ignorability of treatment assignment often applied to nonexperimental studies is ignorability given the observed covariate vector $\mathbf{x}$ (Equation 3.3), also called “selection on observables”:

$$(Y^0, Y^1) \perp\!\!\!\perp Z \mid \mathbf{x} \quad (3.3)$$

Rosenbaum and Rubin used (3.3) as a part of their definition of strong ignorability. Strong ignorability can be assumed when (1) treatment assignment is independent of the potential outcomes given the observed covariates, and (2) all cases have a positive (i.e., nonzero) probability of receiving the treatment (Rosenbaum & Rubin, 1983b). Strong ignorability is also known as unconfoundedness, exogeneity, and conditional independence in the vernacular of various research traditions (Guo & Fraser, 2015).

Theory of the assignment mechanism is central to the implications of assuming strong ignorability. If the analyst assumes that all variables related to the selection mechanism are observed and included in $\mathbf{x}$, δ_i can fairly be said to have isolated and captured the treatment effect. However, this is quite an assumption. The analyst assuming strong ignorability in a causal analysis is in effect claiming that treatment assignment is also statistically independent of vector $\mathbf{u}$ of any possible unobserved variables u , as in Equation 3.4 (Robins et al., 2000; Rosenbaum & Rubin, 1983a):

$$(Y^0, Y^1) \perp\!\!\!\perp Z \mid \mathbf{x}, \mathbf{u} \quad (3.4)$$

No matter how sophisticated the model or comprehensive the dataset may be, it is impossible to know or prove that a confounding influence beyond the observed covariates is biasing the results

(Shadish & Steiner, 2010; Steiner et al., 2010). Steiner and Cook (2013) differentiated between overt bias due to observables and hidden bias due to unobservables. The assertion of no hidden bias (i.e., that x fully explains treatment assignment) is strengthened when the selection mechanism is well understood by the researcher:

This assumption may approximately hold if an extensive set of measures known by subject-matter experts to be predictive of the treatment has been collected at the pretest.

In other applications, this assumption will be highly questionable, for example, if only a few demographic variables are present. (Schafer & Kang, 2008, p. 280; see also Morgan & Winship, 2007; Robins, 1999; Shadish et al., 2002; Steiner & Cook, 2013)

As Rosenbaum (2010) elegantly summarized, people who *look* comparable may not *be* comparable.

Experimental and Observational Studies

An experimental study is one in which the researcher actively assigns participants to the various treatment states in advance, usually randomly (Shadish et al., 2002). In contrast, in an observational study, treatment selection or assignment is not under experimental control (Rosenbaum, 2010). An experimental study in which the researcher has taken steps to rule out sources of confounding (e.g., ensuring that patients are taking their medications as prescribed and not taking any other medications) is known as a randomized controlled trial, or RCT (S. G. West et al., 2008). Randomized experiments are widely considered to be the gold standard for establishing causality. Rubin frequently used the mantra, “No causation without manipulation” (e.g., 1986, p. 962). Nonetheless, whether a study supports the inference of causation is not a matter of black and white, but rather degrees of theoretical support, design rigor, and plausibility of assumptions. As a practical matter, RCTs are rare in education research because “parents

choose the schools their children attend (either directly or in decisions of where to live), teachers decide where they want to teach, and students often make choices about courses to take and colleges to which to apply” (S. E. Graham & Kurlaender, 2011, p. 342). Furthermore, school officials, under pressure to demonstrate student achievement, are often reluctant to give up instructional time to a researcher in order to participate in RCTs (Cook, 2002).

Randomized experiments are also more complicated in education than fields like medicine due to limited “exchangeability” (Greenland & Robins, 2009) of participants. A child taking diabetes medicine in rural Oregon is independent from and exchangeable with a same-age peer in a classroom down the hall and another from Chicago, since the internal biological processes at work are assumed to be unaffected by doctor or regional effects. Conversely, in education, different states, districts, teachers, and peers can all be reasonably assumed to have affected an outcome in some way. In order to maximize external validity in education research, studies are necessarily larger in scope than is typically the norm in medicine or other fields like psychology. Finally, funding for RCTs in education, particularly large-scale, nationally representative RCTs, is highly competitive and unavailable to many researchers.

Whether a study is experimental or observational has critical implications for its ability to support causal inference with strong internal validity. Consider as an example an investigation of whether a proposed intervention, the Coercing And Not Demotivating You (CANDY) program of giving students candy each time they finish a book, improves reading scores relative to an existing program, Kids Are Learning Effectively (KALE). In a prospective experimental design, a researcher could randomly assign students to either KALE or CANDY. With a sufficient sample size, this would yield groups of students who did and not receive the treatment that have covariate balance; i.e., the distribution of $\mathbf{x}$ will be equal between the groups. Crucially, the

groups are also assumed to be balanced on all unobservables $\mathbf{u}$. Randomization also results in each participant having a known probability of assignment to the treatment. For example, if treatment assignment is done by the flip of a coin, all subjects have a probability of $\pi = .5$ that they would be assigned to the CANDY program. This probability of assignment to the treatment, π , is each participant's propensity score. This process of randomization "breaks the link" between treatment assignment and the outcome (Ho et al., 2007, p. 207). The potential outcomes are independent of treatment assignment, as in Equation 3.4, and $E(\delta)$ is assumed to be free from selection bias.

In observational studies, where randomization is impossible, there is nearly always some lingering doubt about the study's internal validity. Confounding is a pervasive concern, requiring more work on the part of the researcher to plausibly rule it out if a causal claim is to be made. Continuing the example, if children had instead been allowed to volunteer for the CANDY program, and a retrospective observational study were conducted, each student's probability of selection into the treatment would be unknown. Furthermore, covariate balance is unlikely, and a naïve estimate of $E(\delta)$ would almost certainly be confounded by selection bias. However, observational studies incorporating methods for convincing internal validity can also serve as a foundation for causal inference: "Subject to certain plausible assumptions, genuine causal inferences can in many cases be distinguished from spurious covariations" (Pearl, 2000, p. 43). Empirical methods such as instrumental variables (Murnane & Willett, 2011) and graph theory (Pearl, 1995, 2000) leverage path analysis to isolate the causal mechanism. Fixed-effects regression models (Allison, 2009; Wooldridge, 2002) attempt to render selection independent of outcomes by using each person as her or his own control. The interrupted time series design can also support causal inferences with longitudinal data (Shadish et al., 2002).

There are also design-based methods that resemble randomized experiments. At times, which individuals qualify for treatments and which do not is haphazard enough to be almost (or “arguably”) random, giving rise to the terms “quasi-experiment” and “natural experiment” (Murnane & Willett, 2011)⁷. A natural experiment is a situation in which experimental conditions arise serendipitously; for example, the effects of a magnet school can be evaluated under conditions that closely approximate an experiment if students are selected for the school by a lottery. A quasi-experiment is a study that resembles an experiment except for the researcher’s ability to assign participants to the treatment conditions. One example of a quasi-experimental design is the regression discontinuity (Shadish et al., 2002). In this design, there is a program, such as school-based reading support, that is offered to students at a certain empirical threshold, such as a certain score on a diagnostic exam. Differences between scores within a narrow range around the cutoff can be attributed to random measurement error. Students with scores just below the threshold are therefore arguably comparable to those with scores just above, and this comparability enables a researcher to provide a quasi-experimental accounting of the program’s efficacy.

In all of these cases, treatment selection is plausibly independent of both $\mathbf{x}$ and $\mathbf{u}$. Still, even though they are innovative and can contribute knowledge that would be difficult to obtain in an experimental setting, they inherently fall somewhat short of the gold standard of experimental control and manipulation. Rosenbaum, an innovator and champion of propensity

⁷ Shadish, Cook, and Campbell (2002) and Cook (2002) enumerated rigorous and careful principles for what constitutes a “quasi-experiment” and regretfully observed that some researchers blithely apply the label to nearly any research that is not a randomized controlled trial, but that they hope will be taken seriously.

score methods for the past 35 years, acknowledged that “many if not most ‘natural experiments’ fail to produce this ideal [comparison]; moreover, even if one ‘natural experiment’ once succeeded, there would be no way of knowing for sure that it had succeeded” (2010, p. 69). Pearl took less of a dogmatic stance:

How safe are our predictions when we recognize three-dimensional objects from their two-dimensional shadows, or from the two-dimensional pictures that objects reflect on our retinas? The answer is: Not absolutely safe, but good enough to tell a tree from a house and good enough to make useful inferences without having to touch every physical object that we see. Returning to causal inference, our question then amounts to assessing whether there are enough discriminating clues...to allow us to make reliable discriminations between cause and effect. (Pearl, 2000, p. 60)

There is a validity trade-off between experimental and observational studies. While experiments have stronger confounder control, their external validity is often limited. It is difficult if not practically impossible to conduct a randomized educational study under generalizable (i.e., “real-world”) conditions. Conversely, the limited internal validity of observational studies is balanced by the presence of all the world’s unobservables in the results, which leads to stronger external validity. The last word here belongs to Rosenbaum and Rubin: “Despite their limitations, observational studies will continue to be a valuable source of information, and therefore it is prudent to develop appropriate methods of analysis for them” (1983a, p. 212).

Propensity Score Foundations

The Theory of Matching

One additional strategy for separating x from the treatment is matching. While related to both the model-based and design-based approaches, matching does not fall neatly into either category. The premise of matching is to create pairs of participants that are as comparable as possible except for treatment selection. (A well-known use of matching is the attempt to separate “nature from nurture” in twin studies.) An analyst hoping to isolate the effect of the CANDY program might create matched pairs of students, with one student in each pair having received the intervention and one functioning as a control in the KALE program. Matching in observational studies uses observed variables. For example, the researcher might create pairs of students, one CANDY and one KALE, who have the same sex and prior reading scores. Assuming that the students in each pair are otherwise comparable, the difference in their outcomes would theoretically represent only the effect of the intervention (Heckman, Ichimura, & Todd, 1998).

In practice, matching is only possible with a limited number of variables. In the CANDY study, assuming a moderate sample size, there may be enough participants to match on sex and prior reading achievement. The assumption of strong ignorability would be strengthened if additional variables (e.g., race, parents’ education, number of close friends, or days absent) were included in the matching, as each represents a potential confounder. Unfortunately, as the number of covariates grows to a larger number n , finding matches will become increasingly impossible as there will be 2^n permutations of the covariates. Attempting to match on more variables leads to more cases that are unmatched and therefore discarded, biasing the results.

Paradoxically, matching on fewer variables may lead to greater overall covariate imbalance, which also biases the results. This issue is known as the problem of dimensionality.

Defining the Propensity Score

In their seminal paper describing propensity score analysis, Rosenbaum and Rubin (1983b) explained that the need for propensity scores came from the problem of high-dimensional matching in observational studies. Rather than achieving covariate balance by attempting to match cases on many variables, the propensity score method involves a single balancing score: the estimated propensity score⁸, notated $\hat{e}(\mathbf{x})$, or each case's conditional probability of receiving the treatment given all relevant covariates $\mathbf{x}$. The propensity score solves the problem of dimensionality by reducing the (often large) vector $\mathbf{x}$ into a single scalar such that “a pair of treated and control participants sharing a similar propensity score are essentially viewed as comparable, even though they may differ on values of specific covariates” (Guo & Fraser, 2015, p. 130). Rubin noted that the propensity score “has the rather remarkable property of adjusting for all of the covariates that went into its estimation, no matter how many there are” (1997, p. 759). The use of propensity scores in published research has increased exponentially since around the year 2000 (Thoemmes & Kim, 2011).

Rubin further wrote that “The propensity score is the observational study analogue of complete randomization in a randomized experiment...[serving] to eliminate systematic biases in

⁸ The term “propensity score” technically refers to a participant's probability of receiving the treatment, regardless of the type of study. In experimental studies, the propensity score is known and is often (but not always) .5 as in a typical two-group randomized design. However, “propensity score” is most often used in the context of observational studies, in which case it is shorthand for “estimated propensity score” as the true propensity score is not known.

treatment-control comparisons” (2008, p. 820). Rather than assuming alleviating selection bias by conditioning on the full covariate vector $\mathbf{x}$, propensity score analysis is based on the potential outcomes being conditionally independent of treatment assignment given the estimated propensity score $\hat{e}(\mathbf{x})$:

$$(Y^0, Y^1) \perp\!\!\!\perp Z \mid \hat{e}(\mathbf{x}) \quad (3.5)$$

Empirical study has shown that matching on the propensity score leads to less bias than matching on sets of observed variables (Heckman, Ichimura, & Todd, 1998). Propensity scores are most commonly estimated using logit models, with variables believed to be related to selection serving as predictors. Cases with equal propensity scores are presumed to be equally likely to have received the treatment, and the control and treatment groups are expected to have covariate balance in the aggregate.

As noted above, randomized experiments are regarded as the gold standard for causal inference because they asymptotically achieve balance on all observed and unobserved covariates. However, propensity score methods only incorporate measured variables into the analysis. As Rosenbaum noted, “Randomization provides a basis for believing that an unobserved covariate u is balanced, but matching on propensity scores provides no basis for believing that” (2010, p. 73). Thus, propensity scores only attenuate selection bias to the extent that the assumption of strong ignorability holds, which is predicated upon the unverifiable belief that the covariates accurately modeled treatment selection. Furthermore, propensity score methods do not explicitly ensure that the covariate distributions are equivalent, only the means, bringing the question of whether the groups are truly comparable in an empirical sense (Heckman, Ichimura, Smith, et al., 1998). In sum, “Propensity score methods certainly do not

represent a magic solution to the problem of selection bias” (S. E. Graham & Kurlaender, 2011, p. 351).

The Stable Unit-Treatment Value Assumption

Another assumption underpinning propensity score analysis and other causal frameworks based upon potential outcomes is the Stable Unit Treatment Value Assumption, or SUTVA (e.g., Rubin, 1986). SUTVA “is a basic assumption of causal effect stability that requires that the potential outcomes of individuals be unaffected by potential changes in the treatment exposures of other individuals” (Morgan & Winship, 2007, p. 37). Put another way, δ_i is assumed to be stable and independent of the assignment mechanism or the treatments any other individuals receive. SUTVA is also violated when unrepresented versions of the treatment exist and the assumed-constant δ_i depends on which version of the treatment each participant receives (Rubin, 1986). SUTVA is another strong assumption that may not hold in many applied research contexts. In the CANDY program, for example, it may not be reasonable to assume that the program’s putative benefits would remain constant if the program size ballooned, or if the ways that different teachers administered the program were correlated with the outcome. Rather than serving to undermine most studies’ conclusions, “SUTVA reveals the limitations of observational data and the perils of immodest causal modeling rather than the limitations of the counterfactual model itself” (Morgan & Winship, 2007, p. 38).

Propensity Score Estimation

Covariate Selection

The degree to which estimated propensity scores plausibly satisfy the independence assumption in Equation 3.5 is dependent on the covariates in $\mathbf{x}$ that predict them. A rich set of variables, containing multiple measures from as broad a range of constructs as possible, is

necessary to establish strong ignorability when the precise selection mechanism is not known—and it typically is not (Steiner et al., 2010; Stuart, 2010). Parsimony is not needed in the propensity score equation. Methodologists (Rubin & Thomas, 1996; Shadish & Steiner, 2010; Steiner et al., 2010) are unanimous on the point that even seemingly redundant variables should be included as predictors, particularly variables that covary with the outcome (Brookhart et al., 2006; Kelcey, 2011). The exception to this rule is any covariates that perfectly predict treatment assignment (Stuart, 2010). Unreliability in the covariates can reduce propensity scores' ability to reduce bias (Steiner et al., 2011), but the effects of unreliability diminish as more predictors are used (Cook & Steiner, 2010). An additional issue is that when using secondary data, or data collected for a purpose other than the current study, the risk of inadequate prediction of assignment is acute:

The uncertainty about the ignorability of the assignment mechanism is even larger for observational studies that rely on archival data or other forms of secondary data. Such data are typically collected independently of the researchers' specific causal questions and without thought about the selection processes corresponding with the treatment at this time and in this setting. Although measures of demographics, prior academic achievement, and personality are often available from national or local archives...they will sometimes, perhaps often, fail to remove all selection bias.” (Steiner et al., 2010, p. 262)

Rosenbaum (1984) elaborated the bias resulting from including variables that many have been affected by the treatment in the propensity score equation. Covariates measured after the treatment began are not necessarily unusable, however. The variable could be unchanging, or changes could be unrelated to the treatment. Further, even if the variable might have been

affected by treatment, it may be reasonable to believe that the effect of treatment on the variable is small in comparison with its effect on the outcome. Still, analysts must obey “the fundamental rule for avoiding selection bias—not selecting on the dependent variable” (Ho et al., 2007, p. 212). Pretest measures of the dependent variable, however, are effective at reducing bias since in educational research they are often correlated with both the outcome and the selection process (Bifulco, 2012; Cook & Steiner, 2010; Hallberg et al., 2018; Wong et al., 2017). Covariate selection is more important for reducing bias than choices about analytic approaches (Cook et al., 2009).

Estimation Using Logit Models

Many methods exist for finding the ideal propensity score equation; “there is no definitive procedure” (Guo & Fraser, 2015, p. 141). Nevertheless, content analyses have shown that the vast majority of studies using propensity scores estimate them through a logistic regression model (e.g., Austin, 2009; King & Nielsen, 2018). Logistic regression models are generalized linear models for binary outcomes (Long, 1997), such as treatment condition vs. control condition. In this application, the exponentiated logit represents the propensity score, $\hat{e}(\mathbf{x})$ or $\hat{e}(\mathbf{x}_i)$: the conditional probability of having received the treatment for individual i with a given distribution of the observed covariates $\mathbf{x}_i$, following Equation 3.6.

$$\hat{e}(\mathbf{x}_i) = P(Z_i = 1 | X_i) = \frac{e^{\mathbf{x}_i\beta}}{1 + e^{\mathbf{x}_i\beta}} = \frac{1}{1 + e^{-\mathbf{x}_i\beta}} \quad (3.6)$$

The resulting estimated probabilities of selection into the treatment condition that are logically bounded by $0 < \hat{e}(\mathbf{x}) < 1$.

Estimation of the propensity score using logit models is subject to all of the assumptions and limitations typically associated with regression. Procedures such as model χ^2 and Hosmer-

Lemeshow testing, and checking for multicollinearity and outliers, are necessary but insufficient steps to establish that the model is performing well, as they cannot determine the extent to which the estimated propensity score reflects the true propensity score (Guo & Fraser, 2015). Perhaps the most pressing concern is that bias results from misspecification of the prediction equation (Heckman, Ichimura, Smith, et al., 1998; Heckman, Ichimura, & Todd, 1998; Lenis et al., 2018). Particularly with large numbers of predictors, equations including transformations, polynomials, and interactions can be difficult to specify properly but are necessary in order to achieve trustworthy results. Similar to the issue of covariate selection, since the true propensity score (i.e., the probability of receiving the treatment) is not known in practice, there is no test to determine how well the model calculated each participant's estimated propensity score.

Estimation Using Nonparametric Methods

In addition to the standard logit model, nonparametric methods for propensity score estimation include kernel matching (Heckman, Ichimura, & Todd, 1998), genetic matching (Diamond & Sekhon, 2013), near-far matching (Rigdon et al., 2018), and classification trees and their derivatives (Strobl et al., 2009). Nonparametric models require fewer assumptions, avoiding some of the ways that propensity score equations can be misspecified (Ho et al., 2007). As Guo and Frasier summarized, "If specifying functional forms can be avoided, then the search for a best model involves fewer subjective decisions and, therefore, may lead to a more accurate prediction of treatment probability" (2015, p. 144). However, "familiarity may predispose investigators to using logistic regression even when better alternatives may be available" (Westreich et al., 2010, p. 827).

CART and Ensemble Methods

A growing area of interest in propensity score estimation is CART, or classification and regression trees, an algorithmic approach associated with machine learning (Berk, 2006). CART is based on recursive partitioning, which is an algorithm-driven approach based on repeatedly classifying cases according to various combinations of the covariates in order to find the best possible prediction method. “Classification trees” canonically refer to partitioning by values of categorical variables and “regression trees” usually refer to continuous variables, although “classification” can also be used to refer to both. While classification trees do not provide fit statistics, coefficients, or prediction, they offer computational advantages over parametric regression models (Berk, 2006). On their own, classification trees do not always perform well as they can be prone to technical issues including overfitting. This led to the development of “ensemble methods” such as bootstrapped aggregation (“bagging”) and random forests, which is another resampling approach to classification trees (Strobl et al., 2009; see Loh, 2011 for a review of common algorithms). As Berk elaborated:

The term *ensemble methods* is commonly reserved for bundled fits produced by a stochastic algorithm, the output of which is some combination of a large number of passes through the data. Such methods are loosely related to iterative procedures, on one hand, and to bootstrap procedures, on the other. An example is the average of a large number of kernel smoothes of a given variable, each based on a bootstrap sample from the same data set. The idea is that a “weak” procedure can be strengthened if given an opportunity to operate “by committee.” (2006, p. 264)

Ensemble methods enable variables to perform differently in different settings—either with respect to variable selection or functional form (Berk et al., 2005). Therefore, ensemble

methods can perform better than logistic regression for estimating propensity scores (Berk et al., 2005), especially with large numbers of predictors (Strobl et al., 2009). Overfitting (i.e., the algorithm becoming overly reliant on certain predictors that happen to come up first) is an issue with CART, but less so with ensemble methods; due to the aggregation of results from many individual trees, local maxima are unlikely to persistently affect the final outcome (Strobl et al., 2009). While still not providing model coefficients or fit statistics, ensemble methods do identify which variables are most predictive (Strobl et al., 2009), potentially leading to future understanding of how the observed variables affected the selection process.

Generalized Boosted Modeling

Another adaptive nonparametric approach to propensity score estimation is generalized boosted modeling (GBM), also known as generalized boosted regression. Similar to CART, GBM does not require advance specification of a functional form, allowing for flexible relationships between the predictors and outcome (McCaffrey et al., 2004; Westreich et al., 2010). In addition, boosting can more easily accommodate mixed data types (continuous, categorical, count, or semicontinuous) that can be cumbersome to specify in a logit model (Burgette et al., 2015). The flexible nonparametric approach is most helpful nonlinear and especially nonmonotone relationships, but since there is no way to know the relationships in advance, GBM is recommended in all applications.

The GBM algorithm can best be understood as a variation on iteratively aggregating classification and regression trees in order to understand each variable's relationship to treatment status and estimate the propensity score (McCaffrey et al., 2004). Shadish and Steiner endorsed ensemble methods and GBM, noting that "Given a highly nonlinear selection model, reason exists to think these methods do indeed provide results that more accurately predict treatment

condition than binomial regression models” (2010, p. 21). While estimation methods based on classification trees are an active area of research, and their properties are not fully known, they hold promise for researchers using propensity scores (Luellen et al., 2005; Westreich et al., 2010) despite their rarity in extant scholarship (e.g., Thoemmes & Kim, 2011).

Covariate Balance

The primary goal of using propensity scores is to create groups of individuals that did and did not receive the treatment who are balanced on key covariates. Conditional independence given the propensity score (Equation 3.5) is predicated on covariate balance, which is checked after the propensity scores are estimated. Rubin (2008) defined covariate balance as the absence of systematic differences in the covariate distributions between the control and treatment groups. In experiments, randomization asymptotically achieves covariate balance; any differences are definitionally due to chance. Covariate balance with propensity scores is often better than covariate balance with randomization, although randomization also matches on unobserved covariates and there is no reason to assume that nonrandomized methods achieve balance on $\mathbf{u}$ (Rosenbaum, 2010). Interestingly, covariate balance with estimated propensity scores is also typically somewhat better than with true propensity scores (Rosenbaum, 1987; Rubin & Thomas, 1996).

There are a number of empirical guidelines for establishing the adequacy of balance between groups. Two classical measures are Rubin’s B , the number of standard deviations between the means of the groups, and Rubin’s R , is the ratio of treatment variance to control variance (Rubin, 2001). More recently, methodologists have proposed the absolute standardized difference (Rosenbaum, 2010), the absolute standardized difference in covariance means (Haviland et al., 2007), and normalized difference scores (Imbens & Wooldridge, 2009).

Graphical methods are also useful (e.g., Ridgeway et al., 2017; Rosenbaum, 2010). Hypothesis testing and p -values are inappropriate for use with covariate balance, since they are only in-sample statistics with no reference population (Stuart, 2010). Insufficient covariate balance can be a sign that the propensity score model is improperly specified. If a model resulting in balance cannot be found, it may indicate that subsequent inferences are not valid (Harder et al., 2010).

The Region of Common Support

In randomized experiments, all members of the sample have a nonzero chance of assignment to the control and treatment conditions. In order for the propensity score model to enable the assumption of strong ignorability, there must be sufficient overlap in the distributions of the propensity scores between the control and treatment groups. The overlapping area of the estimated propensity score distributions between the lowest estimated values of $\hat{e}(\mathbf{x})$ in the treatment group and the highest values of $\hat{e}(\mathbf{x})$ in the comparison group is known as the region of common support. In the counterfactual tradition, identification of the average treatment effect, or ATE (Equation 3.1) is predicated on the possibility of members of both the control and treatment groups being exposed to both potential outcomes, Y^0 and Y^1 .

Accordingly, the analyst must discard cases in either group without an equivalent propensity score in the other group, as the counterfactual model is predicated upon cases in both groups with equal probabilities of selection (Rubin, 2001). Discarding cases outside of the region of common support reduces statistical power and limits generalizability to certain ill-defined subpopulations (Rubin, 2008; Shadish & Steiner, 2010), but without common support, the analysis becomes an extrapolation rather than the desired and assumed interpolation between overlapping distributions (Haviland et al., 2007). In some applications, insufficient overlap or unsuitable covariate balance may indicate that a propensity score analysis is unlikely to produce

trustworthy results (Harder et al., 2010; Ho et al., 2007). This situation arises more often when the assignment mechanism is inconsistent across the population under study (Rubin, 2008) and when a characteristic of the treatment group related to the predictors in the propensity score model makes them highly likely to receive the treatment.

Notably, estimation of the average treatment effect on the treated (ATT) only requires “weak overlap,” defined as $\Pr(Z = 1 \mid \mathbf{x}) < 1$ (Imbens, 2004). This assumption is sufficient for estimation of the ATT because cases with a zero probability of treatment selection, while problematic for calculation of an ATE, will not figure in to the ATT. Furthermore, ATTs require a weaker ignorability assumption of “partial ignorability” (Morgan & Winship, 2007). If $Y^1 \perp\!\!\!\perp Z \mid \mathbf{x}$, then the ATT is unbiased by selection even when Equation 3.3 (which establishes strong ignorability for the ATE) does not hold.

Analytic Approaches Using Estimated Propensity Scores

The estimated propensity score can be used to adjust for baseline differences between treatment and comparison groups in several ways. The most commonly used methods, reviewed here, are matching, subclassification, and weighting. Regression adjustment is less commonly used and not treated in this section. For information about regression adjustment, readers are encouraged to consult Rubin (1979) or Schafer and Kang (2008).

Matching

Matching is the most intuitive method for analyzing propensity score data. The concept is straightforward: create a pair of participants with equivalent propensity scores, one from the control group (individual k) and one from the treatment group (individual ℓ). Under the ignorability assumption, these cases represent both of the potential outcomes Y^0 and Y^1 for an

individual with a given distribution of the covariates $\mathbf{x}$. In matched pairs, however, there are observed rather than hypothetical values for both quantities, Y_k for the control case in the pair and Y_ℓ for the treatment case. Under typical propensity score assumptions, this enables testing of the counterfactual question of interest, as the two individuals are assumed to be exchangeable. Equation 3.1 for individual-level counterfactual differences δ_i can thus be rewritten as Equation 3.7:

$$\delta_i = Y_\ell - Y_k \mid \hat{e}(\mathbf{x}_\ell) \equiv \hat{e}(\mathbf{x}_k) \quad (3.7)$$

The estimated average treatment effect (ATE, Equation 3.8) across the sample is the average of the differences in the outcome between each matched pair:

$$\widehat{ATE} = \bar{\delta} = \sum \frac{Y_\ell - Y_k}{n} \quad (3.8)$$

While matching might seem routine to perform, in practice there are a number of complications that arise. Consequently, a number of matching algorithms have arisen, each with its own strengths and limitations (for reviews, see Guo & Fraser, 2015; Rosenbaum, 2010; and for a critique of matching, see King & Nielsen, 2018). The major differences are (1) whether to prioritize closer matches or prioritize matching more cases, and (2) what kinds of postmatching analyses are possible. Cases with propensity scores above or below the region of common support do not have matches and need to be discarded, potentially biasing the results (Rosenbaum & Rubin, 1985). Within the region of common support, 1:1 matching facilitates estimation and interpretation but has two drawbacks. Unless the group sizes are equal, some cases will be unmatched and discarded, which may alter the marginal distributions of some covariates and potentially also lead to bias (Rosenbaum & Rubin, 1985), and some matched pairs may have propensity scores that are further away than others. However, simpler matching algorithms enable more complex (i.e., multivariate) analyses.

A comparison between the greedy matching and optimal full matching algorithms, while not comprehensive, may be instructive. The greedy matching algorithm finds the pair of values of $\hat{e}(\mathbf{x})$ between the treatment and control groups that are the closest, matches them, pairs the remaining cases with the closest values of $\hat{e}(\mathbf{x})$, and so on, without reconsidering previous matches if the last few remaining cases do not have a good match in the other group. Greedy matching also discards any unmatched cases. In contrast, optimal full matching uses variable-ratio matching instead of 1:1 matching; for example, a participant in the treatment group could be matched to several controls with similar propensity scores, and vice versa. The algorithm seeks to use the full sample while minimizing the total distance between all of the matches. Greedy matching's drawbacks must be weighed against the fact that postmatching analysis can involve any type of regression-based model, while analysis after optimal matching is limited to comparison of means using the nonparametric Hodges-Lehmann Aligned Rank Test (Rosenbaum, 1993).

Subclassification

A generalization of matching is subclassification, originally proposed by Cochran (1968). Subclassification on propensity scores involves dividing the sample into some number of quantiles, or strata, with similar propensity scores within each. Subclassification provides unbiased causal estimates if the within-stratum differences are small enough that the propensity scores are effectively constant, thus approximating random assignment (Imbens & Wooldridge, 2009). Cochran found that using five strata removes 90% of bias, and many subsequent applications have used quintiles. A more current recommendation is to use a flexible number of strata to maximize overlap and minimize bias (e.g., Harder et al., 2010). Cases within smaller strata are more similar, but this must be considered against the loss of precision with smaller

within-stratum sample sizes. Strata that are too small may also fail to achieve balance on important individual covariates.

After dividing the sample into k strata, average treatment effects are calculated separately within each stratum and aggregated into a single estimate of the ATE:

$$\widehat{ATE} = \sum_{k=1}^K \frac{n_k}{N} [\bar{Y}_{1k} - \bar{Y}_{0k}] \quad (3.9)$$

An advantage of subclassifying on the propensity score is that, similar to greedy matching, the type of analysis is not restricted. Because cases are not discarded during stratification even with unequal control and treatment group sizes, subclassification is widely viewed as a sound approach for combining propensity scores with complex analyses, particularly structural equation models (Guo & Fraser, 2015). Subclassification also avoids one of the traditional challenges with propensity score weighting (reviewed next), which is the problem of outsized weights when the propensity score is near 0 or 1 (Schafer & Kang, 2008).

Weighting

An alternative to matching and subclassification is propensity score weighting (Austin & Stuart, 2015; Cole & Hernán, 2008; Lunceford & Davidian, 2004). Propensity score weighting involves using the inverse probability of receiving the treatment as a sampling weight w_i . These weights are known as inverse probability of treatment selection weights (IPTW). As Robins et al. (2000, p. 553) explained:

The denominator of w_i is informally the probability that a subject had his or her own observed treatment...The effect of weighting is to create a pseudopopulation consisting of w_i copies of each subject i . That is, if, for a given subject, $w_i = 4$, the subject contributes four

copies of him- or herself to the pseudopopulation. This new pseudopopulation has the following two important properties. First, in the pseudopopulation, unlike the actual population, [treatment selection] is unconfounded by the measured covariates [$\mathbf{x}$]. Second, [$\Pr(Z = 1)$ and $\Pr(Z = 0)$] in the pseudopopulation are the same as in the true study population so that the causal [estimands] are the same in both populations. Hence, it follows that we can unbiasedly estimate the causal [estimands] by a standard crude analysis in the pseudopopulation.

IPTW enables researchers to conduct simple or complex analyses while retaining all cases (Schafer & Kang, 2008). Combining IPTW with regression-based models is a common approach. The propensity score model and a subsequent regression model may or may not involve some of the same covariates, but covariates function differently in each step. Covariates in the weighting equation control for selection bias, while in the regression they serve to control for confounding of the treatment effect (Robins et al., 2000). Models in which either the propensity score model or the regression model sufficiently control for unobserved confounding are known as “doubly robust” (Bang & Robins, 2005; Kang & Schafer, 2007; Lunceford & Davidian, 2004; Robins & Rotnitzky, 1995). Although propensity score weighting was originally proposed by Rosenbaum (1987), it has only begun to appear regularly in published research since around 2008 (Austin & Stuart, 2015).

The weighting formula for estimating the average treatment effect is in Equation 3.10. The weights are the reciprocal of the probability of receiving treatment, not simply the reciprocal of the propensity score—a common misunderstanding (Austin & Stuart, 2015).

$$w_{ATE} = \frac{Z}{\hat{e}(x)} + \frac{1-Z}{1-\hat{e}(x)} \quad (3.10)$$

For cases in the treatment group ($Z=1$), Equation 3.10 reduces to $w = \frac{1}{\hat{e}(x)}$. For those in the control group ($Z=0$), the weights are $w = \frac{1}{1-\hat{e}(x)}$. The weights in Equation 3.10 can be used to calculate an average treatment effect as in Equation 3.11, which is reminiscent of the subclassification ATE estimator in Equation 3.9. This formula is the same as the original weight estimator for stratified survey sampling (Horvitz & Thompson, 1952), from which the concept for IPTW was developed. The ATE is the sum of each participant's value of Y_i , using the “pseudopopulation” created by propensity score weighting.

$$\widehat{ATE} = \frac{1}{N} \sum_{i=1}^N \left[\frac{Z_i Y_i}{\hat{e}(\mathbf{x}_i)} - \frac{(1 - Z_i) Y_i}{1 - \hat{e}(\mathbf{x}_i)} \right] \quad (3.11)$$

The same principles apply with respect to the ATT. The ATT weighting formula is:

$$w_{ATT} = Z + (1 - Z) \frac{\hat{e}(\mathbf{x})}{1 - \hat{e}(\mathbf{x})} \quad (3.12)$$

This reduces to $w = \frac{\hat{e}(\mathbf{x})}{1 - \hat{e}(\mathbf{x})}$ in the control group, and $w = 1$ in the treatment group. As the control group weight is the familiar $P/(1-P)$ odds formula, estimation of an ATT using propensity score weights is usually not called IPTW but rather “weighting by the odds.” By definition, treatment participants have 1:1 odds of being treated, and control participants will be “upweighted” when they have higher propensity scores (i.e., when they appear more similar to participants who did receive the treatment). The treatment effect on the treated (Equation 3.13) is again the sum of all cases, weighted by their odds of receiving the treatment condition:

$$\widehat{ATT} = \frac{\sum_{i=1}^n Y_i Z_i}{\sum_{i=1}^n Z_i} - \frac{\sum_{i=1}^n Y_i \cdot (1 - Z_i) \cdot \hat{e}(\mathbf{x})}{\sum_{i=1}^n (1 - Z_i) \cdot (1 - \hat{e}(\mathbf{x}))} = \sum_{i=1}^n Y_i - \sum_{i=1}^n \left[\frac{Y_i \cdot \hat{e}(\mathbf{x})}{1 - \hat{e}(\mathbf{x})} \right] \quad (3.13)$$

Technical information and derivations for weighting formulas can be found in Hirano and Imbens (2001), Hirano, Imbens, and Ridder (2003), and Imbens and Wooldridge (2009).

Correct specification of the propensity score model is essential for analyses incorporating propensity score weights (Robins et al., 2007; Schafer & Kang, 2008). One of the most pressing issues with weighted estimation is cases with propensity scores near 0 or 1, which can result in outlier cases with huge influence that make results unstable (Harder et al., 2010; Kang & Schafer, 2007; Rosenbaum, 1987). Two procedures that can ameliorate this issue are trimming and stabilization (Cole & Hernán, 2008; B. K. Lee et al., 2011). Trimming involves capping large weights at a certain threshold, such that no single case has an outlying influence. Stabilization involves multiplying all of the weights by a constant in order to reduce the occurrence of weights near the parameter boundaries.

However, methodologists (e.g., Burgette et al., 2015; Harder et al., 2010; Hirano & Imbens, 2001; Ridgeway & McCaffrey, 2007) increasingly recommend the use of a nonparametric estimator for propensity scores intended for use in weighting, as they tend to avoid the problem by producing fewer propensity scores near the bounds of 0 and 1. In one simulation study (B. K. Lee et al., 2011), boosting and ensemble methods worked as well as non-trimmed logistic regression, and trimming did not substantially help the nonparametric models, as they tend to have fewer extreme weights in the first place. Weighting by the odds to estimate an ATT is also less sensitive to outlier weights than using IPTW to estimate an ATE, since the propensity score is only in the denominator of the equation for the control group. Thus, if the ATT is a causal estimand of interest, one of the most robust methods for propensity score weighting is to combine GBM with weighting by the odds (Harder et al., 2010; Ridgeway & McCaffrey, 2007).

Weighting With Complex Samples

Multistage stratified sampling is often used to generate nationally representative samples in datasets made available by governmental agencies such as the National Institutes of Health and National Center for Education Statistics for use by researchers. In order to yield unbiased estimates representative of the population and not just the sample, propensity score estimation must account for the structure of multistage sampling designs at all stages of the analysis (Hahs-Vaughn, 2015; Lenis et al., 2018). Bias increases as treatment effects become more heterogeneous across the population (Ashmead, 2014).

There is disagreement in the literature about how to account for the sampling design during propensity score estimation. The most rigorous method is to use a design-based method such as Taylor series linearization, or TSL (Hahs-Vaughn, 2015; Hahs-Vaughn & Onwuegbuzie, 2006; Ridgeway et al., 2015). A competing view (Dugoff et al., 2014; Zanutto, 2006) is that since the propensity scores themselves are not intended to be generalizable, and variance parameters from their estimation are not used, TSL may not be necessary. In one simulation, Lenis et al. (2017) found that the particular method of incorporating survey weights into propensity score estimation did not have an appreciable effect on outcome bias. It may be more expedient to merely include the survey weights, stratum or cluster identifiers, etc. as predictors in the propensity score model. On the other hand, Ashmead's (2014) simulation found that bias increased when design weights were not used to estimate the propensity scores. Whatever the method, misspecified models or those resulting in poor covariate balance yield biased estimates in both simple and complex samples (Lenis et al., 2017, 2018).

There is unanimity in the literature that sampling weights must be used in the outcome analysis. Methods for propensity score matching with complex samples have been examined by

Dugoff et al. (2014), Lenis et al. (2017) and Austin et al. (2018). However, because propensity score weighting already incorporates design-based weights into the analysis, it is a natural candidate for use with complex samples. Propensity score weights can be adjusted for the sample design by multiplying each participant's IPTW (or, alternatively, odds of treatment) times her or his survey sampling weight w_i (Hahs-Vaughn & Onwuegbuzie, 2006; Ridgeway et al., 2015).

Equation 3.12 uses weighting by the odds for survey-weighted estimation of an ATT.

$$\begin{aligned}
 \widehat{PATT} &= \frac{\sum_{i=1}^n Y_i \cdot Z_i \cdot (1/w_i)}{\sum_{i=1}^n Z_i \cdot (1/w_i)} - \frac{\sum_{i=1}^n Y_i \cdot (1 - Z_i) \cdot \hat{e}(\mathbf{x}) \cdot (1/w_i)}{\sum_{i=1}^n (1 - Z_i) \cdot (1 - \hat{e}(\mathbf{x})) \cdot (1/w_i)} \\
 &= \sum_{i=1}^n Y_i \cdot \left(\frac{1}{w_i} \right) - \frac{\sum_{i=1}^n Y_i \cdot \hat{e}(\mathbf{x}) \cdot (1/w_i)}{\sum_{i=1}^n (1 - \hat{e}(\mathbf{x})) \cdot (1/w_i)} \quad (3.14) \\
 &= \sum \left[\frac{\bar{Y}_{treated}}{SamplingWeight} \right] - \sum \left[\frac{\bar{Y}_{control} \cdot OddsOfTreatment}{SamplingWeight} \right]
 \end{aligned}$$

The use of propensity scores to estimate unconfounded treatment effects in multistage stratified samples is an ongoing research area without definitive best practices. Simulation studies of “complex samples” do not always use a sample as complex as those found in NCES datasets (Austin et al., 2018). Furthermore, even the most recent literature does not address challenges or recommendations when the propensity scores are estimated with advanced CART-based techniques. Still, due to the near certainty of bias when survey weights are ignored, it seems prudent to use them if population-level inferences are desired.

Sensitivity Analysis

The foundational assumption for propensity score analysis to be unbiased is that the observed covariates adequately predict treatment selection. However, as elaborated earlier, this

assumption is susceptible to hidden bias, or the effects of a covariate u that is related to treatment assignment but not incorporated into propensity estimation. The effect of hidden bias is that individuals who appear to be the same based on $\mathbf{x}$ or $\hat{e}(\mathbf{x})$ still have unequal probabilities of treatment selection (McCaffrey et al., 2004). In some instances, even a small amount of hidden bias would have a noticeable impact on the results. In others, the hidden bias would have to be enormous to change a study's conclusions. If the potential effects of u can be shown to be small, the study's causal conclusions are strengthened (Rosenbaum, 2010).

The critical parameter in sensitivity analysis is Γ , the degree to which the naïve model's assumption that two subjects k and ℓ with equal propensity scores have the same probability of having received the treatment (i.e., $\pi_k = \pi_\ell \mid \mathbf{x}_k \equiv \mathbf{x}_\ell$) is false. Γ is a multiplier ≥ 1 where subject k could be up to Γ times more likely than subject ℓ to receive the treatment due to bias from u . $\Gamma = 1$ is the naïve model: balancing on observed covariates, subjects k and ℓ were equally likely to receive the treatment. A value of $\Gamma = 2$ implies that subject k 's odds of receiving the treatment were twice as high as subject ℓ even though $\mathbf{x}_k \equiv \mathbf{x}_\ell$ (Rosenbaum, 2010). Sensitivity analysis involves calculating the value of Γ that would result in a change in the study's conclusions. For example, a sensitivity analysis might find that values of Γ less than 3 do not cause the treatment effect p -value to rise above $\alpha = .05$, but a hidden bias of $\Gamma \geq 3$ does render the treatment effect nonsignificant. Put another way, the study's finding of a significant treatment effect is robust to hidden bias that causes some subjects to be up to three times as likely as others with equivalent propensity scores (a large bias) to receive the treatment, but explicitly leaves open the possibility that a confounder with a large influence could be at play. This allows the reader to draw conclusions based upon both the magnitude and robustness of the effect.

Although sensitivity analysis is most developed and commonly applied for point estimates and effect sizes in studies involving matching, it has also been described for propensity score weighting (e.g., Brumback et al., 2004; Cole et al., 2005; Shen et al., 2011), and implemented (but still under development) in R software (Ridgeway, 2006; Ridgeway et al., 2017). Hidden bias is the central impediment to causal inference with propensity scores. As sensitivity analysis makes the potential effects of hidden bias transparent (Haviland et al., 2007), it is often a recommended step in propensity score analysis. Still, sensitivity analysis does not determine whether the bias actually exists (Robins, 1999; Steiner et al., 2010). Thus, while it is an important step toward causal conclusions, it is not a cure-all.

Summary

Propensity score adjustment is a technique for bias reduction in observational studies, supporting causal inferences under the assumption of no unobserved confounding when experiments are impossible or infeasible. Propensity score methods involve the reduction of a vector of observed covariates that predict treatment selection into a single dimension, which is designed to yield comparable distributions of these variables between treatment and control groups such that the observed variables are unrelated to treatment selection. The many algorithms for estimation, matching, subclassification, weighting, and sensitivity analysis offer the researcher a wide variety of techniques. Methodologists have expressed contradictory preferences for certain propensity score methods relative to others (e.g., Austin, 2009; Austin et al., 2007; Austin & Mamdani, 2006; Leite et al., 2018; Lunceford & Davidian, 2004; Schafer & Kang, 2008; Smith & Todd, 2001; Stuart, 2010), and there is no consensus as to the best way to reduce bias using them.

Even though the assumptions of SUTVA and unconfoundedness are often at least somewhat violated in practice, the results of a rigorous propensity score analysis may at worst be “a reasonable approximation” (Hirano & Imbens, 2001, p. 262) of the true causal forces at work. In the absence of the ability to conduct the large-scale randomized experiments they would have *preferred* to conduct (Rubin, 2007), propensity scores are one of the tools for causal inference most likely to be usable to music education researchers given their typical research questions and available data. An RCT is ideal for establishing a causal relationship, but propensity score analysis is undeniably more rigorous than making no attempt to account for selection bias. Rosenbaum (2010, p. 69) observed, “That is what you do with ideals: you attempt to move towards them.” Chapter 4 provides the details of how this study attempted to move toward the ideal of an unconfounded examination of music’s putative benefits.

Chapter 4: Research Methods and Design

Overview

Leaders in the field of causal inference in the potential outcomes tradition (e.g., Rosenbaum, 2010; Rubin, 2008) advocate observational studies designed to recreate the conditions of the hypothetical experiments they would have conducted if an experiment were possible. For this study, the experiment I would have conducted would be to randomly assign students to either not participate or participate in their home schools' music programs, under the conditions in which the students in those programs typically participate, and examine any differences in their academic or social-emotional outcomes. Since randomization was impossible for this study, I attempted to replicate randomization by using propensity scores to control for observed selection bias and selecting the ATT rather than an ATE as the estimand. Prior to undertaking this research, I sought and received the approval of Ohio State University's Institutional Review Board for the Behavioral and Social Sciences.

Data

Data Source

Overview

The data source for this study was the High School Longitudinal Study of 2009, or HSLs:09 (Duprey et al., 2018; Ingels et al., 2011, 2013, 2015), which was conducted by the National Center for Education Statistics (NCES) at the Institute of Education Sciences in the United States Department of Education. The HSLs:09 is NCES's most recent longitudinal study

covering students' high school years. It follows the Educational Longitudinal Study of 2002 (ELS:2002), the National Education Longitudinal Study of 1988 (NELS:88), the High School and Beyond Longitudinal Study of 1980 (HS&B:80), and the National Longitudinal Study of 1972 (NLS-72). I also utilized supplementary school-level information from NCES's Common Core of Data (Chen et al., 2012), a repository of basic school- and district-level data for all schools associated with the Department of Education.

Waves of the HSLS:09

Because it was a longitudinal study, data collection for the HSLS:09 took place in multiple waves. Base year (BY) data collection took place in the fall of 2009. The first follow-up wave (F1) occurred during the spring of students' 11th-grade years, in 2012. Both the BY and F1 waves included extensive survey interviews, including questions from which the social-emotional scale scores were derived, with sampled students and their parents or guardians, and one counselor and one administrator from each school. Student participants also completed a proprietary mathematics reasoning exam during BY and F1 data collection. Data collection for the 2013 Update wave (U13) took place in the summer and fall of 2013, focusing on college entrance exam scores, postsecondary plans, and participants' high school transcripts. A second follow-up in 2016, which was not analyzed for this study, added postsecondary transcripts and financial aid records.

Foci of the HSLS:09

While the HSLS:09 was designed to be applicable to a broad range of educational researchers and decision-makers, it emphasized students' abilities and experiences in science, technology, engineering, and mathematics (STEM), and their transitions from high school into college and the work force (Duprey et al., 2018). As with any general-purpose dataset, the study

designers needed to balance a desire for comprehensiveness with a need to avoid participant fatigue, which limited the number of questions that were asked. Undoubtedly, each researcher undertaking a study with an NCES dataset wishes that just a few more questions could have been added to the surveys, but federal funding for nationally representative in-depth studies for each individual subject area has not been allocated. The lack of music-specific information must be weighed against the scale and quality of the other information available in the HSLS:09, particularly considering the lack of an alternative.

Sampling Procedure

The sample frame for the HSLS:09 was “regular” public schools,⁹ charter schools, and private schools in all 50 states and the District of Columbia. Sampling involved a two-stage stratified design. In the first stage, schools were randomly sampled within strata according to school type (public or private), region of the country, and locale or urbanicity. In the second stage, students within each school were randomly sampled. The study team oversampled certain groups of students to ensure accuracy, and sampled additional schools into a “release pool” that guarded against school-level attrition. An average of approximately 27 students from each school’s ninth grade enrollment lists were surveyed.

In addition, the sampling procedure was structured to facilitate state-level reporting for ten larger states (California, Florida, Georgia, Michigan, North Carolina, Ohio, Pennsylvania, Tennessee, Texas, and Washington). The BY sample included 941 schools, resulting in an initial sample size of $N = 25,206$ student participants entering high school in the fall of 2009. By 2013,

⁹ A “regular” public school for this study excluded nontraditional school settings; e.g., schools in detention centers, and schools serving only homeschooled or only special education students.

25,167 students remained study-eligible and had a transcript entered into the database. NCES personnel undertook intensive efforts to ensure a representative sample, reduce nonresponse, and collect accurate data (Duprey et al., 2018).

High School Transcripts

As the analysis centered on participants' music coursetaking, processing of the HSLS:09's transcript data file was critical. A team of coders was hired by NCES to review each transcript collected for U13 and enter a classification number for each class's subject matter according to the School Courses for the Exchange of Data system, or SCED (Bradby et al., 2007). The coders also recorded each transcript entry's term and final course grade.

Restricted-Use Data

Because the HSLS:09 included confidential and personally identifying information, there are two versions of the data. A public-use file is available to the public, from which identifying information has been redacted and the high school transcripts have been removed. This dataset was insufficient to answer my research questions. There is also a restricted-use data file licensed only to researchers after review of an application. The application involves a description of the research and research team, a Security Plan, and notarized Affidavits of Nondisclosure from study personnel. Research reports containing restricted-use data must use rounded sample sizes and are reviewed by NCES to ensure that personally identifying information is not disclosed. I requested and received a data license, abided by the security procedures, and submitted this report to NCES for publication approval.

Sample Restrictions

This study was intended to examine the effects of music instruction in traditional public schools. Specialty and alternative schools were not sampled by NCES. Students who attended

private or charter schools, or career and technical education centers, during one or more terms were excluded from the analysis. I also excluded students who changed schools during the study period, dropped out of school, or stopped attending for one or more semesters (including early graduation). The transcript file contained a small number of courses from before or after participants' typical four years of high school. I limited the analysis to courses occurring between the fall of 2009 and spring of 2013. Since they represented two primary dependent measures, I also excluded students who did not take any college entrance examinations.

Data Preparation

The data file for the HSLs:09 contained several waves of data for most variables. In order to simplify the analysis and interpretation, I sought to reduce participant demographic variables to a single value for each. I wrote Stata programs titled `cors`, `x12`, and `fix12` to assist me with examining the variables' distributions at Time 1 and Time 2, calculating Cronbach's α to ensure that the nonmissing values covaried, and specifying whether the BY or the F1 value was appropriate. Code for the Stata programs is in Appendix C. In most cases, I used the BY value when it was nonmissing and substituted the F1 value when BY was missing. At other times, I used the higher of the two values even when they were both observed (e.g., parents' education level). I used the same approach with other student- and school-level variables that could reasonably be assumed to be relatively constant from 2009 to 2011. In addition to addressing transcripts and missingness, which are detailed subsequently, data preparation involved recoding some categorical variables with sparse values and examining each variable's univariate distribution to ensure that it was appropriate to use in the analysis. I used Stata 14 (StataCorp, 2015) for data preparation and descriptive analyses.

High School Transcripts

A preliminary examination of the coded high school transcript information revealed enough misclassified music courses to warrant extensive efforts aimed at correctly characterizing each. I began the process of cleaning the transcript file by identifying a series of keywords that would be likely to indicate that a class would belong to each music SCED code. One example is that the word “wind” would likely indicate a band course. I added to the list of keywords throughout the course cleaning process. The full list of keywords is in Table 4.1. Additional variations on search terms were used. Some search terms were used in combination with the prefix “mus” in order to exclude, for example, English “comp” classes that overwhelmed the music results.

I began my process by reviewing the list of course titles that were classified under each of the music SCED codes. Some of the classes were music courses but almost certainly in the wrong category; e.g., an orchestra course titled “wind ensemble.” I wrote a Stata program, titled `cleanup`, to search the course titles for a given SCED code for keywords that would indicate that they may have been misclassified and assign them to the correct music type. Code these Stata programs is in Appendix C. I then reviewed each course title not containing the SCED’s keywords, which was facilitated by a program I wrote called `wordlist`, and manually reclassified each course whose title did not contain those keywords but still appeared to be misclassified using a program I wrote called `swap`. In some cases, this was facilitated by other information in the transcripts; e.g., an ambiguous course titled “Jazztones” taken in a student’s senior year, when the student had taken three previous years of band and no other music courses,

Table 4.1

Initial Course Classification Information

Music Subject	SCED Code(s)	Representative Search Terms	Category for Analysis
Band	05101, 05102, 05103	“band” “wind” “drum” “perc” “march” “guard” “flag” “flute” “clarinet” “sax” “oboe” “bassoon” “trumpet” “trombone”	Instrumental
Orchestra	05104	“orch” “string” “violin” “viola” “cello” “bass” “early music”	Instrumental
Jazz	05105	“jazz” “improv” “stage”	Instrumental
Instrumental Music	05106, 05109	“instr” “inst mus” “instru tecnqs” “instren” “inen”	Instrumental
Choir	05110, 05111, 05112, 05121	“cho” “chrs” “ch mus” “voc” “voice” “sing” “madrigal” “glee” “acap” “capp” “barbershop” “men” “women” “mix” “bel canto” “treble” “baritone”	Vocal
Music Theory, History, or Appreciation	05113, 05114, 05115, 05116, 05117, 05118	“theory” “apprec” “ib music” “history” “ear” “pop music” “general” “gen” “harm”	General
Piano	05107	“piano” “pno” “keyboard” “kbd” “keybd”	General
Guitar	05108	“guitar” “gtr”	General
Composition	05119	“comp” “songw” “garage”	General
Music Technology	05123, 05124	“midi” “tech” “prod” “record”	General
Music Business	05148	“bus” “indus”	Not analyzed
Music—Other	05124, 05147, 05149		General

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

seemed likely to be a jazz band rather than, for example, vocal jazz. I wrote another program, `sked`, to quickly review the students' coursetaking history. I also used `sked` to see the coursetaking history of other sampled students in the class, which was useful for classification.

After the cleanup of courses initially coded as music classes, I had an extensive list of keywords. I then searched the entire list of courses initially coded as non-music classes to locate course titles containing those keywords using another program I wrote, `orphan`. I reclassified many such courses. Afterward, I went through the cleanup process again until I was confident that the music courses in my working dataset were classified accurately. It is possible that I missed some music courses that were given other SCED codes, but any such courses would have to have had titles that were different from any of those identified by the HSLS staff or my search process—i.e., highly idiosyncratic.

The transcript file had a length for each course. My initial approach was to convert these to year units, with a year-long course counted as 1, a semester-long course as 0.5, a trimester-long course as 0.33, and a quarter-long course as 0.25. 1.42% of the music courses in the database had an unspecified length, which I assumed to be one semester as that was the modal value (72%). I assigned 0.25 year-units to summer courses. Next, I summed the course lengths for each subject (instrumental, choral, and general) to create three new variables measuring total years of participation. The resulting distributions had wide variation, with some students reported as having more than ten semesters in a subject.

Upon closer examination, I found that discrepancies in individual schools' course coding was responsible for at least some of the differences. For example, marching band was sometimes a separate credit-bearing course and sometimes not. This represented unobserved confounding in the critical measurement of music participation. Whether, for example, marching band was

included separately in the transcript had resulted in some instrumental music being double-counted for bureaucratic rather than substantive reasons. Another example was orchestra students who had both “orchestra” and “string technique” on their transcripts each semester. There was no way to know if other students had supplemental string technique courses that were equivalent in practice but not listed on their transcripts, or whether “string technique” represented as much music participation as additionally registering for band, which would have counted the same. This relates to one of the limitations of general-purpose datasets raised in Chapter 1: a music-focused survey would have probed the characteristics of the sampled music programs, but the HSLS:09 did not. I decided not to use the course-counting measures.

I subsequently pursued an alternative strategy for operationalizing the extent of students’ musical involvement by counting the number of different school years in which students enrolled in a music course in each category, regardless of the number of registrations they had in that year. Each student then had three variables for participation in instrumental music, vocal music, and general music that ranged from zero (did not participate in that subject) to four (took at least one course in that subject each year of high school). This operationalization meant that, for example, a student registered for an additional credit-bearing jazz band course was not measured differently than if the student had participated only in concert band. This measure appeared to be much more consistent across schools, which outweighed the loss of information in the instances where “double-counting” would have been appropriate.

Afterward, I checked to see that the participants had completed all of the music courses. I reclassified music courses for which students had earned grades of F (1.29% of music courses), “Unsatisfactory” (.07%), “Withdrew” (.07%), or “Incomplete” (.02%) as non-music courses so that they would not register as music participation. Since over 98% of the music class grades

were not one of these conditions (itself a remarkable finding deserving of further study), music courses with a recorded transcript grade of “Missing” (2.1%) were assumed passing and retained in the dataset. Although music course grades are far from a perfect measure of treatment compliance (Shaw, 2018), restricting the operational definition of music participation to having taken and passed the class was a feasible first step toward the assumption of fidelity.

Missing Data

Overview

Missing data is inevitable with large-scale surveys. Missingness is commonly defined as the case of a “true” value that is missing due to nonresponse rather than a reason like a logical skip (Schafer & Graham, 2002). Missingness in longitudinal studies can be caused by unit nonresponse (e.g., attrition) or by item response (e.g., a skipped question). A critical step in handling missing data, first described by Rubin (1976), is the assumption the analyst makes about the relationship between the missingness pattern R , the observed dependent variables X , and the outcomes Y (Schafer & Graham, 2002). The situation in which the missingness R is correlated with the outcome Y even after accounting for the observed variables X is known as MNAR, or Missing Not at Random. Under MNAR, complete-case analyses of Y are biased by the nonresponse pattern. As an example, suppose that the lowest-performing students in the hypothetical CANDY program introduced in Chapter 3 had unobserved discipline problems that were made worse by all of the candy and were suspended during the reading post-test, resulting in missing data. The observed values of Y would then be biased upward, and conclusions drawn from any analysis would be dubious.

If R is related to X or even Y , as it often is, the resulting inferences may still be unbiased. Bias results only when R is related to Y in ways not captured by X . In other words, in many

situations, X sufficiently explains the missingness such that the data may be assumed Missing at Random, or MAR. Continuing the example, if students' disciplinary histories had been included in the analysis, then observed estimates of the reading score may well not have been biased after conditioning on disciplinary history and all other predictors. A special case of MAR is Missing Completely at Random (MCAR), in which the missingness pattern R is assumed to be independent not only of Y but also independent of X .

Brief Review of Missing Data Methods

A common approach to missing data is to delete cases with missing values, but this not recommended by modern methodologists (J. W. Graham, 2009). Under MCAR, the best-case scenario, listwise deletion is inefficient. In all other situations, listwise deletion can result in bias. Imputation of missing values to generate complete cases helps to retain power and avoid bias. Imputation involves creating a Bayesian regression model for each predictor using the available data for each variable from all cases. Single imputation, or using the model to simply predict the missing values in the original dataset, can result in standard errors that are biased downward, since “imputed values from single imputation always lie right on the regression line” (J. W. Graham, 2009, p. 556). Multiple imputation algorithms create several datasets that add a measure of uncertainty to the imputed values, leading to more honest standard error calculations (Allison, 2001).

Traditional approaches to multiple imputation assume a multivariate normal distribution. This is rarely defensible with large datasets containing many types of variables, often with dependencies among them: “Constructing a joint distribution for all the variables with missing values with all such complexities is nearly impossible” (Raghunathan et al., 2018, p. 10). A more flexible method is fully conditional specification, also known as chained equations and

sequential regression (Enders et al., 2018; Raghunathan et al., 2018). Rather than imputing all variables at once from a joint multivariate distribution, fully conditional specification draws from a series of univariate posterior distributions for each variable, allowing the most accurate imputation of each. The algorithm continues to update as all of the other variables are imputed, using all available data for progressively more realistic estimates of each case's missing values and the uncertainty with which they were estimated.

Missing Data Strategy in This Study

The HSLS:09 experienced missing data from unit nonresponse, attrition, and item nonresponse. Analytic weights were calculated to reduce bias from unit nonresponse and the study team reported greater than 90% bias reduction across variables (Ingels et al., 2013). The study team also examined item nonresponse and found that while there were identifiable differences between responders and nonresponders, the differences were small and unlikely to adversely affect analyses. Certain variables were imputed using a combination of other items where applicable, while others were imputed using weighted sequential hot-deck imputation using a CART (Classification And Regression Trees) algorithm. In instances of unit nonresponse, such as parents who refused even an abbreviated interview, imputation was not performed by the study team. The final data file contained a predictable amount of missing data, which was greatly reduced when excluding nonqualifying schools and students without complete transcripts. Information on missingness can be found in Appendix B, with further details available in the documentation from NCES (e.g., Duprey et al., 2018; Ingels et al., 2011). Given the extensive set of observed variables and intensive efforts undertaken by the study team to reduce nonresponse, it seemed reasonable to assume that the missingness was MAR.

Still, in light of NCES's documentation of differences between the complete and incomplete cases, I used two procedures to fill in missing values to generate complete cases for analysis and avoid casewise deletion. The first strategy was to substitute a known value from a different data source. Certain survey questions (e.g., demographics) were asked at more than one wave. If the variable was missing at BY but observed at F1, I used the F1 value per the HSLS:09 data manual (Ingels et al., 2013) as detailed previously. I was also able to incorporate certain school-level variables in NCES's Common Core of Data (Chen et al., 2012) with the HSLS:09 dataset using a common NCES school identifier.

Where there was still missingness, I employed multiple imputation to estimate a set of $m = 20$ complete datasets containing estimated values for the missing data using chained equations in IVEware 0.3 (Raghunathan et al., 2016). I specified the imputation model to account for the complex sampling design and dependencies and restrictions among the variables. I began with a fully conditional model (i.e., all of the variables are used to predict all of the other variables), with 10 iterations for each of the 20 imputed datasets, but this proved computationally prohibitive as the procedure would have taken more than three weeks to run. I resolved this issue by using a stepwise regression procedure that stopped adding predictors when the marginal R^2 was less than .01 (Raghunathan et al., 2018). This had the effect of greatly speeding up the computation time by excluding predictors that had insignificant marginal contributions to the equations.

I used the imputed datasets for all descriptive and inferential analyses, including propensity score estimation, that accounted for the uncertainty associated with imputed data. The only instance in which I used listwise deletion was the transcript information. Since the analytic strategy revolved around propensity scores, itself a prediction model for music participation, I

was worried that imputing the music participation variables would result in overfitting and distort the propensity score estimates. I also did not impute missing values for the college entrance examination scores, since many students do not take these tests and there was no way to determine which missing values resulted from nonparticipation. The analyses were limited to cases with observed values on those variables.

Analytic Strategy

Attenuating Selection Bias

In order to infer a causal relationship between music instruction and any purported benefits, a study must have strong internal validity, defined as the extent to which it establishes that music participation alone led to any differences in the outcomes (Shadish et al., 2002). Experimental studies involving randomized treatment assignment are one way to ensure that the treatment under study, not pre-existing differences between the control and treatment groups, was the sole reason for any outcome differences (Murnane & Willett, 2011). However, in this observational study, I could not assign the student participants to certain classes. Rather, they mainly self-selected into or out of them for any number of reasons, many of which may not be reflected in the dataset. Extant research has demonstrated a host of differences between students who typically enroll in music and the general student population (e.g., Elpus & Abril, 2011, 2019; Kinney, 2010, 2018). Estimation of the effects of music instruction is therefore susceptible to confounding by systematic pre-treatment differences between control and treatment groups, known as selection bias (Lohr, 1999; B. Schneider et al., 2007). Comparisons between music and non-music students that do not adjust for these differences are suspect.

In order to mitigate selection bias, I used propensity scores to render treatment selection independent of the outcomes conditional on the observed variables in the dataset. Propensity

score analysis (Guo & Fraser, 2015; Rosenbaum, 2010; Rubin, 2006) involves reducing a vector of covariates into a single value representing each individual's probability of having received the treatment. Thus, each treated participant's outcomes can be compared with another participant from the control group who has a comparable distribution of all of the observed variables in the dataset. Under certain assumptions, propensity score adjustment can support causal inferences by rendering treatment effects independent of effects related to selection into the treatment.

Plausibility of Propensity Score Assumptions

Propensity score analysis is based on the potential outcomes framework (e.g., Holland, 1986; Morgan & Winship, 2007) in which each participant has an observed outcome in either the treatment or control state, and an unobserved potential outcome (the counterfactual) in the other state. The theory is that the observed values of the outcome for control and treatment participants with equivalent propensity scores represent the potential outcomes for any individual with that same propensity score. The averaged difference between the treatment and control outcomes is then the effect of the treatment. While this is an elegant theory of inferred causality, it requires certain assumptions that can be difficult to meet in applied research.

Strong Ignorability

The first assumption is termed “strong ignorability” (Rosenbaum & Rubin, 1983b), which operationalizes the potential outcomes framework detailed in Chapter 3. Two conditions are required to establish ignorability. First, the observed covariates must adequately predict treatment selection such that the potential outcomes are independent of pre-treatment differences. Second, each case must have a nonzero probability of receiving the treatment. I addressed this second condition by ensuring that each student functioning as a control attended a school that offered the particular type of music course (instrumental, choral, or general music) used in each

analysis. The only way to accomplish this was to see if, for each sampled school, any sampled students had enrolled in each type of music. This was a conservative approach, as it was likely that the school-level HSLS sample did not capture each school's entire course catalog. 12.1% of schools were classified as having no vocal music program, 10.0% as having no instrumental music program, and 29.5% as having no general music program. As students within each school were selected at random, this may have reduced efficiency but probably did not introduce bias. I assumed that if a school offered a type of music, there was a possibility, however small, that a student could have pursued it.

A strong theoretical and empirical knowledge of what predicts selection into music would aid in satisfying the first condition (Morgan & Winship, 2007). However, while the extant research into enrollment and persistence in school music programs is valuable and ongoing, it is a stretch to say that the individual-level selection mechanisms are “known” at this time. Given the myriad individual-level and school-level factors leading to decisions about course enrollments, it may be impossible for researchers to ever completely “know” what variables predict enrollment in music. Therefore, it is difficult to claim that this variable set, or any variable set, explains treatment selection sufficiently to rule out the possible presence of hidden bias, or the confounding effects of unobserved forces related to both treatment selection and the outcomes (Steiner & Cook, 2013). Unfortunately, there is no test for the assumption of no unobserved confounding related to selection (Shadish & Steiner, 2010; Steiner et al., 2010). Even though hidden bias cannot be completely ruled out, this study can provide some causal evidence. The HSLS:09 contains a broad range of covariates known to be related to both selection into music courses and the outcome variables. It also contains limited information on academic achievement

from before high school. These types of measures have been shown to substantially reduce both overt and hidden bias (Hallberg et al., 2018; Steiner et al., 2010; Stuart, 2010).

One issue left unresolved, however, is the looming effect of music participation prior to or outside of high school. The majority of students in high school music ensembles also enrolled in middle school music programs, and some likely began their time in music in elementary school ensembles. A student beginning to study her instrument in fifth grade has already been participating for four years before the high school orchestra participation that is the subject of this research. Any benefits of high school music inferred in this study are conflated with potential benefits of earlier musical involvement. They are also conflated with extracurricular and non-school music participation. There is a single general survey question about outside-of-school music participation at F1, but it is insufficient to differentiate between, for example, a student participating in a curricular concert band but not an extracurricular marching band, or between casual piano playing at home and serious piano study.

The issue of unobserved music participation has a second deleterious effect on potential-outcomes inference. If the variables in the propensity score model have been affected by the treatment, the unconfoundedness assumption may no longer hold (Rosenbaum, 1984). Many of the covariates in the HSLS:09 can be assumed to be independent of prior or non-school music participation, such as demographics, but others (e.g., 8th grade mathematics achievement or number of friends) may not. Rosenbaum describes this situation as a compromise between a less complete covariate set, which is susceptible to hidden bias, and the additional assumption that the effects of the treatment on the covariate are small relative to the covariate's importance to treatment selection or the outcome. For this study, given how early into students' high school careers the BY data collection occurred, and given that unobserved music participation is already

a limiting potential confounder, I chose to include all of the BY variables related to treatment selection in the propensity score model.

Stable Unit-Treatment Value Assumption

Another assumption related to potential outcomes is the stable unit-treatment value assumption, or SUTVA (Rubin, 1986). SUTVA implies that the treatment effect is constant regardless of the ways it is administered, whether and how other participants also receive the treatment, and any pre-treatment characteristics of the participants. None of the elements of SUTVA hold in this study. It is easy to imagine variations in the ways that individual schools' music programs operate (e.g., program size; competitive orientation; teacher experience and dispositions) being related to those programs' benefits, and also that those benefits accrue in a differential fashion according to participant characteristics such as personality, prior achievement, musical giftedness, or enrollment in private lessons. In addition, music's benefits may well change depending on how many and which students are enrolled. The violation of SUTVA is a limitation of this study. SUTVA will also be violated in any future studies of music participation that do not include measures of music "treatment" fidelity.

Consequences of Violated Assumptions

The fact that some of these assumptions are violated means that this study cannot provide ironclad inferences about the universal effects of music instruction in American high schools. Common sense would already lead to the conclusion that there is not a tidy universal relationship between phenomena as variegated as school music instruction and as complex and poorly understood as cognitive and social-emotional development. Still, it is important to understand the results of this study in light of these limitations on their generalizability, as unfounded causal claims can have pernicious effects on the field (O'Connell & Gray, 2011). For example, since

SUTVA does not hold, this study would not support the notion that any benefits resulting from music instruction would accrue to the same extent if music were required for all students, although it is easy to see how advocates or policymakers may reasonably arrive at such a conclusion.

Propensity Score Estimation

The last assumption is that the propensity score model is correctly specified. This can be difficult to accomplish using the traditional approach of logistic regression. Nonlinear relationships, possible interactions and polynomial terms, and a potentially large number of predictors make specification difficult, especially in the absence of empirical knowledge of the selection process. However, misspecification can result in biased conclusions (Heckman, Ichimura, Smith, et al., 1998; Heckman, Ichimura, & Todd, 1998; Lenis et al., 2018). For this study, I used generalized boosted modeling, or GBM (McCaffrey et al., 2004; Westreich et al., 2010), to estimate the propensity scores. GBM is an adaptive nonparametric classification method derived from the field of machine learning that, unlike regression, does not require specification of a functional form (Berk, 2006; Strobl et al., 2009). I estimated the propensity scores with the `twang` package (Ridgeway et al., 2017) for the R statistical environment (R Core Team, 2018). The variables I included in the propensity score model are listed in Appendix A. I addressed the few categorical variables that did not perform well by combining categories with small counts (e.g., not forcing balance on Dominican nationality in the variable `X1HISPTYPE`). I also split two skewed continuous variables into quartiles.

Following recommendations for propensity score estimation with complex survey data (e.g., Hahs-Vaughn & Onwuegbuzie, 2006; Ridgeway et al., 2015), I incorporated the design weights into the propensity score estimation. I limited the reservoir of control participants to

those who did not have any music participation listed on their transcripts (i.e., a choir student could not serve as a control when orchestra was the treatment). I ran the propensity score estimation a total of 120 times: short-term (1-2 years) and sustained (3-4) involvement for each of the three types of music detailed in the next section on the $m = 20$ imputed datasets.

Measures

The dependent measures I modeled are listed in Table 4.2. All of them were continuous.

Academic Outcomes

The first outcome was college entrance examination scores. While some students took the SAT and others the ACT, the ACT scores were converted to SAT units and reported as such in the HSLS:09 dataset. Only scores for reading and mathematics were reported, so those for writing, science, etc. could not be analyzed. The college entrance exam scores were censored at 200 and 800 (or 400 and 1600), but were otherwise normally distributed. I estimated these models using both OLS and truncated linear regression models. As the results were similar, I reported results from the OLS models for ease of interpretation. The score for mathematics reasoning was a scale score derived from a test that the HSLS:09 study team developed internally.

Social-Emotional Outcomes

Scores for each of the social-emotional variables were figured using item response theory, or IRT, with data from survey responses from all participants. As the HSLS:09 had a STEM focus, the social-emotional outcomes directly related to mathematics and science were measured during both the BY and F1 waves of the HSLS:09. In these instances, I used the F1 (11th grade) variable as the outcome and included the BY (9th grade) variable as a predictor. Regrettably, the broader constructs of school belongingness and engagement were measured only

in participants' 9th grade year, and motivation was measured only in 11th grade. As a result, the “treatment” of high school music participation cannot be said to have affected belongingness or engagement. While I was not able to consider these variables in the propensity-score-weighted analyses, I did regress them on applicable predictors such as demographics and achievement in eighth grade. I was also not able to include a pretest variable when modeling school motivation.

Table 4.2

Dependent Measures

Variable	Category	Grade(s) Measured
College Exam Mathematics Score	Academic	11–12
College Exam Reading Score	Academic	11–12
Mathematics Reasoning Score	Academic	9, 11
Sense of School Motivation	Social-Emotional	11
Mathematics Self-Efficacy	Social-Emotional	9, 11
Science Self-Efficacy	Social-Emotional	9, 11
Sense of School Belonging	Social-Emotional	9
Sense of School Engagement	Social-Emotional	9

Note. College entrance examination scores were reported in SAT score units.

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Certain individual-level covariates from the propensity score model were also included as predictors in the regression models. Of course, the most important predictor variables were related to curricular music participation, as measured by participants' high school transcripts. I combined the twenty-six music SCED codes into twelve categories as described in Table 4.1. However, further aggregation was necessary. Course offerings were not equivalent across the sample. Few, if any, schools offered all of the listed courses. A student enrolling in guitar may

have been equally likely to enroll in piano if it had been offered instead, or vice versa. Some schools offered band and orchestra instruction, some offered band only, and for some it was impossible to determine. In addition, many of the course titles were ambiguous, with unobservable course content likely spilling across tidy boundaries, and were much more likely to be accurately assigned to broad categories rather than specific courses. As a result, I created three subclasses of music participation: instrumental music, vocal music, and general music¹⁰. I chose to exclude the few courses coded as “music business” from the operational definition of music participation as they seemed less likely to involve music making or art-focused study.

For each subclass, I created two further categories for the duration of music participation: one for 1-2 years in the subject, and the other for 3-4 years.¹¹ This led to six “treatment” groups. In order to directly estimate the “treatment” effects of music participation, I used only students who never registered for a high school music course as the “control” group. Furthermore, to avoid possible confounding, I excluded $n = 60$ students who registered for more than one type of music from the analysis. The number of excluded students was small enough to have minimal effects on the results.

¹⁰ I acknowledge that the term “general music” is unsatisfying (e.g., Abril, 2017); however, it is currently widely recognized.

¹¹ Some of the outcome variables were measured during participants’ 11th grade years and some in their 12th grade years. Since it is foundational that outcomes should not affect the propensity scores, it would be a problem if a dependent variable measured during 11th grade could have predicted music participation in 12th grade. However, all students who took music for four years also took it for three years, alleviating this concern. Furthermore, this implies that the category of 3-4 years of music is safe to use for outcomes measured during 11th grade.

Analysis Using Propensity Score Weighting

In order to calculate an unbiased estimate of music enrollment on the outcomes, I combined propensity score weighting with regression modeling in a “doubly robust” approach, meaning that the results will be unbiased if either the propensity score model or analytic model is unbiased (Bang & Robins, 2005; Lunceford & Davidian, 2004; Schafer & Kang, 2008). With the ATT as the estimand, I used the “weighting by the odds” method (Austin & Stuart, 2015; Harder et al., 2010). Non-music students’ weights were figured using the estimated propensity score $\hat{e}(\mathbf{x})$ such that at a given probability of treatment selection, the weights were equal to $w = \frac{\hat{e}(\mathbf{x})}{1-\hat{e}(\mathbf{x})}$ while music students’ weights were all set to $w = 1$.

The combination of the weighting by the odds approach with propensity score estimation using GBM makes the resulting inferences robust to several of the traditional challenges with weighting (Harder et al., 2010; Ridgeway & McCaffrey, 2007). Still, propensity score weighting can be sensitive to outlier weights when calculating the inverse of a very small probability of treatment selection (Kang & Schafer, 2007; Robins et al., 2000). In light of this sensitivity, I trimmed the weights within each treatment group (Cole & Hernán, 2008; Schafer & Kang, 2008) by setting weights above the 99th percentile equal to the value at the 99th percentile. This reduced the possibility of potential undue influence from outliers while avoiding the bias associated with discarding cases. Stata’s `mi estimate: svy: regress` command produced results that accounted for the increased uncertainty of parameter estimates using the multiply imputed data. As propensity scores do not have variances, I manually averaged them rather than using aggregation procedures for multiple imputation that are focused on variance estimation.

Regression Models

The regression approach was to estimate three nested models two times each: “unadjusted,” using only the HSLs:09’s sample weights, and “adjusted,” which incorporated both the sampling and IPT weights. Model 1 was unconditional, and functionally equivalent to a t test:

$$\text{Outcome}_i = \alpha + \beta_1 \text{Music}_i + \varepsilon_i \quad (4.1)$$

Model 2 added a common set of pretreatment predictors. The first three were demographic variables: an indicator for whether the student was identified as male, an indicator for whether the student’s race was recorded as white, and a standardized, continuous measure of socioeconomic status (SES). I also added a square term for SES in the models for college entrance examination reading scores, as preliminary analyses showed that this fit the data well. Model 2 also included a continuous measure of students’ 9th grade school engagement, and whether they had received a grade of C, D, or F in their 8th grade mathematics or science course. I used the mathematics course grade when modeling college examination mathematics scores, mathematics reasoning, and mathematics self-efficacy. I used the science course grade for the other outcomes because it likely relied less on any single set of skills and dispositions. I also included pretest variables where applicable: 9th grade mathematics reasoning for the college entrance examination scores and 11th grade mathematics reasoning models, and the relevant 9th grade self-efficacy variable for the 11th grade mathematics and science self-efficacy models.

$$\begin{aligned} \text{Outcome}_i = & \alpha + \beta_1 \text{Music}_i + \gamma_1 \text{Male} + \gamma_2 \text{White} + \gamma_3 \text{SES} + \gamma_4 \text{SES}^2 \\ & + \gamma_5 \text{SchoolEngagement} + \gamma_6 \text{LowGrade} (+ \gamma_7 \text{Pretest}) + \varepsilon_i \end{aligned} \quad (4.2)$$

Model 3 added interaction terms between the music enrollment variable and the sex, race, socioeconomic status, and eighth grade achievement variables. It also included coefficients for the interaction between music enrollment and pretest variables when they were included in Model 2. When the pretest variable was 9th grade mathematics reasoning, I also included a square term based upon a preliminary specification search. Many of these terms proved nonsignificant after the models were run. I kept them in the models in order to keep the estimation strategy consistent and to estimate as accurate a comparison between the music groups and the comparison group as possible, since a parsimonious model was not central to the purpose or research questions.

$$\begin{aligned} \text{Outcome}_i = & \alpha + \beta_1 \text{Music}_i + \gamma_1 \text{Male} + \delta_1 \text{Male} * \text{Music} + \gamma_2 \text{White} + \delta_2 \text{White} * \text{Music} \quad (4.3) \\ & + \gamma_3 \text{SES} (+ \gamma_4 \text{SES}^2) + \gamma_5 \text{SchoolEngagement} + \gamma_6 \text{LowGrade} + \delta_3 \text{LowGrade} * \text{Music} \\ & (+ \gamma_7 \text{Pretest} + \delta_4 \text{Pretest} * \text{Music}) + \varepsilon_i \end{aligned}$$

I selected the ATT, the average treatment effect on the treated, as the causal estimand. As such, the results are not generalizable to all students, but only to those who did participate in their school's music programs during the study period.

Procedures for Incorporating the Complex Sampling Design

The multistage stratified cluster sampling design of the HSLS:09 required additional steps to make valid inferences (Lohr, 1999; Stapleton & Thomas, 2008). Students were grouped by a primary sampling unit, or PSU, at the school level, within strata constructed by region of the country, locale or urbanicity, and school type (i.e., public, Catholic, or other private). Analytic weights were used to account for probability of selection and survey nonresponse (Heeringa et al., 2017). I used Stata's suite of `svy:` commands to use the HSLS:09's design weights for

accurate variance estimation using Taylor series linearization and to ensure national representativeness. Since the data were multiply imputed, I used the command `mi estimate: svy: regress.`

Since the analysis involved both propensity score weights and survey analytic weights, the final propensity-score-adjusted weights for analysis were the product of the two. Each individual's survey design weight is a multiplier indicating how heavily the individual should be considered in order for the analysis to be representative of the target population (Lohr, 1999). Each individual's propensity score weight is a multiplier indicating how heavily the individual should be considered in order for the analysis to be unconfounded by selection bias (Robins et al., 2000). Therefore, using analytic weights that combined design weights with propensity score weights should have produced nationally representative estimates that were free from overt selection bias. Descriptive statistics and these estimates of the effect of music participation on students' academic and social-emotional outcomes are detailed in the next chapter.

Chapter 5: Results

Propensity Score Estimation and Diagnostics

The purpose of this investigation was to compare the observed academic and social-emotional achievement of students who enrolled in high school music between 2009 and 2013 with those who did not enroll in school music. Because selection bias is a pervasive concern with this type of research, the analytic strategy involved comparing the scores of music students with a comparison group as close as possible to the music students in terms of their observed characteristics. Due to the documented differences in the academic and demographic profiles of students enrolled in various types of music courses, I constructed the comparison groups separately for general, choral, and instrumental music.

Rather than attempting to match on a small number of supposedly representative variables, I used propensity scores to incorporate information from each of the more than one hundred student-level variables and twenty-seven school-level variables listed in Table 4.2. A successful propensity score estimation results in covariate balance between the control and comparison groups; i.e., the distributions of the variables are the same across groups. Following an unsuccessful propensity score estimation, t tests of many covariate means between the control and treatment groups will show significant differences. Given the large sample sizes, number of variables involved, and the 5% error rate of typical hypothesis testing, some tests are almost certain to be significant even if the groups are balanced. Accordingly, while t tests between

groups can be informative, the absolute standardized mean difference d_x is a preferred metric for assessing covariate balance (Haviland et al., 2007; Rosenbaum, 2010).

I estimated the propensity scores using the R package “Toolkit for Weighting and Analysis of Nonequivalent Groups,” known as `twang` (Ridgeway et al., 2017). Figures in this chapter show three types of graphic diagnostic output from `twang`. Figures 5.1, through 5.12 show boxplots of the estimated propensity scores for the “music” (on the top) and “control” (on the bottom) groups. As detailed in Chapter 3, the desideratum is sufficient overlap in the estimated propensity scores between the control group of students who did not enroll in the particular type of music and the treatment group of students who did. However, given that the ATT (Average Treatment Effect on the Treated) is the causal estimand of interest, “weak overlap” (Imbens, 2004), defined as no cases with propensity scores equal to 1.0, is a sufficient standard.

Propensity Score Results for Choral Music

Figures 5.1 and 5.3 indicate that there is adequate overlap in the distributions of the propensity scores between the control and treatment groups for both “doses” of choral music, with no perfect prediction of treatment group membership. Furthermore, Figures 5.2 and 5.4 reveal that the absolute standardized mean difference d_x between the two groups was less than the desired caliper 0.2 for all of the variables in the propensity score equation, indicating that the estimation algorithm was successful at creating covariate balance between the groups. Examination of Figures 5.1 and 5.3 reveals that, when considering all available information, the profiles of choral music students are manifestly different than the students who did not enroll in choral music.

Figure 5.1

Propensity Score Overlap Plot for 1-2 Years of Choral Music

Note. Treatment group “1” is the non-music group, and treatment group “2” is the listed music group.

SOURCE: *Public Elementary/Secondary School Universe Survey, 2009–10*, Common Core of Data (CCD), Version Provisional 2a, U.S. Department of Education, National Center for Education Statistics; and U.S. Department of Education, National Center for Education Statistics, *High School Longitudinal Study of 2009 (HSL:09)*, “Base-Year to Second Follow-Up, 2018.”

Figure 5.2

Absolute Standardized Mean Difference (d_x) Plot for 1-2 Years of Choral Music

Note. Closed circles indicate statistically significant differences between groups; open circles indicate nonsignificant differences. The absolute standardized difference is the figure of most interest as it is not affected by sample size.

SOURCE: *Public Elementary/Secondary School Universe Survey, 2009–10*, Common Core of Data (CCD), Version Provisional 2a, U.S. Department of Education, National Center for Education Statistics; and U.S. Department of Education, National Center for Education Statistics, *High School Longitudinal Study of 2009 (HSL:09)*, “Base-Year to Second Follow-Up, 2018.”

Figure 5.3

Propensity Score Overlap Plot for 3-4 Years of Choral Music

SOURCE: *Public Elementary/Secondary School Universe Survey, 2009–10*, Common Core of Data (CCD), Version Provisional 2a, U.S. Department of Education, National Center for Education Statistics; and U.S. Department of Education, National Center for Education Statistics, *High School Longitudinal Study of 2009 (HSL:09)*, “Base-Year to Second Follow-Up, 2018.”

Figure 5.4

Absolute Standardized Mean Difference (d_x) Plot for 3-4 Years of Choral Music

SOURCE: *Public Elementary/Secondary School Universe Survey, 2009–10*, Common Core of Data (CCD), Version Provisional 2a, U.S. Department of Education, National Center for Education Statistics; and U.S. Department of Education, National Center for Education Statistics, *High School Longitudinal Study of 2009 (HSL:09)*, “Base-Year to Second Follow-Up, 2018.”

Propensity Score Results for Instrumental Music

As illustrated by Figures 5.5 and 5.7, there was an adequate region of common support for both 1-2 and 3-4 years of instrumental music. No cases had propensity scores equal to 1.0. d_x was also under 0.2 for all of the covariates in the matching equation (Figures 5.6 and 5.8), establishing acceptable covariate balance between both pairs of control and treatment groups. As

was the case with the choral music students, Figures 5.5 and 5.7 show that the propensity scores of the instrumental music students were markedly different than those of the non-music students. This indicates that other characteristics are different between ensemble music students and the general population, which may only be exacerbated by the fact that the control population in this study includes an unknown number of students that initially enrolled in music but ceased their participation prior to high school.

Figure 5.5

Propensity Score Overlap Plot for 1-2 Years of Instrumental Music

SOURCE: *Public Elementary/Secondary School Universe Survey, 2009–10*, Common Core of Data (CCD), Version Provisional 2a, U.S. Department of Education, National Center for Education Statistics; and U.S. Department of Education, National Center for Education Statistics, *High School Longitudinal Study of 2009 (HSL:09)*, “Base-Year to Second Follow-Up, 2018.”

Figure 5.6

Absolute Standardized Mean Difference (d_x) Plot for 1-2 Years of Instrumental Music

SOURCE: *Public Elementary/Secondary School Universe Survey, 2009–10*, Common Core of Data (CCD), Version Provisional 2a, U.S. Department of Education, National Center for Education Statistics; and U.S. Department of Education, National Center for Education Statistics, *High School Longitudinal Study of 2009 (HSL:09)*, “Base-Year to Second Follow-Up, 2018.”

Figure 5.7

Propensity Score Overlap Plot for 3-4 Years of Instrumental Music

SOURCE: *Public Elementary/Secondary School Universe Survey, 2009–10*, Common Core of Data (CCD), Version Provisional 2a, U.S. Department of Education, National Center for Education Statistics; and U.S. Department of Education, National Center for Education Statistics, *High School Longitudinal Study of 2009 (HSL:09)*, “Base-Year to Second Follow-Up, 2018.”

Figure 5.8

Absolute Standardized Mean Difference (d_x) Plot for 3-4 Years of Instrumental Music

SOURCE: *Public Elementary/Secondary School Universe Survey, 2009–10*, Common Core of Data (CCD), Version Provisional 2a, U.S. Department of Education, National Center for Education Statistics; and U.S. Department of Education, National Center for Education Statistics, *High School Longitudinal Study of 2009 (HSL:09)*, “Base-Year to Second Follow-Up, 2018.”

Propensity Score Results for General Music

While Figures 5.9 and 5.10 show that the overlap and covariate balance requirements were met for 1-2 years of general music, the algorithm was unsuccessful at creating a comparable control group for students who enrolled in general music for three or four years (Figures 5.11 and 5.12). As a result, the assumptions required for valid inferences resulting from

these propensity scores were not met. This was likely due to the small number of observations in the dataset who had sustained enrollment in general music and the considerable heterogeneity not only among those students but also among the classes falling under the general music umbrella. Consequently, I did not use students who enrolled in 3-4 years of general music in any analyses.

Figure 5.9

Propensity Score Overlap Plot for 1-2 Years of General Music

SOURCE: *Public Elementary/Secondary School Universe Survey, 2009–10*, Common Core of Data (CCD), Version Provisional 2a, U.S. Department of Education, National Center for Education Statistics; and U.S. Department of Education, National Center for Education Statistics, *High School Longitudinal Study of 2009 (HSL:09)*, “Base-Year to Second Follow-Up, 2018.”

Figure 5.10

Absolute Standardized Mean Difference (d_x) Plot for 1-2 Years of General Music

SOURCE: *Public Elementary/Secondary School Universe Survey, 2009–10*, Common Core of Data (CCD), Version Provisional 2a, U.S. Department of Education, National Center for Education Statistics; and U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.11

Propensity Score Overlap Plot for 3-4 Years of General Music

SOURCE: *Public Elementary/Secondary School Universe Survey, 2009–10*, Common Core of Data (CCD), Version Provisional 2a, U.S. Department of Education, National Center for Education Statistics; and U.S. Department of Education, National Center for Education Statistics, *High School Longitudinal Study of 2009 (HSL:09)*, “Base-Year to Second Follow-Up, 2018.”

Figure 5.12

Absolute Standardized Mean Difference (d_x) Plot for 3-4 Years of General Music

SOURCE: *Public Elementary/Secondary School Universe Survey, 2009–10*, Common Core of Data (CCD), Version Provisional 2a, U.S. Department of Education, National Center for Education Statistics; and U.S. Department of Education, National Center for Education Statistics, *High School Longitudinal Study of 2009 (HSL:09)*, “Base-Year to Second Follow-Up, 2018.”

Descriptive Statistics

While the reporting in this chapter and in Appendix E does include significance tests and p values, null hypothesis significance testing has well documented limitations (e.g., Kline, 2013; Trafimow & Marks, 2015; Wasserstein, 2016; Ziliak & McCloskey, 2008). To complement the

classical statistical tests, I also calculated effect sizes for each contrast using Hedges's g (Hedges, 1981), which is appropriate in cases such as this where the control and treatment groups have dissimilar sample sizes (Ellis, 2010). In contrast with the commonly used Cohen's d that assumes homogeneity of variances, Hedges's g uses a weighted pooled standard deviation to report the magnitude of group differences in standard deviation units. I wrote a Stata program, `getsvysd`, to obtain measures of the standard deviations for each outcome variable that took the complex sample design and multiple imputation into account, since the base `estat esize` command does not. I also wrote a Stata program, `getsvyg`, to calculate the adjusted values of Hedges's g . The syntax for these commands is in Appendix D. Even where p values were greater than .05, I reported effects where Hedges's g was larger than .15. I reported effects with p values greater than .05 and Hedges's g values less than .15 as "nonsignificant."

Sample sizes, rounded as prescribed by NCES, are in Table 5.1. A total of 3,290 cases in the analytic sample were coded as music students, representing a population of 755,360 American high school music students. Each music "treatment" group had a comparison group of several thousand sampled students, representing a total of over a million students in the population. As detailed in Chapter 4, while there was no overlap between the music "treatment" groups (i.e., students who enrolled in multiple types of music were excluded from the sample), the control group was essentially the same. The slight variation between control groups is the result of (1) discarding cases when students' schools did not verifiably offer the given type of curricular music, and (2) the multiple imputation being estimated separately for each of the five combinations of music and control students. I compared the regression predictors and outcomes between the music and control groups for each music type. The results were consistent with extant literature.

Table 5.1

Sample Sizes

	Treatment Group		Control Group		Total	
	<i>n</i>	Population Size	<i>n</i>	Population Size	<i>n</i>	Population Size
Choral, 1-2 Years	630	131,530	5,520	1,244,860	6,150	1,376,390
Choral, 3-4 Years	440	90,730	5,520	1,244,860	5,950	1,335,600
Instrumental, 1-2 Years	530	110,810	6,040	1,328,430	6,560	1,439,240
Instrumental, 3-4 Years	790	172,150	6,040	1,328,430	6,830	1,500,570
General, 1-2 Years	1,060	250,140	4,610	1,034,910	5,680	1,285,050

Note. Sample sizes are rounded in accordance with NCES reporting requirements.

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Regression Predictors

Choral Music

Recent research (Elpus & Abril, 2019; Kinney, 2018) has found that choral music students tend to be disproportionately white and female, and similar in socioeconomic status to the population. Descriptive statistics for the choral music students and the comparison group are in Tables 5.2 and 5.3. The population of choral music students largely reflects the observable demographic characteristics of the student population as a whole. Even with the large sample sizes involved, there were almost no significant differences between students who enrolled in 1-2 years of choral music and those who did not enroll in music. The only exception was the proportion of male students, which was higher in the control group than the choral music group ($g = -.47; p = .000$).

There were differences between the control group and students who enrolled in 3-4 years of choral music, however. White ($g = +.38$; $p = .000$) and female ($g = +.62$; $p = .000$) students enrolled in 3-4 years of choral music at significantly higher rates, and these students had significantly higher socioeconomic status than those in the control group ($g = +.27$; $p = .001$). There were no significant differences between either choral music group and the control group with respect to mathematics or science self-efficacy or students' likelihood of having received a grade of C, D, or F in their 8th grade mathematics or science courses, although their 9th grade mathematics reasoning scores were higher ($g = +.17$; $p = .013$). These results indicate that it may be a fallacy to refer to American choral music education students as a single population. Students with sustained enrollment in choral music were demographically different from students in the control group, even though most of their pretreatment academic and social-emotional scores were similar.

Table 5.2

Regression Predictors for 1-2 Years of Choral Music

Variable	Music		Control		Unadjusted Comparison				
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	Music vs. Control	<i>SE</i>	<i>t</i>	<i>p</i>	<i>g</i>
White Proportion	0.53	0.04	0.55	0.02	-0.02	0.04	-0.46	.647	-.03
Male Proportion	0.29	0.03	0.53	0.01	-0.23	0.04	-6.32	.000	-.47
Socioeconomic Status	-0.10	0.05	-0.05	0.02	-0.05	0.05	-1.08	.280	-.07
9th Grade Math Reasoning Score	-0.06	0.07	0.06	0.03	-0.12	0.07	-1.64	.102	-.13
9th Grade Math Self-Efficacy	0.02	0.07	0.06	0.02	-0.04	0.07	-0.60	.550	-.05
Low Grade in 8th Grade Math	0.22	0.03	0.23	0.01	-0.01	0.03	-0.32	.749	-.02
9th Grade Science Self-Efficacy	-0.08	0.08	-0.01	0.03	-0.07	0.08	-0.84	.405	-.07
Low Grade in 8th Grade Science	0.18	0.01	0.20	0.02	-0.03	0.03	-0.91	.364	-.06

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HLS:09), "Base-Year to Second Follow-Up, 2018."

Table 5.3

Regression Predictors for 3-4 Years of Choral Music

Variable	Music		Control		Unadjusted Comparison				
	Linearized		Linearized		Music vs. Control	SE	<i>t</i>	<i>p</i>	<i>g</i>
White Proportion	0.74	0.03	0.55	0.02	0.19	0.04	5.27	.000	+.38
Male Proportion	0.22	0.04	0.52	0.01	-0.30	0.04	-8.14	.000	-.62
Socioeconomic Status	0.14	0.05	-0.05	0.02	0.20	0.06	3.46	.001	+.27
9th Grade Math Reasoning Score	0.22	0.06	0.06	0.03	0.16	0.06	2.50	.013	+.17
9th Grade Math Self-Efficacy	-0.03	0.07	0.06	0.02	-0.09	0.08	-1.19	.237	-.10
Low Grade in 8th Grade Math	0.21	0.03	0.23	0.01	-0.02	0.03	-0.69	.493	-.06
9th Grade Science Self-Efficacy	-0.01	0.07	-0.01	0.03	0.00	0.08	-0.05	.960	+.00
Low Grade in 8th Grade Science	0.18	0.02	0.20	0.01	-0.03	0.03	-1.12	.265	-.07

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Instrumental Music

Descriptive statistics for the instrumental music and comparison samples are in Tables 5.4 and 5.5. In contrast to the choral music sample, students who enrolled in instrumental music for 1-2 years had numerous significant differences from the control group. They had higher 9th grade science self-efficacy ($g = +.15$; $p = .031$) and mathematics reasoning scores ($g = +.16$; $p = .023$), and a lower probability of receiving a grade of C, D, or F in their 8th grade science course ($g = -.22$; $p = .000$). Students who enrolled in 3-4 years of instrumental music were significantly different from the control group in every category except sex and 9th grade mathematics self-efficacy. These findings are in line with previous music education demography, which has shown that instrumental music students have higher average socioeconomic status than the general population, and also to be disproportionately white (Elpus & Abril, 2019; Kinney, 2018).

Table 5.4

Regression Predictors for 1-2 Years of Instrumental Music

Variable	Music		Control		Unadjusted Comparison				
	Linearized		Linearized		Music vs. Control	SE	<i>t</i>	<i>p</i>	<i>g</i>
White Proportion	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>					
White Proportion	0.53	0.04	0.56	0.02	-0.02	0.04	-0.56	.575	-.05
Male Proportion	0.50	0.03	0.52	0.01	-0.02	0.03	-0.58	.562	-.04
Socioeconomic Status	0.02	0.06	-0.03	0.02	0.06	0.06	0.92	.361	+.08
9th Grade Math Reasoning Score	0.20	0.07	0.05	0.03	0.15	0.07	2.29	.023	+.16
9th Grade Math Self-Efficacy	0.14	0.06	0.07	0.02	0.07	0.06	1.17	.244	+.07
Low Grade in 8th Grade Math	0.19	0.03	0.23	0.01	-0.04	0.03	-1.32	.188	-.09
9th Grade Science Self-Efficacy	0.15	0.06	0.00	0.03	0.15	0.07	2.18	.031	+.15
Low Grade in 8th Grade Science	0.10	0.02	0.19	0.01	-0.09	0.02	-3.86	.000	-.22

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), "Base-Year to Second Follow-Up, 2018."

Table 5.5

Regression Predictors for 3-4 Years of Instrumental Music

Variable	Music		Control		Unadjusted Comparison				
	Linearized		Linearized		Music vs. Control	SE	<i>t</i>	<i>p</i>	<i>g</i>
White Proportion	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>					
White Proportion	0.65	0.03	0.56	0.02	0.10	0.03	2.84	.005	+.19
Male Proportion	0.50	0.02	0.52	0.01	-0.03	0.03	-1.06	.289	-.05
Socioeconomic Status	0.22	0.05	-0.03	0.02	0.25	0.05	5.09	.000	+.35
9th Grade Math Reasoning Score	0.55	0.05	0.05	0.03	0.49	0.05	9.55	.000	+.52
9th Grade Math Self-Efficacy	0.18	0.06	0.07	0.02	0.11	0.07	1.68	.094	+.12
Low Grade in 8th Grade Math	0.11	0.02	0.23	0.01	-0.12	0.02	-5.80	.000	-.30
9th Grade Science Self-Efficacy	0.26	0.06	0.00	0.03	0.27	0.06	4.36	.000	+.28
Low Grade in 8th Grade Science	0.08	0.01	0.19	0.01	-0.11	0.02	-6.73	.000	-.28

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), "Base-Year to Second Follow-Up, 2018."

General Music

The general music group (Table 5.6) largely mirrored the control group in terms of observable characteristics. The only difference that even approached statistical significance was that the general music group had somewhat lower socioeconomic status ($g = -.14$; $p = .014$). Comparisons for all other variables were nonsignificant. National figures for general music demography are scarce in the literature.

Table 5.6

Regression Predictors for 1-2 Years of General Music

Variable	Music		Control		Unadjusted Comparison				
	Linearized		Linearized		Music vs. Control	SE	<i>t</i>	<i>p</i>	<i>g</i>
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>					
White Proportion	0.49	0.03	0.55	0.02	-0.06	0.04	-1.55	.123	-.12
Male Proportion	0.54	0.02	0.52	0.02	0.03	0.03	0.84	.400	+.05
Socioeconomic Status	-0.13	0.04	-0.03	0.02	-0.10	0.05	-2.13	.034	-.14
9th Grade Math Reasoning Score	0.01	0.07	0.03	0.03	-0.03	0.07	-0.39	.699	-.03
9th Grade Math Self-Efficacy	0.13	0.06	0.05	0.03	0.08	0.06	1.24	.218	+.08
Low Grade in 8th Grade Math	0.20	0.04	0.23	0.01	-0.03	0.04	-0.81	.420	-.07
9th Grade Science Self-Efficacy	0.07	0.06	-0.01	0.03	0.08	0.06	1.42	.156	+.09
Low Grade in 8th Grade Science	0.16	0.02	0.19	0.01	-0.04	0.03	-1.28	.203	-.09

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HLS:09), “Base-Year to Second Follow-Up, 2018.”

Sustained Involvement

I compared the predictors between the students who took 1-2 (“shorter enrollment”) and 3-4 years (“sustained enrollment”) of choral and instrumental music. These comparisons are displayed in Tables 5.7 and 5.8. While the sustained-enrollment choral music group had a proportion of males 7 percentage points lower than for the shorter-enrollment choral music group

($g = -.16$), the difference was not statistically significant ($p = .154$). The sustained-enrollment choral group did have significantly and substantially higher 9th grade mathematics reasoning scores, percentage of white students, and socioeconomic status. These results imply that treating high school choral music students as a single entity may be obscuring important differences between those who participated for different durations.

Table 5.7

Choral Music Predictor Comparison

Variable	1-2 Years		3-4 Years		Unadjusted Comparison				
	<i>M</i>	Linearized <i>SE</i>	<i>M</i>	Linearized <i>SE</i>	3-4 Years vs. 1-2 Years	<i>SE</i>	<i>t</i>	<i>p</i>	<i>g</i>
White Proportion	0.53	0.04	0.74	0.03	0.20	0.05	4.51	.000	+.43
Male Proportion	0.29	0.03	0.22	0.04	-0.07	0.05	-1.43	.154	-.16
Socioeconomic Status	-0.10	0.05	0.14	0.05	0.24	0.07	3.38	.001	+.34
9th Grade Math Reasoning Score	-0.06	0.07	0.22	0.06	0.28	0.09	3.23	.002	+.31
9th Grade Math Self-Efficacy	0.02	0.07	-0.03	0.07	-0.05	0.10	-0.47	.638	-.05
Low Grade in 8th Grade Math	0.22	0.03	0.21	0.03	-0.01	0.05	-0.29	.776	-.03
9th Grade Science Self-Efficacy	-0.08	0.08	-0.01	0.07	0.07	0.11	0.63	.533	+.07
Low Grade in 8th Grade Science	0.18	0.02	0.18	0.02	0.00	0.03	-0.09	.932	-.01

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

There were several significant differences between the instrumental music groups. Students who had sustained instrumental music enrollment had a lower incidence of low grades in 8th grade mathematics ($g = -.24$; $p = .013$) and higher 9th grade mathematics reasoning scores ($g = +.37$; $p = .000$) relative to those with shorter enrollment. Students in the sustained enrollment group were also more likely to be white ($g = +.24$; $p = .006$), and had higher

socioeconomic status on average ($g = +.25$; $p = .002$). Similar to the choral music students, treating instrumental music students as a monolith in future research may be inadvisable.

Table 5.8

Instrumental Music Predictor Comparison

Variable	1-2 Years		3-4 Years		Unadjusted Comparison				
	<i>M</i>	Linearized <i>SE</i>	<i>M</i>	Linearized <i>SE</i>	3-4 Years vs. 1-2 Years	<i>SE</i>	<i>t</i>	<i>p</i>	<i>g</i>
White Proportion	0.53	0.04	0.65	0.03	0.12	0.04	2.76	.006	+.24
Male Proportion	0.50	0.03	0.50	0.02	-0.01	0.04	-0.21	.830	-.02
Socioeconomic Status	0.02	0.06	0.22	0.05	0.19	0.06	3.15	.002	+.25
9th Grade Math Reasoning Score	0.20	0.07	0.55	0.05	0.34	0.08	4.37	.000	+.37
9th Grade Math Self-Efficacy	0.14	0.06	0.18	0.06	0.04	0.08	0.52	.601	+.04
Low Grade in 8th Grade Math	0.19	0.03	0.11	0.02	-0.08	0.03	-2.50	.013	-.24
9th Grade Science Self-Efficacy	0.15	0.06	0.26	0.06	0.12	0.08	1.49	.140	+.13
Low Grade in 8th Grade Science	0.10	0.02	0.08	0.01	-0.02	0.03	-0.79	.432	-.07

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Comparisons Between Groups

Figures 5.13 through 5.20 display the means and 95% confidence intervals for the regression predictors across all of the music groups and the instrumental music control group, which was chosen because it was the largest. General music students were broadly similar to the control group. The general music sample had a somewhat lower socioeconomic status than the control group, but marginally better eighth grade achievement. General music had the lowest proportion of white students out of all of the categories. General music students also had slightly higher self-efficacy scores than the control group.

Both of the short-term ensemble music groups were near the control group mean in terms of race. The sustained enrollment choral and instrumental groups had noticeably higher proportions of white students, with the students enrolled in 3-4 years of choral music exhibiting the highest proportion of white students among all of the music groups. Unsurprisingly, the choral music groups had the lowest proportion of male students. Socioeconomic status, 9th grade mathematics reasoning, and 8th grade achievement generally followed the same pattern. Students enrolled in 3-4 years of choral and instrumental music had higher achievement and socioeconomic status than their counterparts who enrolled for 1-2 years. Students enrolled in 3-4 years of choir had a roughly similar profile to students enrolled in 1-2 years of instrumental music, and those enrolled in 3-4 years of instrumental music were the highest achievers and had the highest socioeconomic status. Self-efficacy scores for choral music students were slightly lower than the control group, while those for instrumental music students were slightly higher.

Figure 5.13

Sample Comparison: Proportion of Students Identified as White

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), "Base-Year to Second Follow-Up, 2018."

Figure 5.14

Sample Comparison: Proportion of Students Identified as Male

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), "Base-Year to Second Follow-Up, 2018."

Figure 5.15

Sample Comparison: Mean Socioeconomic Status

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.16

Sample Comparison: Mean 9th Grade Mathematics Reasoning Score

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.17

Sample Comparison: Mean 9th Grade Mathematics Self-Efficacy Score

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.18

Sample Comparison: Proportion Receiving Low Grade in 8th Grade Mathematics

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.19

Sample Comparison: Mean 9th Grade Science Self-Efficacy Score

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.20

Sample Comparison: Proportion Receiving Low Grade in 8th Grade Science

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Outcomes

Choral Music

Descriptive statistics for the outcomes among the choral music students are shown in Tables 5.9 and 5.10. There were few significant differences in the outcomes of interest between the students who enrolled in 1-2 years of choral music and the control group, but the choral music students had lower performance in mathematics. These students had lower college examination mathematics scores ($g = -.16$; $p = .060$) and 11th grade mathematics reasoning scores ($g = -.15$; $p = .034$). Students who took 3-4 years of choral music had significantly higher college entrance examination reading scores ($g = +.21$; $p = .037$) and 11th grade mathematics reasoning scores ($g = +.16$; $p = .048$) than their counterparts who did not enroll in music. These findings accord with prior research (Kinney, 2018). Social-emotional variables were comparable between the choral music and control groups.

Table 5.9

Outcomes for 1-2 Years of Choral Music

Variable	Music		Control		Unadjusted Comparison				
	Linearized		Linearized		Music vs. Control	SE	<i>t</i>	<i>p</i>	<i>g</i>
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>					
College Exam Math Score	479.69	8.14	496.18	3.85	-16.50	8.72	-1.89	.060	-.16
College Exam Reading Score	481.61	7.56	486.18	4.18	-4.57	8.69	-0.53	.599	-.05
Mathematics Reasoning	0.60	0.08	0.76	0.03	-0.17	0.08	-2.13	.034	-.15
School Motivation	0.13	0.06	0.08	0.02	0.05	0.06	0.77	.442	+.05
Mathematics Self-Efficacy	0.02	0.08	0.02	0.02	0.00	0.09	0.05	.963	+.00
Science Self-Efficacy	0.04	0.08	-0.03	0.02	0.07	0.08	0.86	.391	+.07

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HLS:09), "Base-Year to Second Follow-Up, 2018."

Table 5.10

Outcomes for 3-4 Years of Choral Music

Variable	Music		Control		Unadjusted Comparison				
	<i>M</i>	Linearized <i>SE</i>	<i>M</i>	Linearized <i>SE</i>	Music vs. Control	<i>SE</i>	<i>t</i>	<i>p</i>	<i>g</i>
College Exam Math Score	500.84	9.63	496.18	3.85	4.66	9.99	0.47	.641	+.05
College Exam Reading Score	507.28	9.75	486.18	4.18	21.10	10.07	2.10	.037	+.21
Mathematics Reasoning	0.93	0.08	0.76	0.03	0.17	0.09	1.99	.048	+.16
School Motivation	0.20	0.07	0.08	0.02	0.12	0.07	1.62	.106	+.13
Mathematics Self-Efficacy	-0.11	0.08	0.02	0.02	-0.13	0.08	-1.69	.093	-.13
Science Self-Efficacy	-0.10	0.06	-0.03	0.02	-0.07	0.06	-1.06	.292	-.07

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Instrumental Music

Tables 5.11 and 5.12 contain descriptive statistics for the instrumental music students’ outcomes of interest. Instrumental music students who enrolled for 1-2 years had higher college entrance examination scores in reading ($g = +.36$; $p = .000$) and mathematics ($g = +.22$; $p = .056$) than the control group. They also had higher 11th grade mathematics reasoning scores ($g = +.17$; $p = .024$). The instrumental music education students who enrolled for 3-4 years were significantly different from the non-music group in almost every respect: both college entrance examination mathematics ($g = +.38$; $p = .000$) and reading ($g = +.48$; $p = .000$) scores, mathematics reasoning ($g = +.56$; $p = .000$), school motivation ($g = +.20$; $p = .000$) and science self-efficacy ($g = +.25$; $p = .000$). Mathematics self-efficacy was the only outcome for which there was not a statistically significant difference ($g = +07$.; $p = .322$). Prior research has found instrumental music students to have high test scores and to receive high grades in school (Elpus & Abril, 2019; Kinney, 2018).

Table 5.11

Outcomes for 1-2 Years of Instrumental Music

Variable	Music		Control		Unadjusted Comparison				
	Linearized		Linearized		Music vs. Control	SE	t	p	g
College Exam Math Score	519.89	11.50	497.77	3.92	22.12	11.52	1.92	.056	+.22
College Exam Reading Score	522.27	9.87	485.33	4.12	36.94	10.14	3.64	.000	+.36
Mathematics Reasoning	0.94	0.08	0.75	0.03	0.19	0.08	2.27	.024	+.17
School Motivation	0.05	0.07	0.08	0.02	-0.03	0.07	-0.45	.656	-.04
Mathematics Self-Efficacy	0.11	0.06	0.02	0.02	0.09	0.07	1.28	.201	+.09
Science Self-Efficacy	0.12	0.07	-0.02	0.02	0.14	0.08	1.82	.070	+.14

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HLS:09), “Base-Year to Second Follow-Up, 2018.”

Table 5.12

Outcomes for 3-4 Years of Instrumental Music

Variable	Music		Control		Unadjusted Comparison				
	Linearized		Linearized		Music vs. Control	SE	t	p	g
College Exam Math Score	535.78	7.76	497.77	3.92	38.01	8.37	4.54	.000	+.38
College Exam Reading Score	534.35	8.73	485.33	4.12	49.03	9.52	5.15	.000	+.48
Mathematics Reasoning	1.35	0.06	0.75	0.03	0.60	0.06	10.55	.000	+.56
School Motivation	0.26	0.04	0.08	0.02	0.17	0.04	4.09	.000	+.20
Mathematics Self-Efficacy	0.09	0.06	0.02	0.02	0.07	0.07	0.99	.322	+.07
Science Self-Efficacy	0.23	0.06	-0.02	0.02	0.25	0.06	4.06	.000	+.25

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HLS:09), “Base-Year to Second Follow-Up, 2018.”

General Music

As Table 5.13 indicates, there were no significant differences between the general music students’ outcomes and the non-music group. The students who enrolled in 1-2 years of general

music in this study were not appreciably different from the broader student population in terms of the studied observable characteristics.

Table 5.13

Outcomes for 1-2 Years of General Music

Variable	Music		Control		Unadjusted Comparison				
	Linearized		Linearized		Music vs. Control	SE	<i>t</i>	<i>p</i>	<i>g</i>
College Exam Math Score	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>					
College Exam Math Score	494.37	7.51	495.20	3.93	-0.83	7.94	-0.10	.917	-.01
College Exam Reading Score	483.08	5.87	484.47	3.71	-1.39	6.76	-0.21	.837	-.01
Mathematics Reasoning	0.72	0.07	0.71	0.04	0.00	0.07	0.05	.964	+.00
School Motivation	0.07	0.06	0.08	0.02	-0.02	0.07	-0.22	.828	-.02
Mathematics Self-Efficacy	0.03	0.05	0.01	0.03	0.02	0.06	0.31	.755	+.02
Science Self-Efficacy	0.07	0.06	-0.03	0.03	0.10	0.07	1.41	.159	+.10

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Sustained Involvement

An examination of the differences between shorter and sustained enrollment in choral music (Table 5.14) revealed that the sustained-enrollment group had higher college entrance examination scores in reading ($g = +.27$; $p = .032$) and mathematics ($g = +.21$; $p = .074$). The sustained-enrollment group also had higher scores in mathematics reasoning ($g = +.30$; $p = .003$). No large or significant differences were observed with respect to the social-emotional outcomes. Table 5.15 indicates that sustained-enrollment instrumental music students had higher school motivation ($g = +.25$; $p = .007$), mathematics reasoning ($g = +.30$; $p = .003$), and college examination mathematics scores ($g = +.15$; $p = .255$) than those who only enrolled for one or two years. No other differences met the reporting threshold.

Table 5.14

Choral Music Outcome Comparison

Variable	1-2 Years		3-4 Years		Unadjusted Comparison				
	<i>M</i>	Linearized <i>SE</i>	<i>M</i>	Linearized <i>SE</i>	3-4 Years vs. 1-2 Years	<i>SE</i>	<i>t</i>	<i>p</i>	<i>g</i>
College Exam Math Score	479.69	8.14	500.84	9.63	21.16	11.69	1.81	.074	+.21
College Exam Reading Score	481.61	7.56	507.28	9.75	25.67	11.75	2.19	.032	+.27
Mathematics Reasoning	0.60	0.08	0.93	0.08	0.34	0.11	3.06	.003	+.30
School Motivation	0.13	0.06	0.20	0.07	0.07	0.09	0.78	.436	+.08
Mathematics Self-Efficacy	0.02	0.08	-0.11	0.08	-0.14	0.11	-1.24	.216	-.13
Science Self-Efficacy	0.04	0.08	-0.10	0.06	-0.13	0.08	-1.59	.115	-.13

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Table 5.15

Instrumental Music Outcome Comparison

Variable	1-2 Years		3-4 Years		Unadjusted Comparison				
	<i>M</i>	Linearized <i>SE</i>	<i>M</i>	Linearized <i>SE</i>	3-4 Years vs. 1-2 Years	<i>SE</i>	<i>t</i>	<i>p</i>	<i>g</i>
College Exam Math Score	519.89	11.50	535.78	7.82	15.89	13.88	1.14	.255	+.15
College Exam Reading Score	522.27	9.87	534.36	8.73	12.09	12.85	0.94	.349	+.11
Mathematics Reasoning	0.94	0.08	1.35	0.06	0.41	0.10	4.08	.000	+.38
School Motivation	0.05	0.07	0.26	0.04	0.21	0.08	2.73	.007	+.25
Mathematics Self-Efficacy	0.11	0.06	0.09	0.06	-0.02	0.07	-0.25	.801	-.02
Science Self-Efficacy	0.12	0.07	0.23	0.06	0.11	0.08	1.48	.140	+.12

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Comparisons Between Groups

Between-group comparisons for the outcome variables mirrored the trends for the predictors. The general music group was nearly identical to the control group on most measures. The lone exception was the slightly higher 11th grade science self-efficacy score, which is

unsurprising since general music students' 9th grade science self-efficacy was also higher than control. Both college entrance examination scores increased from 1-2 years of choral music, to 3-4 years of choral music, to 1-2 years of instrumental music, to 3-4 years of instrumental music, which were the highest. Both self-efficacy scores were slightly lower among students enrolled in 3-4 years of choral music than those enrolled in 1-2 years. While instrumental music students' mathematics self-efficacy scores were not appreciably different, instrumental music students who enrolled for 3-4 years had higher science self-efficacy scores. Students enrolled in 3-4 years of instrumental music had the highest science self-efficacy scores out of any of the groups. School motivation was higher among both sustained enrollment groups than the short-term enrollment groups. Again, the students who had sustained enrollment in instrumental music had the highest scores for school motivation.

Figure 5.21

Sample Comparison: Unadjusted College Examination Mathematics Score

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), "Base-Year to Second Follow-Up, 2018."

Figure 5.22

Sample Comparison: Unadjusted College Examination Reading Score

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.23

Sample Comparison: Unadjusted 11th Grade Mathematics Reasoning Score

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.24

Sample Comparison: Unadjusted 11th Grade School Motivation Score

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HLS:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.25

Sample Comparison: Unadjusted 11th Grade Mathematics Self-Efficacy Score

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HLS:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.26

Sample Comparison: Unadjusted 11th Grade Science Self-Efficacy Score

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Social-Emotional Variables Measured Only in Ninth Grade

I compared the base year scale scores for school belongingness and school engagement between the music groups and the control groups. It is important to note that these associations are only associational, and no causal relationships can be established. Since these variables were measured at the beginning of 9th grade only, the high school music “treatments” cannot have been said to have affected them, and thus the propensity score approach was not applicable. It is likely that previous music enrollments (i.e., in middle school) are highly conflated with 9th grade “baseline” scores on these metrics among the instrumental music students, but potentially less conflated among the choral music students, and hardly or not conflated among the general music students.

Choral Music

The group of students who enrolled in 3-4 years of choral music had higher 9th grade school belongingness than the control group ($g = +.25$; $p = .000$), but not the students who enrolled for 1-2 years ($g = +.01$; $p = .865$). Choral music students who enrolled for 1-2 and 3-4 years had a sense of school engagement that was $g = +.13$ standard deviations higher. However, while this was statistically significant for students who enrolled for 1-2 years ($p = .031$), it was nonsignificant for students enrolling for 3-4 years ($p = .076$).

Table 5.16

Belongingness and Engagement for 1-2 Years of Choral Music

Variable	Music		Control		Unadjusted Comparison				
	Linearized		Linearized		Music vs. Control	SE	<i>t</i>	<i>p</i>	<i>g</i>
Sense of School Belonging	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>					
	0.01	0.07	-0.01	0.03	0.01	0.08	0.17	.865	+.01
Sense of School Engagement	0.18	0.05	0.05	0.03	-0.13	0.06	2.17	.031	+.13

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), "Base-Year to Second Follow-Up, 2018."

Table 5.17

Belongingness and Engagement for 3-4 Years of Choral Music

Variable	Music		Control		Unadjusted Comparison				
	Linearized		Linearized		Music vs. Control	SE	<i>t</i>	<i>p</i>	<i>g</i>
Sense of School Belonging	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>					
	0.23	0.06	-0.01	0.03	0.24	0.06	3.94	.000	+.25
Sense of School Engagement	0.18	0.01	0.05	0.03	0.12	0.07	1.79	.076	+.13

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), "Base-Year to Second Follow-Up, 2018."

Instrumental Music

Instrumental music students scored higher on school engagement among both students who enrolled for 1-2 years ($g = +.17$; $p = .016$) and 3-4 years ($g = +.18$; $p = .001$). While the instrumental music students who participated for 3-4 years had a higher belongingness score ($g = +.28$; $p = .000$), the students who enrolled for only 1-2 years did not ($g = +.04$; $p = .505$).

Table 5.18

Belongingness and Engagement for 1-2 Years of Instrumental Music

Variable	Music		Control		Unadjusted Comparison				
	Linearized		Linearized		Music vs. Control	SE	<i>t</i>	<i>p</i>	<i>g</i>
Sense of School Belonging	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>					
	0.04	0.05	0.00	0.03	0.04	0.06	0.67	.505	+.04
Sense of School Engagement	0.23	0.07	0.07	0.03	0.16	0.07	2.42	.016	+.17

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table 5.19

Belongingness and Engagement for 3-4 Years of Instrumental Music

Variable	Music		Control		Unadjusted Comparison				
	Linearized		Linearized		Music vs. Control	SE	<i>t</i>	<i>p</i>	<i>g</i>
Sense of School Belonging	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>					
	0.27	0.05	0.00	0.03	0.26	0.05	5.09	.000	+.28
Sense of School Engagement	0.24	0.05	0.07	0.02	0.17	0.05	3.26	.001	+.18

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

General Music

While students who enrolled in 1-2 years of general music did not have higher belongingness scores, they did have barely higher scores for school engagement ($g = +.09$; $p = .022$).

Table 5.20

Belongingness and Engagement for 1-2 Years of General Music

Variable	Music		Control		Unadjusted Comparison				
	Linearized		Linearized		Music vs. Control	SE	t	p	g
Sense of School Belonging	0.06	0.07	0.00	0.03	0.06	0.08	0.71	.478	+.06
Sense of School Engagement	0.16	0.05	0.07	0.03	0.09	0.03	2.32	.022	+.09

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), "Base-Year to Second Follow-Up, 2018."

Sustained Enrollment

There were no differences with respect to school engagement between students who enrolled in 1-2 and 3-4 years of either choral or instrumental music. Students who enrolled for 3-4 years had higher school belongingness scores than peers who enrolled for 1-2 years, both for choral music ($g = +.23$; $p = .022$) and instrumental music ($g = +.26$; $p = .002$).

Table 5.21

Choral Music Belongingness and Engagement Comparison

Variable	1-2 Years		3-4 Years		Unadjusted Comparison				
	Linearized		Linearized		3-4 Years vs. 1-2 Years	SE	t	p	g
Sense of School Belonging	0.01	0.07	0.23	0.06	0.22	0.10	2.31	.022	+.23
Sense of School Engagement	0.18	0.05	0.18	0.07	0.00	0.08	-0.05	.962	.00

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), "Base-Year to Second Follow-Up, 2018."

Table 5.22

Instrumental Music Belongingness and Engagement Comparison

Variable	1-2 Years		3-4 Years		Unadjusted Comparison				
	<i>M</i>	Linearized <i>SE</i>	<i>M</i>	Linearized <i>SE</i>	3-4 Years vs. 1-2 Years	<i>SE</i>	<i>t</i>	<i>p</i>	<i>g</i>
Sense of School Belonging	0.04	0.05	0.27	0.05	0.23	0.07	3.17	.002	+.26
Sense of School Engagement	0.23	0.07	0.24	0.05	0.01	0.08	0.11	.912	+.01

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Comparisons Between Groups

Students who enrolled for 1-2 years of any type of music had similar school belongingness to the control group (Figure 5.27). However, the choral and instrumental music students who enrolled for 3-4 years had higher scores on the belongingness metric. As shown in Figure 5.28, all of the music groups had slightly elevated school engagement relative to the control group, with the instrumental music students having the highest observed levels of school engagement.

Figure 5.27

Sample Comparison: Unadjusted 9th Grade School Belongingness Score

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.28

Sample Comparison: Unadjusted 9th Grade School Engagement Score

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Regression Results for Academic Outcomes

As detailed in Chapter 4, the heart of this study was a series of nested regression analyses. Each model was estimated twice, once unadjusted and once adjusted for observable selection bias using inverse probability of treatment selection weights (IPTW) derived from the propensity scores and trimmed at the 99th percentile. Both the adjusted and unadjusted models accounted for the HSLs:09's complex sampling design using Stata's `mi estimate: svy: regress` commands. The complete regression tables are in Appendix E.

Since the indicator variable for treatment condition was used in interactions in these models, reporting or testing its coefficients and significance alone would defeat the purpose of a complex regression model and ultimately be uninformative. In addition, the main effects of predictors not related to treatment condition (e.g., socioeconomic status) did not answer this study's research questions. Therefore, the coefficients of central interest in this investigation resulted from postestimation comparisons of conditional predicted values that I obtained using Stata's `pwcompare` command. Analogous to post hoc contrasts in analysis of variance, these postestimation comparisons were *t* tests between the predicted values of the outcome variable between students in the music and propensity-score-weighted non-music groups, after accounting for the effects of the other predictors including interactions and polynomials. The Stata add-on `mimrgns` (Klein, 2016) facilitated computing and graphing these predicted values.

For each of the five music groups (1-2 years of choral music, 3-4 years of choral music, 1-2 years of instrumental music, 3-4 years of instrumental music, and 1-2 years of general music), and for each of the six outcomes (college entrance examination mathematics and reading scores, mathematics reasoning ability, school motivation, mathematics self-efficacy, and science self-efficacy), I reported an effect if the propensity-score-adjusted conditional difference

between the music and non-music groups met the previously described significance thresholds. The first subpopulation effect I tested was between those students who did and did not earn a C, D, or F in the relevant 8th grade course (mathematics or science). Further potential subpopulation effects were those between students identified as male and female, and students identified as non-Hispanic white and students identifying as any of the additional races and ethnicities.

Choral Music

Table 5.23 and 5.24 contain selected results from the predicted conditional mean comparisons for academic outcomes.

Main Effects

Most of the comparisons between students who participated in choral music and students who did not participate in music were nonsignificant. There was an unadjusted reading score main effect of $g = +.23$ for students who enrolled in choral music for 3-4 years, but this was rendered nonsignificant by propensity score weighting.

Table 5.23

Academic Conditional Mean Comparisons for 1-2 Years of Choral Music

Dependent Variable	Unadjusted				PS Adjusted			
	Main Effect	8 th Grade	Sex	Race	Main Effect	8 th Grade Math	Sex	Race
College Exam Math Score	n.s.	C, D, or F: $g = +.36^{**}$	n.s.	n.s.	n.s.	C, D, or F: $g = +.37^*$	n.s.	n.s.
College Exam Reading Score	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Mathematics Reasoning	n.s.	n.s.	n.s.	Students of Color: $g = -.15^\dagger$	n.s.	n.s.	n.s.	Students of Color: $g = -.15^*$

$^\dagger p \geq .05$; $^* p < .05$; $^{**} p < .01$; $^{***} p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Table 5.24

Academic Conditional Mean Comparisons for 3-4 Years of Choral Music

Dependent Variable	Unadjusted				PS Adjusted			
	Main Effect	8 th Grade	Sex	Race	Main Effect	8 th Grade	Sex	Race
College Exam Math Score	n.s.	C, D, or F: $g = +.15^\dagger$	Male: $g = +.18^\dagger$	n.s.	n.s.	n.s.	Male: $g = +.21^\dagger$	n.s.
College Exam Reading Score	$g = +.23^\dagger$	A or B: $g = +.22^\dagger$ C, D, or F: $g = +.26^\dagger$	Female: $g = +.18^\dagger$ Male: $g = +.32^\dagger$	Students of Color: $g = +.27^\dagger$ White: $g = +.21^\dagger$	n.s.	n.s.	Male: $g = +.31^\dagger$	Students of Color: $g = +.19^\dagger$
Mathematics Reasoning	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	Male: $g = +.18^\dagger$	n.s.

$^\dagger p \geq .05$; * $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Achievement in Eighth Grade

Students who enrolled in 1-2 years of choral music and who received a grade of C, D, or F in their 8th grade mathematics courses had significantly higher scores on the mathematics portion of their college entrance examinations even after accounting for the other predictors (Figure 5.29). The unadjusted effect size was $g = +.36$, which was nearly identical to the result of $g = +.37$ that accounted for observable group differences ($p = .006$). However, this effect was nonsignificant among students who enrolled in choir for 3-4 years when incorporating the propensity scores (Figure 5.30).

Figure 5.29

Predicted College Entrance Examination Mathematics Scores by 8th Grade Mathematics

Achievement for 1-2 Years of Choral Music

Note. Students who took the ACT had their scores converted to SAT units.

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.30

Predicted College Entrance Examination Mathematics Scores by 8th Grade Mathematics

Achievement for 3-4 Years of Choral Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Sex

In all three academic areas, male students who participated in 3-4 years of curricular choral music had higher scores than otherwise-comparable male students in the control group (Figures 5.31 through 5.33). While all of the effect sizes were noticeable, none were statistically significant. No adjusted differences met the effect size or statistical significance thresholds for male students enrolled in only one or two years of choir.

Figure 5.31

Predicted College Entrance Examination Mathematics Scores by Sex for 3-4 Years of Choral Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.32

Predicted College Entrance Examination Reading Scores by Sex for 3-4 Years of Choral Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.33

Predicted Mathematics Reasoning Scores by Sex for 3-4 Years of Choral Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Race

Students of color had slightly but significantly lower *ceteris paribus* mathematics reasoning scores when they enrolled for 1-2 years of choral music ($p = .025$), but not for 3-4 years ($p = .875$). Students of color who participated in choir for 3-4 years had slightly higher reading scores at $g = +.19$, although this effect was statistically nonsignificant at $p = .291$. These results are illustrated in Figures 5.34, 5.35, and 5.36.

Figure 5.34

Predicted Mathematics Reasoning Scores by Race for 1-2 Years of Choral Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HLS:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.35

Predicted Mathematics Reasoning Scores by Race for 3-4 Years of Choral Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.36

Predicted College Entrance Examination Reading Scores by Race for 3-4 Years of Choral Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Instrumental Music

Notable results of the academic outcome regression models for instrumental music students are in Tables 5.25 and 5.26.

Table 5.25

Academic Conditional Mean Comparisons for 1-2 Years of Instrumental Music

Dependent Variable	Unadjusted				PS Adjusted			
	Main Effect	8 th Grade	Sex	Race	Main Effect	8 th Grade	Sex	Race
College Exam Math Score	n.s.	C, D, or F: $g = +.44^\dagger$	n.s.	White: $g = +.15^\dagger$	n.s.	C, D, or F: $g = +.39^\dagger$	Female: $g = +.18^*$	n.s.
College Exam Reading Score	$g = +.30^{**}$	A or B: $g = +.30^{**}$ C, D, or F: $g = +.33^*$	Female: $g = +.38^{**}$ Male: $g = +.21^\dagger$	Students of Color: $g = +.30^{**}$ White: $g = +.30^{**}$	$g = +.23^{**}$	A or B: $g = +.26^{**}$ C, D, or F: $g = +.23^*$	Female: $g = +.32^{**}$	Students of Color: $g = +.23^*$ White: $g = +.24^{**}$
Mathematics Reasoning	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.

$^\dagger p \geq .05$; $^* p < .05$; $^{**} p < .01$; $^{***} p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), "Base-Year to Second Follow-Up, 2018."

Table 5.26

Academic Conditional Mean Comparisons for 3-4 Years of Instrumental Music

Dependent Variable	Unadjusted				PS Adjusted			
	Main Effect	8 th Grade	Sex	Race	Main Effect	8 th Grade	Sex	Race
College Exam Math Score	n.s.	C, D, or F: $g = -.15^\dagger$	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
College Exam Reading Score	$g = +.32^{***}$	A or B: $g = +.34^{***}$ C, D, or F: $g = +.23^*$	Female: $g = +.40^{***}$ Male: $g = +.24^*$	Students of Color: $g = +.31^{**}$ White: $g = +.33^{***}$	$g = +.21^{**}$	A or B: $g = +.22^{**}$	Female: $g = +.27^{**}$ Male: $g = +.15^\dagger$	Students of Color: $g = +.22^{**}$ White: $g = +.21^{**}$
Mathematics Reasoning	$g = +.18^{***}$	A or B: $g = +.16^{***}$ C, D, or F: $g = +.23^{**}$	Male: $g = +.28^{***}$	Students of Color: $g = +.19^{***}$ White: $g = +.16^{***}$	$g = +.13^{***}$	A or B: $g = +.12^{**}$ C, D, or F: $g = +.20^*$	Male: $g = +.25^{***}$	Students of Color: $g = +.13^{***}$ White: $g = +.12^{**}$

$^\dagger p \geq .05$; $^* p < .05$; $^{**} p < .01$; $^{***} p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), "Base-Year to Second Follow-Up, 2018."

Main Effects

Students participating in instrumental music had noticeably higher reading examination scores (Figures 5.37 and 5.38), regardless of the duration of their enrollment or other attributes. The propensity-score-adjusted main effect for 1-2 years of instrumental music participation was $g = +.23$ ($p = .004$) and for 3-4 years of participation it was $g = +.21$ ($p = .004$). There was also a significant adjusted main effect for mathematics reasoning among students who participated in instrumental music for 3-4 years ($g = +.18$; $p = .000$), but this effect was zero among the students who enrolled for only 1-2 years ($g = .00$; $p = .914$). Figures 5.39 and 5.40 display these results. The main effects for college examination mathematics scores, however, were nonsignificant.

Figure 5.37

Predicted College Entrance Examination Reading Scores for 1-2 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.38

Predicted College Entrance Examination Reading Scores for 3-4 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.39

Predicted Mathematics Reasoning Scores for 1-2 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.40

Predicted Mathematics Reasoning Scores for 3-4 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Achievement in Eighth Grade

There was not a discernible pattern related to students' achievement in eighth grade. Students who participated in 1-2 years of instrumental music and who earned low grades in their 8th grade mathematics courses scored substantially higher than comparable peers on the mathematics portion of their college entrance examinations ($g = +.39$; $p = .099$). This did not hold true for students who enrolled for 3-4 years, however ($g = -.13$; $p = .597$). Graphs of these results are in Figures 5.41 and 5.42. The other academic outcomes were not appreciably different according to eighth grade achievement.

Figure 5.41

Predicted College Entrance Examination Mathematics Scores by 8th Grade Mathematics Achievement for 1-2 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HLS:09), "Base-Year to Second Follow-Up, 2018."

Figure 5.42

Predicted College Entrance Examination Reading Scores by 8th Grade Mathematics

Achievement for 3-4 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Sex

After controlling for overt selection bias, students identified as female were predicted to have higher scores on both the mathematics ($g = +.18$; $p = .028$) and reading ($g = +.32$; $p = .004$) portions of their college examinations when they were enrolled in instrumental music for 1-2 years, as illustrated in Figure 5.43 and 5.45. This trend continued among female students who participated for 3-4 years for reading scores ($g = +.27$; $p = .004$) but the mathematics scores for these students were not appreciably higher (Figures 5.44 and 5.46). Male students who had sustained enrollment had higher ceteris paribus scores on the mathematics reasoning exam ($g = +.25$; $p = .000$), but not those with shorter-term enrollments ($g = -.08$; $p = .257$). Figures 5.47 and 5.48 display these results graphically.

Figure 5.43

Predicted College Entrance Examination Mathematics Scores by Sex for 1-2 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.44

Predicted College Entrance Examination Mathematics Scores by Sex for 3-4 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.45

Predicted College Entrance Examination Reading Scores by Sex for 1-2 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), "Base-Year to Second Follow-Up, 2018."

Figure 5.46

Predicted College Entrance Examination Reading Scores by Sex for 3-4 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), "Base-Year to Second Follow-Up, 2018."

Figure 5.47

Predicted 11th Grade Mathematics Reasoning Scores by Sex for 1-2 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.48

Predicted 11th Grade Mathematics Reasoning Scores by Sex for 3-4 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Race

Students' academic outcomes did not differ significantly by race. Both racial groups contributed approximately equally to the observed main effects.

General Music

Table 5.27 contains noteworthy regression results for general music academic outcomes.

Main Effects

Taking 1-2 years of general music coursework was not associated with any large or significant main effects.

Table 5.27

Academic Conditional Mean Comparisons for 1-2 Years of General Music

Dependent Variable	Unadjusted				PS Adjusted			
	Main Effect	8 th Grade	Sex	Race	Main Effect	8 th Grade	Sex	Race
College Exam Math Score	n.s.	n.s.	Female: $g = +.14^*$	n.s.	n.s.	n.s.	n.s.	n.s.
College Exam Reading Score	n.s.	n.s.	Female: $g = +.20^*$	n.s.	n.s.	n.s.	Female: $g = +.17^*$	n.s.
Mathematics Reasoning	n.s.	A or B: $g = +.08^*$ C, D, or F: $g = -.17^\dagger$	n.s.	n.s.	n.s.	A or B: $g = +.08^*$ C, D, or F: $g = -.15^\dagger$	n.s.	n.s.

$^\dagger p \geq .05$; $^* p < .05$; $^{**} p < .01$; $^{***} p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), "Base-Year to Second Follow-Up, 2018."

Achievement in Eighth Grade

While students' eighth grade achievement did not predict differential performance on college entrance examinations, it did predict differences in mathematics reasoning (Figure 5.49).

While students who earned a grade of A or B in their eighth grade mathematics courses had slightly higher adjusted scores on the mathematics reasoning exam than control students who earned As or Bs ($g = .08$; $p = .022$), students who earned lower grades were predicted to score lower than their comparable peers ($g = -.15$; $p = .104$).

Figure 5.49

Predicted 11th Grade Mathematics Reasoning Scores by 8th Grade Mathematics Achievement for 1-2 Years of General Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Sex

Female general music students scored higher than comparable female peers in the control group on average ($g = +.20$; $p = .024$) on the reading portion of their college examinations (Figure 5.50). This was the only observed difference by sex.

Figure 5.50

Predicted College Entrance Examination Reading Scores by Sex for 1-2 Years of General Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Race

No observable differences arose between general music students in different racial categories.

Regression Results for Social-Emotional Outcomes

Choral Music

The results of the social-emotional regression models for choral music are contained within Tables 5.28 and 5.29.

Main Effects

There were no observable main effects on social-emotional outcomes resulting from participation in choral music.

Table 5.28

Social-Emotional Conditional Mean Comparisons for 1-2 Years of Choral Music

Dependent Variable	Unadjusted				PS Adjusted			
	Main Effect	8 th Grade	Sex	Race	Main Effect	8 th Grade	Sex	Race
School Motivation	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Mathematics Self-Efficacy	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Science Self-Efficacy	n.s.	C, D, or F: $g = +.19^{\dagger}$	n.s.	Students of Color: $g = +.28^*$	n.s.	n.s.	n.s.	Students of Color: $g = +.18^*$

$\dagger p \geq .05$; $* p < .05$; $** p < .01$; $*** p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table 5.29

Social-Emotional Conditional Mean Comparisons for 3-4 Years of Choral Music

Dependent Variable	Unadjusted				PS Adjusted			
	Main Effect	8 th Grade	Sex	Race	Main Effect	8 th Grade	Sex	Race
School Motivation	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Mathematics Self-Efficacy	n.s.	C, D, or F: $g = -.24^{\dagger}$	Male: $g = -.23^{\dagger}$	n.s.	n.s.	C, D, or F: $g = -.15^{\dagger}$	Male: $g = -.16^{\dagger}$	n.s.
Science Self-Efficacy	n.s.	n.s.	n.s.	n.s.	n.s.	C, D, or F: $g = -.19^*$	n.s.	n.s.

$\dagger p \geq .05$; $* p < .05$; $** p < .01$; $*** p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Achievement in Eighth Grade

Enrollment in 3-4 years of choral music was associated with lower ceteris paribus mathematics self-efficacy ($g = -.15$; $p = .399$) and science self-efficacy ($g = -.19$; $p = .025$)

among students who earned low grades in their eighth grade courses. Figures 5.51 and 5.52 depict these results. These effects were not observed among students who participated for only 1-2 years.

Sex

There were few differences by sex. Male students who participated in choir for three or four years had slightly lower average mathematics self-efficacy ($g = -.16$), although this was not statistically significant at $p = .257$. This effect was not observed among students who participated for 1-2 years or for any other dependent variables.

Figure 5.51

Predicted 11th Grade Mathematics Self-Efficacy by 8th Grade Mathematics Achievement for 3-4 Years of Choral Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), "Base-Year to Second Follow-Up, 2018."

Figure 5.52

Predicted 11th Grade Science Self-Efficacy by 8th Grade Mathematics Achievement for 3-4 Years of Choral Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Race

Students of color who enrolled in 1-2 years of choral music were predicted to have higher science self-efficacy than comparable peers who did not enroll in music ($g = +.18$; $p = .041$), although this effect did not persist for those who enrolled for 3-4 years ($g = +.03$; $p = .809$). These results are graphed in Figures 5.53 and 5.54. While this result indicates that participation in choral music is at least somewhat associated with increased science self-efficacy among students of color, the effect was not observed in those students who participated for 3-4 years and its magnitude was small. Follow-up study would be required in order to establish the existence of such an effect for this population.

Figure 5.53

Predicted 11th Grade Science Self-Efficacy by Race for 1-2 Years of Choral Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.54

Predicted 11th Grade Science Self-Efficacy by Race for 3-4 Years of Choral Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Instrumental Music

Noteworthy results from the social-emotional outcome regressions are displayed in Tables 5.30 and 5.31.

Table 5.30

Social-Emotional Conditional Mean Comparisons for 1-2 Years of Instrumental Music

Dependent Variable	Unadjusted				PS Adjusted			
	Main Effect	8 th Grade	Sex	Race	Main Effect	8 th Grade	Sex	Race
School Motivation	n.s.	n.s.	Male: $g = -.20^{\dagger}$	n.s.	n.s.	n.s.	n.s.	n.s.
Mathematics Self-Efficacy	n.s.	C, D, or F: $g = +.20^{\dagger}$	Male: $g = +.15^{\dagger}$	n.s.	n.s.	n.s.	Male: $g = +.16^*$	n.s.
Science Self-Efficacy	n.s.	C, D, or F: $g = +.15^{\dagger}$	Male: $g = +.19^*$	n.s.	n.s.	n.s.	Male: $g = +.18^*$	n.s.

$\dagger p \geq .05$; $* p < .05$; $** p < .01$; $*** p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table 5.31

Social-Emotional Conditional Mean Comparisons for 3-4 Years of Instrumental Music

Dependent Variable	Unadjusted				PS Adjusted			
	Main Effect	8 th Grade	Sex	Race	Main Effect	8 th Grade	Sex	Race
School Motivation	n.s.	A or B: $g = +.11^*$	n.s.	White: $g = +.11^*$	$g = +.11^{**}$	A or B: $g = +.11^{**}$	Male: $g = +.17^{**}$	Students of Color: $g = +.10^*$ White: $g = +.11^{**}$
Mathematics Self-Efficacy	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Science Self-Efficacy	$g = +.14^*$	A or B: $g = +.13^*$ C, D, or F: $g = +.17^\dagger$	Male: $g = +.16^*$	White: $g = +.14^{**}$	$g = +.12^*$	A or B: $g = +.12^*$	Male: $g = +.15^*$	Students of Color: $g = +.11^*$ White: $g = +.12^*$

$\dagger p \geq .05$; $^* p < .05$; $^{**} p < .01$; $^{***} p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), "Base-Year to Second Follow-Up, 2018."

Main Effects

There were no main effects of instrumental music participation on students' mathematics self-efficacy. There were small but significant main effects for students who enrolled in instrumental music for three or four years on school motivation ($g = +.11$; $p = .009$) and science self-efficacy ($g = +.12$; $p = .015$), illustrated in Figures 5.56 and 5.58. These effects were not observed in the population of students who enrolled in instrumental music for only one or two years, illustrated in Figures 5.55 and 5.57.

Figure 5.55

Predicted 11th Grade School Motivation for 1-2 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.56

Predicted 11th Grade School Motivation for 3-4 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.57

Predicted 11th Grade Science Self-Efficacy for 1-2 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.58

Predicted 11th Grade Science Self-Efficacy for 3-4 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Achievement in Eighth Grade

Eighth grade achievement was not a significant predictor of social-emotional outcomes for students who enrolled in 1-2 years. Among students who enrolled for 3-4 years, those who had earned a grade of A or B in their eighth grade class appeared to be driving the main effect for school motivation (Figure 5.59). Students who had lower grades in eighth grade did not have significantly higher school motivation ($g = +.05$; $p = .474$). Science self-efficacy effect sizes were similar (Figure 5.60), although the effect for students with lower grades was nonsignificant ($p = .087$).

Figure 5.59

Predicted 11th Grade Mathematics Self-Efficacy by 8th Grade Mathematics Achievement for 3-4 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HLS:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.60

Predicted 11th Grade Science Self-Efficacy by 8th Grade Science Achievement for 3-4 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Sex

Male students had several social-emotional advantages relative to comparable peers in the control group, but students who were identified as female did not have any. School motivation (Figures 5.61 and 5.62) was higher among male students who enrolled for 3-4 years ($g = +.17$; $p = .005$) but not for 1-2 years ($g = -.12$; $p = .233$). Mathematics self-efficacy (Figures 5.63 and 5.64) was higher among male students who enrolled for 1-2 years ($g = +.16$; $p = .011$) but not 3-4 years ($g = +.05$; $p = .462$). Male students had higher science self-efficacy (Figures 5.65 and 5.66) for enrollment in both 1-2 years ($g = +.16$; $p = .011$) and 3-4 years ($g = +.15$; $p = .019$).

Figure 5.61

Predicted 11th Grade School Motivation by Sex for 1-2 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.62

Predicted 11th Grade School Motivation by Sex for 3-4 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.63

Predicted 11th Grade Mathematics Self-Efficacy by Sex for 1-2 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.64

Predicted 11th Grade Mathematics Self-Efficacy by Sex for 3-4 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.65

Predicted 11th Grade Science Self-Efficacy by Sex for 1-2 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Figure 5.66

Predicted 11th Grade Science Self-Efficacy by Sex for 3-4 Years of Instrumental Music

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Race

Subpopulation effects for racial groups were not substantially different from each other or the main effects.

General Music

As shown in table 5.32, enrollment in 1-2 years of general music had no noteworthy effects on students' social-emotional outcomes.

Table 5.32

Social-Emotional Conditional Mean Comparisons for 1-2 Years of General Music

Dependent Variable	Unadjusted				PS Adjusted			
	Main Effect	8 th Grade	Sex	Race	Main Effect	8 th Grade	Sex	Race
School Motivation	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Mathematics Self-Efficacy	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Science Self-Efficacy	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.

† $p \geq .05$; * $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), "Base-Year to Second Follow-Up, 2018."

Regression Results for Base-Year Social-Emotional Variables

Music Groups Compared With Control Groups

Finally, I estimated linear regression models to determine whether a limited set of control variables moderated the comparisons described in Tables 5.16 through 5.20. Although I did use the HSLs:09 design weights, I did not use the propensity scores since these variables were measured only prior to the "treatment" of high school music participation. The predictors I used were race (student of color or white student), sex (dichotomous), socioeconomic status

(continuous), and whether the student had received a grade of C, D, or F in their eighth grade mathematics course (dichotomous). The nested regression procedure was similar to the other regression analyses. The final models included interactions between each term and the music indicator variable, and I examined postestimation contrasts between the music and non-music groups. Rather than estimating “treatment effects,” these contrasts represented baseline differences between each music group and its corresponding control group.

Table 5.33

Unadjusted Regression Results Summary: 9th Grade School Belongingness and Engagement

Type of Music Enrollment	Belongingness Baseline Differences	Engagement Baseline Differences
Choral Music, 1-2 Years		• $g = +.12^*$
Choral Music, 3-4 Years	• $g = +.19^*$	
Instrumental Music, 1-2 Years		
Instrumental Music, 3-4 Years	• $g = +.20^{**}$	
General Music, 1-2 Years		

† $p \geq .05$; * $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

The regression tables for these analyses are in Appendix E. The t tests in Tables 5.16 through 5.20 revealed several significant differences between music students and the corresponding control groups of students with respect to school engagement. However, even this simple regression moderated the difference in school engagement to nonsignificance for both of the instrumental music groups and the general music group. Only the students enrolled in 1-2

years of choral music had higher baseline levels of school engagement after accounting for the effects of the predictors, although the difference was small ($g = +.12$; $p = .042$).

Comparisons of school belongingness, however, revealed an unmistakable trend. Music students who enrolled for 3-4 years had higher levels of school belongingness than control even when controlling for race, sex, socioeconomic status, and eighth grade mathematics achievement. These effects were also relatively small, but not trivial: $g = +.19$ for the choral music students, and $g = +.20$ for the instrumental music students. Music students who enrolled for 1-2 years of choral music ($g = -.02$; $p = .839$) and instrumental music ($g = +.02$; $g = .681$) did not have higher regression-adjusted school belongingness than control.

Choral Music

Choral music students had an average school belongingness score that was $g = +.17$ standard deviations higher when they enrolled for 3-4 years than when they enrolled for 1-2 years ($p = .051$). Among students participating for 3-4 years, those who earned low grades in eighth grade mathematics ($g = +.44$; $p = .030$), male students ($g = +.22$; $p = .303$), and students of color ($g = +.26$; $p = .181$) all had higher belongingness scores than comparable students who participated for 1-2 years. The eighth grade mathematics subpopulation result is particularly striking, as those who enrolled for 3-4 years and earned a grade of A or B had belongingness that was barely elevated relative to those who enrolled for 1-2 years ($g = +.10$; $p = .334$).

Choral music students in general did not have substantial differences in school engagement. Certain subgroups of choral music students who enrolled for 3-4 years had lower *ceteris paribus* school engagement than those who enrolled for 1-2 years. These subpopulations mirrored the observed differences in belongingness. Students who received low grades in eighth grade mathematics courses ($g = -.27$; $p = .320$), male students ($g = -.38$; $p = .079$), and students

of color ($g = -.30$; $p = .119$) all had lower school engagement after controlling for the other predictors in the model.

Instrumental Music

Among the instrumental music students, the group who enrolled for higher years had higher levels of school belongingness, even after controlling for race, sex, socioeconomic status, and eighth grade mathematics achievement ($g = +.23$; $p = .008$). These differences manifested differently among different student subpopulations. Students who earned a grade of A or B in eighth grade mathematics, and who enrolled for 3-4 years, had higher belongingness scores ($g = +.26$; $p = .005$) than the students who enrolled for only 1-2 years, but for students who earned a lower grade there were no appreciable differences ($g = +.02$; $p = .900$). Students of color who enrolled for 1-2 and 3-4 years were comparable on school belongingness ($g = -.05$; $p = .681$) but white students' belongingness was much higher when they enrolled for 3-4 years ($g = +.41$; $p = .000$). Male students who enrolled for 3-4 years had a larger difference in school belongingness ($g = +.28$; $p = .048$) than female students ($g = +.18$; $p = .065$).

School engagement was not substantially different between the instrumental music students who participated for 1-2 and 3-4 years. Controlling for the other model terms, certain subpopulations who participated for 3-4 years had lower predicted school engagement: students who had low grades in eighth grade mathematics ($g = -.17$; $p = .561$), female students ($g = -.18$; $p = .144$), and students of color ($g = -.24$; $p = .155$). Notably, there was considerable uncertainty surrounding all of these point estimates.

Comparisons Between Ensemble Music Students According to Duration of Enrollment

I also tested whether there were *ceteris paribus* differences in school belongingness or engagement between the groups of ensemble music students who enrolled for 1-2 and 3-4 years.

The *t* tests in Tables 5.21 and 5.22 showed that students who enrolled for 3-4 years had higher school belongingness, but not higher school motivation, than students who enrolled for 1-2 years. However, I wanted to partial out the effects of demographic or achievement differences between the groups. I used the same regression strategy as in the previous section, but for each type of music I substituted the group enrolled for 1-2 years as the “control” group and the group enrolled for 3-4 years as the “treatment” group. Table 5.34 presents notable results from the resulting postestimation comparisons.

Table 5.34

Social-Emotional Baseline Conditional Mean Comparisons Between Students Enrolled for 3-4 Years and 1-2 Years

Dependent Variable	Choral Music				Instrumental Music			
	Population	8 th Grade	Sex	Race	Population	8 th Grade	Sex	Race
School Belongingness	$g = +.17^{\dagger}$	C, D, or F: $g = +.44^*$	Male: $g = +.22^{\dagger}$	Students of Color: $g = +.26^{\dagger}$	$g = +.23^{**}$	A or B: $g = +.26^{**}$	n.s.	White: $g = +.41^{***}$
School Engagement	n.s.	C, D, or F: $g = -.27^{\dagger}$	Male: $g = -.38^{\dagger}$	Students of Color: $g = -.30^{\dagger}$	n.s.	C, D, or F: $g = -.17^{\dagger}$	Female: $g = -.18^{\dagger}$	Students of Color: $g = -.24^{\dagger}$

$^{\dagger} p \geq .05$; $^* p < .05$; $^{**} p < .01$; $^{***} p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Summary

After accounting for the other predictors, both the choral and instrumental music students who enrolled for 3-4 years had lower average scores than students who enrolled for 1-2 years on the school engagement metric. Yet, as Figure 5.28 indicates, marginal school engagement scores were comparable for both groups. This implies that these differences in students’ levels of school

engagement, while sizeable in some instances, are difficult to interpret and may be regression artifacts. The school belongingness scores tell a different story. School belongingness is associated with persistence in high school ensemble music, even after accounting for the race, sex, socioeconomic status, and eighth grade mathematics achievement of the students. White and male instrumental music students, and those earning an A or B in the mathematics course, who enrolled for 3-4 years had higher predicted school belongingness than those who enrolled for 1-2 years. Conversely, the subpopulations of choral music students who had higher predicted belongingness when they enrolled for 3-4 years were students earning a grade of C, D, or F in eighth grade mathematics, male students, and students of color. These results imply that the students who persist in ensemble music may have different marginal social-emotional profiles.

Chapter 6: Discussion

Review of Purpose

The purpose of this study was to examine the association between high school music participation, academic achievement, and self-efficacy among students in American high schools. The questions guiding this research were:

1. Is there an association between high school music coursetaking and academic achievement, as measured by test scores?
2. Is there an association between high school music coursetaking and social-emotional outcomes such as self-efficacy, belongingness, or motivation?
3. Do these relationships differ by type of music studied or the duration of study?

Review of Methodology

The data source for this study was a secondary dataset, the nationally representative High School Longitudinal Study of 2009 produced by the National Center for Education Statistics at the U.S. Department of Education. Extensive transcript review and classification facilitated the creation of six “treatment” groups of music students: 1-2 years of choral music, 3-4 years of choral music, 1-2 years of instrumental music, 3-4 years of instrumental music, 1-2 years of general music, and 3-4 years of general music. I also created control groups for each type of music consisting of students who had not enrolled in any high school music courses and who attended schools that verifiably offered the type of music in question.

Since prior research has consistently demonstrated pre-treatment differences between students enrolled in certain types of music instruction and their peers who did not enroll, this analysis used propensity scores to control for overt selection bias between the students who did and did not participate in each type of music in high school. I used propensity-score-adjusted ordinary least squares regression models with music participation as an indicator variable and involved in interaction terms with other predictors, and compared postestimation predicted scores between each music “treatment” group and the control group across six outcome variables: college entrance examination mathematics and reading scores, mathematics reasoning, school motivation, mathematics self-efficacy, and science self-efficacy. I also used regression without the propensity score weights to examine the differences between groups of music students who enrolled for 1-2 and 3-4 years.

Summary of Findings

Research Questions 1 and 3: Academic Outcomes

This study sought to determine whether high school music participation was associated with higher standardized test scores, and whether any such associations differed according to type of music or students’ duration of enrollment.

College Examination Mathematics Scores

This study did not find any large or significant main effects of high school music enrollment on college entrance examination mathematics scores (Table 6.1). Both choral and instrumental students who enrolled for 1-2 years and who received a grade of C, D, or F in their 8th grade mathematics courses had substantially higher scores than comparable peers—among the largest effects in the study. This was true even though a higher proportion of choral music students received a low grade and a lower proportion of instrumental music students did. This

finding is tempered somewhat by the fact that these benefits were not present for those who participated for 3-4 years, but would have been expected to intensify if music enrollment were the cause. Male students who enrolled in 3-4 years of choir were predicted to score higher than comparable peers on this measure, although the difference was not statistically significant and male students with shorter enrollments did not enjoy higher college examination mathematics scores. Female general and instrumental students were predicted to have higher *ceteris paribus* college examination mathematics scores than female students in the control group, although again for female students this effect was observed only among those who enrolled for 1-2 years and not for 3-4 years.

Table 6.1

Propensity Score Adjusted Results Summary: College Entrance Examination Mathematics Score

Type of Music Enrollment	Main Effects	Subpopulation Effects
Choral Music, 1-2 Years		<ul style="list-style-type: none"> • C, D, or F in 8th Grade Math: $g = +.37^*$
Choral Music, 3-4 Years		<ul style="list-style-type: none"> • Male: $g = +.21^\dagger$
Instrumental Music, 1-2 Years		<ul style="list-style-type: none"> • C, D, or F in 8th Grade Math: $g = +.39^\dagger$ • Female: $g = +.18^*$
Instrumental Music, 3-4 Years		
General Music, 1-2 Years		<ul style="list-style-type: none"> • Female: $g = +.20^*$

$^\dagger p \geq .05$; $^* p < .05$; $^{**} p < .01$; $^{***} p < .001$

Note. In order to facilitate interpretation, in Tables 6.1 through 6.6, subpopulation effects that did not differ appreciably from main effects are omitted. Subpopulation effects beyond main effects are listed in parentheses.

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

College Examination Reading Scores

Table 6.2 shows that students who enrolled in instrumental music were predicted to have overall higher college entrance examination reading scores than comparable peers, regardless of the duration of their enrollment. The conclusion that participation in instrumental music is the cause of this effect may not be warranted, however. If music were the cause, one would expect to see the effect grow stronger as the duration of participation increased. In this case, this did not occur; the ostensible benefits of 3-4 years of enrollment in instrumental music were slightly smaller. Still, the two populations of instrumental students were significantly different in important respects, and it is noteworthy that this effect was consistent over the two different weightings of the students in the control group.

It is possible that there is a ceiling effect exerting downward pressure on these students' score advantages (consistent with Miksza, 2007), or that the benefits of music on reading scores are real and that there is a diminishing rate of return above a certain "dosage." Furthermore, it is possible that the comparable pseudopopulation of students who participated for 3-4 years has an overrepresentation of students who initially participated in music, but dropped out prior to high school. Finally, it is also noteworthy that while both male and female instrumental students were predicted to score higher than their comparable peers who did not enroll in music regardless of the duration of enrollment, female instrumental students' score advantages were larger in both cases. The presence of such an effect in a nationally representative dataset controlling for both selection and confounding is an important finding from this research.

There were no other significant main effects for reading scores. Male students and students of color who enrolled in choral music for 3-4 years had higher reading scores on

average than comparable peers. However, these differences were not statistically significant. Female general music students had higher college examination reading scores as well.

Table 6.2

Propensity Score Adjusted Results Summary: College Entrance Examination Reading Score

Type of Music Enrollment	Main Effects	Subpopulation Effects
Choral Music, 1-2 Years		
Choral Music, 3-4 Years		<ul style="list-style-type: none"> • Male: $g = +.31^{\dagger}$ • Students of color: $g = +.19^{\dagger}$
Instrumental Music, 1-2 Years	<ul style="list-style-type: none"> • $g = +.23^{**}$ 	<ul style="list-style-type: none"> • (Female: $g = +.32^{**}$)
Instrumental Music, 3-4 Years	<ul style="list-style-type: none"> • $g = +.21^{**}$ 	<ul style="list-style-type: none"> • (Female: $g = +.27^{***}$)
General Music, 1-2 Years		<ul style="list-style-type: none"> • Female: $g = +.17^{*}$

$\dagger p \geq .05$; $* p < .05$; $** p < .01$; $*** p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Mathematics Reasoning

The only mathematics reasoning main effect that met either the threshold for statistical significance or effect size was among instrumental music students who enrolled for 3-4 years (Table 6.3). Interestingly, as this effect appeared to be driven almost entirely by male students, it may be considered less of a main effect and more of a subpopulation effect. It was not observed among students who enrolled in instrumental music for only one or two years. The other subpopulation effects did not exhibit a clear pattern. Among students of color who enrolled in 1-2 years of choral music, mathematics reasoning was predicted to be lower than the control group even after accounting for the other predictors. Male students who enrolled for 3-4 years of choral music were predicted to exhibit higher mathematics reasoning ability than comparable males in

the control group. Students who entered general music study having earned an A or B in 8th grade mathematics were predicted to score higher than comparable peers, while students who entered general music study with a low grade were predicted to score lower.

Table 6.3

Propensity Score Adjusted Results Summary: Mathematics Reasoning Score

Type of Music Enrollment	Main Effects	Subpopulation Effects
Choral Music, 1-2 Years		• Students of Color: $g = -.15^*$
Choral Music, 3-4 Years		• Male: $g = +.18^\dagger$
Instrumental Music, 1-2 Years		
Instrumental Music, 3-4 Years	• ($g = +.13^{***}$)	• Male: $g = +.25^{***}$
General Music, 1-2 Years		• A or B: $g = +.08^*$ • C, D, or F: $g = -.15^\dagger$

$^\dagger p \geq .05$; $^* p < .05$; $^{**} p < .01$; $^{***} p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Summary

This analysis did not find widespread effects of music participation on students’ academic achievement as measured by their college entrance examination and mathematics reasoning scores. The only main effect that was observed for both durations of participation was the reading scores of instrumental music students. While there was not an undeniable pattern or incontrovertible evidence, the results of this analysis did suggest the possibility that music participation was associated with higher academic achievement for certain subpopulations of music students, even after accounting for overt selection bias through propensity score

weighting. Effect sizes for these effects were mainly between +0.2 and +0.3 standard deviations, with some smaller and some larger. The What Works Clearinghouse reporting program of the Institute for Education Sciences at the U.S. Department of Education classified standardized effect sizes of greater than .25 as “substantively important” (What Works Clearinghouse, 2014, p. 23). Several subpopulation effects met this standard, and the estimated main effects of instrumental music education on reading scores were nearly as large.

Table 6.4

Propensity Score Adjusted Results Summary: School Motivation

Type of Music Enrollment	Main Effects	Subpopulation Effects
Choral Music, 1-2 Years		
Choral Music, 3-4 Years		
Instrumental Music, 1-2 Years		
Instrumental Music, 3-4 Years	• $g = +.11^*$	
General Music, 1-2 Years		

† $p \geq .05$; * $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Research Questions 2 and 3: Social-Emotional Outcomes

This study sought to determine whether high school music participation was associated with higher scores on social-emotional measures, and whether any such associations differed according to type of music or students’ duration of enrollment.

School Motivation

Most of the propensity-score-adjusted comparisons between the school motivation scores of the music “treatment” groups and the control group were small (Table 6.4). The only result that met the reporting threshold was that students who enrolled in 3-4 years of instrumental music had slightly but significantly higher school motivation than comparable peers who did not participate in high school music. This was the dependent variable with the lowest number of effects that met the reporting threshold.

Mathematics Self-Efficacy

There were no observed main effects of music participation on students’ mathematics self-efficacy. While three subpopulation effects emerged, there was not a clear pattern among them. The subpopulation effects among students who enrolled in choral music for 3-4 years were both small and negative, while the one for students enrolled in instrumental music for 1-2 years was small and positive. Table 6.5 contains the relevant results for mathematics self-efficacy.

Table 6.5

Propensity Score Adjusted Results Summary: Mathematics Self-Efficacy

Type of Music Enrollment	Main Effects	Subpopulation Effects
Choral Music, 1-2 Years		
Choral Music, 3-4 Years		<ul style="list-style-type: none">• C, D, or F in 8th Grade Math: $g = -.15^{\dagger}$• Male: $g = -.16^{\dagger}$
Instrumental Music, 1-2 Years		<ul style="list-style-type: none">• Male: $g = +.16^*$
Instrumental Music, 3-4 Years		
General Music, 1-2 Years		

$\dagger p \geq .05$; $* p < .05$; $** p < .01$; $*** p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Science Self-Efficacy

Students enrolled for 3-4 years of instrumental music had higher science self-efficacy than comparable peers. However, this effect was not observed among students who enrolled for 1-2 years, nor among any of the other music “treatment” groups. There were no discernible trends among the other three observed subpopulation effects, as shown in Table 6.6, although all were small and positive.

Table 6.6

Propensity Score Adjusted Results Summary: Science Self-Efficacy

Type of Music Enrollment	Main Effects	Subpopulation Effects
Choral Music, 1-2 Years		• Students of Color: $g = +.18^*$
Choral Music, 3-4 Years		• C, D, or F in 8 th Grade Science $g = -.19^*$
Instrumental Music, 1-2 Years		• Male: $g = +.18^*$
Instrumental Music, 3-4 Years	• $g = +.12^*$	
General Music, 1-2 Years		

† $p \geq .05$; * $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Summary

This study also did not find widespread effects of high school music participation on students’ social-emotional abilities, as measured by the HSLS’s proprietary standardized measures of school motivation, mathematics self-efficacy, and science self-efficacy. While this is similar to the summary of the results for academic outcomes, there were even fewer noteworthy main and subpopulation effects for social-emotional outcomes, with none achieving the What

Works Clearinghouse's (2014) "substantively important" threshold. There was also less evidence that was suggestive of possible trends worthy of future investigation, at least with similarly measured variables. It is possible that the measurement of these affective constructs in the HSLS:09 was less than ideal, or perhaps oriented toward a vision of school motivation that was unsuited for detecting particular benefits that arose from music participation. However, it is at least equally likely that high school music participation does not have notable impacts on students' global sense of school motivation, or on mathematics or science self-efficacy.

Propensity score adjustment was not possible for the social-emotional outcomes measured only in 9th grade. Still, it is noteworthy that higher 9th grade school belongingness was observed among students who eventually elected to remain enrolled in ensembles for 3-4 years, relative to students who enrolled for only 1-2 years or who did not enroll, even after regression on several predictors. In addition, both the choral and instrumental music students had higher baseline school engagement scores than the control group. Unfortunately, the fact that these variables were not measured in 11th grade means that there was no way to test for possible effects of high school music participation on belongingness or engagement. The extent to which these results cause music participation, are caused by music participation, or both is a pertinent question for future investigation.

Discussion

The literature is full of studies examining school music's ostensible extramusical benefits (e.g., Catterall et al., 2012; Gouzouasis et al., 2007; Guhn et al., 2020; Helmrich, 2010; Kinney, 2008; Vaughn & Winner, 2000), including analyses of NCES datasets (e.g., Miksza, 2007, 2010; Southgate & Roscigno, 2009). In many cases, students who enrolled in curricular music programs outperformed students who did not on various measures of academic achievement. In

others, mixed or null findings arose. However, in studies using techniques to control for selection bias (e.g., Elpus, 2013b), the differences were nonsignificant. This study largely supports Elpus's conclusion that pretreatment differences are the major reason that music students have higher standardized test scores on average. The exception is the instrumental music students' higher reading scores, which are a central finding of this investigation.

Music educators have long touted music participation as a way to boost social-emotional abilities such as school belongingness, engagement, motivation, and resilience (see Edgar, 2017). As social-emotional learning is increasingly a mainstay in the educational zeitgeist, these beliefs would seem to be a natural element of music education advocacy if they were supported by empirical research. There is a fair amount of speculation regarding music's social-emotional benefits (e.g., Edgar, 2017; Eisner, 1999a), but the only rigorous empirical study of music's connection with social-emotional outcomes (Foster & Jenkins, 2017) led the authors to conclusions similar to Elpus's: that any observed score advantages for music students were due to selection bias.

The results of this investigation cohere with Foster and Jenkins's findings: they do not support the hypothesis that music participation contributes to school motivation, mathematics self-efficacy, or science self-efficacy insofar as they were measured in the HSLS:09. Future studies may uncover social-emotional benefits of music instruction with different outcome variables or different operationalization of these variables. Even so, in the words of Winner and Cooper (2000), it is critical for the profession to "mute those claims" regarding universal social-emotional benefits of music instruction until they can be based upon sound social science.

Implications for Music Education Advocacy

There are serious downsides to leading policymakers and the public to believe that they should support music programs due to instrumental claims of extramusical benefits, even if such a strategy appears effective in the short term (Demorest & Morrison, 2000; Hetland & Winner, 2001; Kohn, 2015; Reimer, 1999, 2005; C. West & Clauhs, 2015). If music is subsequently found to be ineffective for this purpose, which seems likely given the best research evidence currently available, not only will its standing suffer in the eyes of stakeholders, these stakeholders will also have shifted their attention away from music's intrinsic value. Music programs themselves may be twisted toward utilitarian functions such as workplace preparation and away from musicians' core values such as joy, spontaneity, and discovery (Koza, 2006). As Jorgensen presciently observed more than twenty-five years ago:

The profession finds itself caught in somewhat of a bind: it is convinced of music's intrinsic value, and yet, for political reasons at least, it must appeal to music's utilitarian contributions to society and education, and its extrinsic as opposed to intrinsic value.

What is more, these two sets of values do not always mesh easily. (1994, p. 25)

This study does not get the profession out of this "bind." However, it is yet another sign that the benefits music educators and advocates "know" to result from music participation may be the result of overrepresentation of high achievers in music rather than some sort of cognitive rewiring. Utilitarian appeals may be effective advocacy, particularly in the short term, but they will, to a certain extent, be disingenuous until the preponderance of the research evidence changes. There may be something going on, whether cognitive, attitudinal, self-regulatory, or otherwise, but scholars have yet to conclusively identify exactly what it is or how it works, and a truthful accounting of music's benefits necessarily includes this disclaimer.

Researchers, advocates, and organizations desiring quantitative evidence that music education is effective would do well to consider the extent to which they are accepting questionable premises about its aims, and design future studies according to their own values. Discovery of particular approaches or situations that do appear to have measurable benefits for students may provide an evidence base that leads music educators toward improved teaching practice. They may just as easily point music education in a direction with which certain constituencies are uncomfortable. If future research finds that music instruction emphasizing mathematical relationships between note values improves mathematics scores, the implications could be the beginning of a long-term corruption of the values many music educators hold. As another hypothetical example, suppose that future research finds that music does have benefits for certain student profiles, but not for others. Would those students' parents be less inclined to encourage them to participate? What if students best learn persistence and discipline through the teaching style lamented in Allsup and Benedict's (2008) *The Problems of Band*? These questions do not have easy answers, but as music education does not have a great deal of research upon which to base "evidence-based practice," even preliminary evidence could prompt healthy dialogue.

Implications for Future Research on Extramusical Benefits

Elpus's study used fixed-effects regression to attenuate selection bias by using differencing to control for confounding influences, whether observed or unobserved. This study used propensity scores to control for observed confounding, but propensity scores' major weakness is that they do not control for unobserved confounding. While this study primarily coheres with Elpus's findings "opposing the existence of a causal link between high school music course taking and standardized test scores" (2013b, p. 16), some tangible score advantages

did emerge for certain groups of music students. Yet, this study incorporated over one hundred student, family, and school characteristics. One can imagine this kind of data being more than enough for the typical school official, but addressing overt selection bias may be an insufficient strategy to arrive at a satisfactory answer to this research question for researchers. It is sensible to consider potential variables that were uncontrolled in this study, as future research on them or incorporating them would fill major knowledge gaps:

A more pressing concern for the profession, however, is to begin to understand the root cause of the preexisting systematic differences between music and non-music students.

The extant research offers no concrete examination of why students who are more predisposed toward academic success are enrolled in music in greater numbers than are their peers, and there is little extant theory to guide an empirical investigation into this phenomenon. (Elpus, 2013b, p. 17)

The recent literature on the demographics of music education enrollment and retention (Elpus & Abril, 2011, 2019; Kinney, 2008, 2010, 2018) has identified certain characteristics that are frequently available in administrative datasets, but other constructs are certainly involved.

Unobserved Within-Person Variables

Research on music's benefits in the music education literature has been conducted on a somewhat separate track than scholarship in psychological literature. Yet, evidence from psychology unmistakably leads to the conclusion that, in order to truly be convincing, studies in music education should account for certain constructs, many of which arise from the psychological literature, that are typically not contained within even the most rigorous secondary datasets. These include executive function and IQ (Babo, 2004; Corrigall & Schellenberg, 2015; Costa-Giomi, 1999; Schellenberg, 2006, 2011; Schellenberg & Mankarious, 2012), personality

(Corrigan et al., 2013; Corrigan & Schellenberg, 2015; Hille & Schupp, 2015), and certain parent characteristics (Corrigan & Schellenberg, 2015).

Another glaring set of omitted variables has to do with students' musical aptitude, engagement, achievement, and exposure. While there is no single universally accepted measure of students' musical aptitude or ability, it is undeniable that students differ in important respects related to musical success (McPherson, 1995; McPherson et al., 2012) prior to the first day of school music. Furthermore, parents' musical backgrounds affect their children's enrollment, retention, and achievement in music (Corrigan & Schellenberg, 2015; Davidson et al., 1996; Howe et al., 1998; McPherson, 2009; Pitts et al., 2000). In addition, although certain factors related to enrollment and persistence are controlled in this study, other facets known to be or potentially related to enrollment decisions and commitment are uncontrolled. Unwilling participants are assumed to be no different from their peers who strongly identify with school music or invest above-average time and effort. There is no measure of students' satisfaction, feelings of social connection, intrinsic motivation relative to extrinsic motivation, or reasons for pursuing music in the first place. It is plausible but unproven that increased individual-level engagement with music strengthens its associations with other constructs.

Researchers using large datasets such as those available from NCES are accepting that these are omitted variables and assuming that these constructs will be reflected in aggregated results, but this may not be the case. For example, practice has been demonstrated to be central to success in instrumental music (Davidson et al., 1996; McPherson, 2005, 2009). What if sustained music study in school leads to increased self-efficacy for students who score high in conscientiousness in the Five-Factor Model, who are dispositionally well suited for disciplined practice, but leads to decreased self-efficacy for those with low conscientiousness? As

personality is typically an omitted variable, this hypothetical reality would show up as a muddled and likely null effect. In another hypothetical situation, what if music participation is shown to have null effects in the general population, but measurable benefits for children with strong auditory processing ability?

The presence of unobserved within-person characteristics naturally lead to the fascinating hypothetical existence of unobserved subpopulations. Among students who participated in instrumental music for 3-4 years, mathematics reasoning scores showed a “main” effect of $g = +.13$ that was really a result of approximately half of these students (the male subpopulation) that registered as $g = +.25$. What if there are latent traits, which are certainly correlated with these and other observed subpopulations, that may either be driving or suppressing these results? For example, it may be the case that male students who are high in conscientiousness are overrepresented in the population of high school instrumental music students and that these students are the main contributors to the overall higher mathematics reasoning scores among male students. Alternatively, benefits among unobserved subpopulations could be suppressed by aggregation. Consider a hypothetical scenario in which the diligence and persistence required for success in music performance lead to increased academic achievement for male students with low conscientiousness but not for males with high pretreatment conscientiousness. A powerful synergy may arise if the foci of the educational and psychological traditions can be merged.

Unobserved Features of Individual Schools’ Music Instruction

There are as many approaches to teaching music as there are music teachers, and aggregation of hundreds of such approaches and situations into a single “music: yes or no” indicator variable is riddled with unobserved heterogeneity. Teacher experience, expertise, effort, philosophy, and personality could all conceivably affect how students are shaped by their

interactions. Edgar (2017), Kaschub (2002), Shields (2001), and others have described instruction that would seemingly provide social-emotional benefits for music students. On the other hand, the competition and director-centered instruction that have often typified American ensemble instruction may have adverse effects on students' social-emotional outcomes (Evans, 2015; Hendricks, 2014; Kaschub, 2002). It is likely that student, teacher, and program interact with one another in complex and dynamic ways that are not well understood.

Beyond factors relating solely to the teacher, aspects of the instructional setting may also matter. The status of and support for the music program within a school, attrition and private lesson rates, when and how beginners are taught, and other factors may “average out” of these kinds of studies, but they may also be concealing critical associations. What if music programs that are judged to be somehow “better” have more benefits? Johnson and Memmott (2006), and Schneider and Klotz (2000) made initial attempts to answer this research question, but it remains underexplored. Furthermore, what if factors that make a program observably “better” (e.g., large enrollments or competitive success) are not the same factors that cultivate extramusical benefits? What if the factors that make a music program “better” depend on characteristics of the student? Different music programs may offer different levels of social support and belongingness to students with different races, genders and gender identities, and other characteristics. The educational and policy outcomes downstream from these sorts of questions are murky, and datasets not directly intended to support music education research will almost certainly not contain this kind of information.

Unobserved Music Study

Music participation that is not reflected in transcripts is another source of confounding. Many types of formal (e.g., classical piano) and informal (e.g., guitar) music making primarily

occur outside of school. There are also musical activities with social dimensions. Some are organized around religions and even families, and some are informal engagements such as a garage band. Some children attend concerts, create and listen to music through their phones or computers, and may adopt a musician identity even though they do not feel like a part of their school's music program (R. A. MacDonald, 2002).

Even within students clearly participating in their schools' music courses, differences in amount of practice, practice quality, private lesson participation, and even instrument ownership can reasonably be said to be affecting the music "dosage" each student receives. Furthermore, in studies such as this one, the control group comprised of students who did not select music courses in high school undoubtedly contains a sizeable number of students who did participate in school music prior to high school, even though they may have enrolled in middle school for three or four years. This point is central in the interpretation of the fact that ensemble music students had elevated 9th grade school belongingness and engagement. Future research may attempt to determine whether this, too, is a result of selection bias, or whether features of ensemble music instruction make unique contributions to students' social-emotional outcomes. Although some factors related to attrition in the literature were controlled in one or both of the propensity score analysis or regression analysis, all of this makes it impossible to draw tidy boundaries around a "musician" group of students and another "non-musician" group. This methodological hurdle would be greatly reduced with even a small set of music-specific survey questions in future studies.

Emphasis on Subpopulations

Continued research on this topic may do well to focus less on a monolithic conception of student performance, and instead probe the abilities and achievement of disaggregated groups of

students. Even if future research consistently finds an average of no effect, the unmodeled variability surrounding these point estimates will represent missed opportunities to tell a story that is more precise, more accurate, and ultimately more effective for the profession. The possible influences of person and context leave the profession's current understanding of this research area in limbo. Disaggregation may lead us to the overlap of two seemingly competing truths: on one hand, many practitioners, parents, advocates, and students are convinced that they have seen music's benefits with their own eyes, and on the other hand, empirical research says that on average this is not so. The question for researchers may not be *if* music education has detectable extramusical benefits, but rather *for whom*, and *in what situations*, these benefits accrue:

Assuming a homogenous population of band, string, and choir students could bias the conclusion that all music students differ from nonmusic students. Further investigation using disaggregated definitions of music elective participation would help clarify and extend previous findings. (Kinney, 2018, p. 3)

Table 6.7 provides preliminary if underwhelming evidence that academic achievement benefits may accrue to music students who received low grades in eighth grade courses. More broadly, it is possible that music instruction might be associated with higher achievement among some populations of students than other observed characteristics would indicate. This hypothesis would be supported more strongly if the effects for one or two years of choral and instrumental music had persisted or strengthened among students who enrolled for three or four years, or if it had been observed among other outcomes. The finding that general music was associated with higher mathematics reasoning scores among students with an A or B in the relevant eighth grade course but lower scores among students who received a grade of C, D, or F also runs counter to

this hypothesis. Still, future research examining Catterall's (2009) peer-group hypothesis, that music education can surround lower-achieving students with high-achieving peers, would be intriguing.

Table 6.7

Propensity Score Adjusted Subpopulation Summary: Academic Outcomes by Eighth Grade

Achievement

Type of Music Enrollment	A or B	C, D, or F
Choral Music, 1-2 Years		• College Exam Math: $g = +.37^*$
Choral Music, 3-4 Years		
Instrumental Music, 1-2 Years		• College Exam Math: $g = +.39^\dagger$
Instrumental Music, 3-4 Years		
General Music, 1-2 Years	• Mathematics Reasoning: $g = +.08^*$	• Mathematics Reasoning: $g = -.15^\dagger$

$^\dagger p \geq .05$; $^* p < .05$; $^{**} p < .01$; $^{***} p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), "Base-Year to Second Follow-Up, 2018."

Table 6.8

Propensity Score Adjusted Subpopulation Summary: Academic Outcomes by Sex

Type of Music Enrollment	Female Students	Male Students
Choral Music, 1-2 Years		
Choral Music, 3-4 Years		• College Exam Math: $g = +.21^\dagger$ • College Exam Reading: $g = +.31^\dagger$
Instrumental Music, 1-2 Years	• College Exam Math: $g = +.18^*$ • College Exam Reading: $g = +.32^{**}$	
Instrumental Music, 3-4 Years	• College Exam Reading: $g = +.27^{***}$	• Mathematics Reasoning: $g = +.25^{***}$
General Music, 1-2 Years	• College Exam Math: $g = +.20^*$ • College Exam Reading: $g = +.17^*$	

$^\dagger p \geq .05$; $^* p < .05$; $^{**} p < .01$; $^{***} p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), "Base-Year to Second Follow-Up, 2018."

As detailed in Table 6.8, there were many differential outcomes according to students' sex. The fact that male students who enrolled in choral music for 3-4 years had higher *ceteris paribus* college entrance examination scores, while there were no discernible benefits for female students, was noteworthy and worthy of follow-up investigation. Equally intriguing was the association between instrumental music study and females' academic achievement related to female peers, and comparably slight association between instrumental music and males' achievement. A thorough treatment of gendered dynamics at play in the selection and socialization of music students in different performance areas is beyond the scope of this paper, but these dynamics should be considered in future research involving the academic performance of music students.

Table 6.9

Propensity Score Adjusted Subpopulation Summary: Academic Outcomes by Race

Type of Music Enrollment	Students of Color	White Students
Choral Music, 1-2 Years	• Mathematics Reasoning: $g = -.15^*$	
Choral Music, 3-4 Years	• College Exam Reading: $g = +.19^\dagger$ • Mathematics Reasoning: $g = +.18^\dagger$	
Instrumental Music, 1-2 Years		
Instrumental Music, 3-4 Years		
General Music, 1-2 Years		

$^\dagger p \geq .05$; $* p < .05$; $** p < .01$; $*** p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), "Base-Year to Second Follow-Up, 2018."

Table 6.9 reveals that peer comparisons of the academic performance of instrumental and general music students were not differentiated by students' race. It is striking that all three of the

associations between music study academic achievement among students of color were for choral music students. Those who participated for 1-2 years had lower comparable mathematics reasoning, but it was higher for those who participated for 3-4 years. The students of color who enrolled for 3-4 years also had higher college examination reading scores. This may be a fruitful area for future inquiry as well.

There was scant evidence of social-emotional benefits for subpopulations defined by eighth grade achievement, sex, or race in these analyses (Tables 6.10, 6.11, and 6.12). In most of the regression results in Appendix E, the strongest predictors of both 11th grade self-efficacy outcomes were the 9th grade self-efficacy pretests; that is, self-efficacy was relatively stable during the period of the study. It is conceivable that previous music enrollment was related to 9th grade self-efficacy, but that is impossible to determine from this dataset.

Table 6.10

Propensity Score Adjusted Subpopulation Summary: Social-Emotional Outcomes by Eighth Grade Achievement

Type of Music Enrollment	A or B	C, D, or F
Choral Music, 1-2 Years		
Choral Music, 3-4 Years		<ul style="list-style-type: none"> • Mathematics Self-Efficacy: $g = -.15^{\dagger}$ • Science Self-Efficacy: $g = -.19^*$
Instrumental Music, 1-2 Years		
Instrumental Music, 3-4 Years		
General Music, 1-2 Years		

$\dagger p \geq .05$; $* p < .05$; $** p < .01$; $*** p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table 6.11

Propensity Score Adjusted Subpopulation Summary: Social-Emotional Outcomes by Sex

Type of Music Enrollment	Female Students	Male Students
Choral Music, 1-2 Years		
Choral Music, 3-4 Years		• Mathematics Self-Efficacy: $g = -.16^\dagger$
Instrumental Music, 1-2 Years		• Mathematics Self-Efficacy: $g = +.16^\dagger$ • Science Self-Efficacy: $g = +.18^*$
Instrumental Music, 3-4 Years		
General Music, 1-2 Years		

$^\dagger p \geq .05$; $^* p < .05$; $^{**} p < .01$; $^{***} p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HLS:09), “Base-Year to Second Follow-Up, 2018.”

Table 6.12

Propensity Score Adjusted Subpopulation Summary: Social-Emotional Outcomes by Race

Type of Music Enrollment	Students of Color	White Students
Choral Music, 1-2 Years	• Science Self-Efficacy: $g = +.18^*$	
Choral Music, 3-4 Years		
Instrumental Music, 1-2 Years		
Instrumental Music, 3-4 Years		
General Music, 1-2 Years		

$^\dagger p \geq .05$; $^* p < .05$; $^{**} p < .01$; $^{***} p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HLS:09), “Base-Year to Second Follow-Up, 2018.”

Propensity Score Adjustment and Regression Adjustment

This study used the “doubly robust” framework (Bang & Robins, 2005; see also Ho et al., 2007; Rubin, 2001) which combines propensity score adjustment and regression adjustment. In many instances the regression adjustment had similar effects to the propensity score adjustment. In such situations, with the benefit of hindsight there would have been no need for both the

regression models and the propensity score weighting. However, as detailed in Table 6.13, a number of effect estimates changed dramatically as a result of one or both of these procedures, demonstrating the value of the doubly robust approach. For example, reading scores for 1-2 years of choral music were similar after regression adjustment with and without the propensity scores. However, in this situation the propensity-score-adjusted t test would have been inadequately controlled. Conversely, regression alone barely moved the unadjusted or adjusted reading scores for 3-4 years of instrumental music, but weighting by propensity scores reduced the unadjusted effect size by more than half.

A second consequence of the doubly robust approach is revealed by the results in Table 6.14. Several noteworthy subpopulation effects emerged only in the propensity-score-adjusted models. Even more unadjusted subpopulation effects, primarily but not only those with reportable effect sizes but those that were not statistically significant, were moderated to nonsignificance by the propensity scores. Further study will be necessary to conclude the importance of this point based on whether these or similar results are observed in other samples. Still, in future investigations of extramusical benefits, it seems prudent to control for confounding with respect to selection and treatment effects.

Table 6.13

Selected Main Effects With and Without Regression and Propensity Score Adjustment

Treatment Group	Outcome	Not Propensity Score Adjusted		Propensity Score Adjusted	
		Unconditional	Regression Adjusted	Unconditional	Regression Adjusted
Choral, 1-2 Years	College Exam Reading Score	-16.495	-13.666	-2.171	-12.076
Choral, 3-4 Years	College Exam Reading Score	+21.102	+18.183	+8.092	+16.120
Choral, 3-4 Years	Mathematics Reasoning	+0.171	-0.075	0.057	-0.069
Instrumental, 1-2 Years	College Exam Mathematics Score	+22.119	+3.491	+19.848	+12.132
Instrumental, 3-4 Years	College Exam Mathematics Score	+38.006	+10.803	+18.405	+9.194
Instrumental, 1-2 Years	College Exam Reading Score	+36.943	+41.720	+30.214	+30.629
Instrumental, 3-4 Years	College Exam Reading Score	+49.027	+53.192	+24.271	+24.897
Instrumental, 3-4 Years	Mathematics Reasoning	+0.599	+0.070	+0.318	+0.068
General, 1-2 Years	Mathematics Reasoning	+0.003	+0.084	+0.050	+0.139
Choral, 3-4 Years	School Motivation	+0.117	+0.072	+0.032	+0.080
Choral, 3-4 Years	Mathematics Self-Efficacy	-0.133	+0.054	-0.071	+0.128
Choral, 1-2 Years	Science Self-Efficacy	+0.065	+0.237	+0.062	+0.167
Instrumental, 3-4 Years	School Motivation	+0.175	+0.035	+0.096	+0.089
Instrumental, 3-4 Years	Mathematics Self-Efficacy	+0.070	-0.247	+0.021	-0.166
General, 1-2 Years	School Motivation	-0.015	-0.122	-0.034	-0.051
General, 1-2 Years	Mathematics Self-Efficacy	+0.019	-0.111	+0.035	-0.029

† $p \geq .05$; * $p < .05$; ** $p < .01$; *** $p < .001$

Note. “Unconditional” comparisons are comparisons between the raw scores of the listed music group and the control group. “Regression Adjusted” comparisons are the differences in postestimation predicted conditional means the listed music group and the control group.

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table 6.14

Subpopulation Effect Differences Between Unadjusted and Adjusted Regression Results

Treatment Group	Outcome	Unadjusted	PS Adjusted
Choral, 3-4 Years	College Exam Mathematics Score	C, D, or F: $g = +.15^\dagger$	n.s.
Choral, 3-4 Years	Mathematics Reasoning	n.s.	Male: $g = +.18^\dagger$
Instrumental, 1-2 Years	College Exam Mathematics Score	White: $g = +.15^\dagger$	n.s.
Instrumental, 1-2 Years	College Exam Mathematics Score	n.s.	Female: $g = +.18^*$
Instrumental, 3-4 Years	College Exam Mathematics Score	C, D, or F: $g = -.15^\dagger$	n.s.
General, 1-2 Years	College Exam Mathematics Score	Female: $g = +.14^*$	n.s.
Choral, 1-2 Years	Science Self-Efficacy	C, D, or F: $g = +.19^\dagger$	n.s.
Choral, 3-4 Years	Science Self-Efficacy	n.s.	C, D, or F: $g = -.19^*$
Instrumental, 1-2 Years	School Motivation	Male: $g = -.20^\dagger$	n.s.
Instrumental, 1-2 Years	Mathematics Self-Efficacy	C, D, or F: $g = +.20^\dagger$	n.s.
Instrumental, 1-2 Years	Science Self-Efficacy	C, D, or F: $g = +.15^\dagger$	n.s.
Instrumental, 3-4 Years	School Motivation	n.s.	Male: $g = +.17^{**}$

$^\dagger p \geq .05$; $* p < .05$; $** p < .01$; $*** p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), "Base-Year to Second Follow-Up, 2018."

Null Hypothesis Significance Testing

The main and subpopulation effects detailed in Tables 6.1 through 6.6 were a mix of results above and below the traditional $p = .05$ threshold. Large effect sizes accompanied by extensive heterogeneity would not be considered by an analyst concerned only with statistical significance, but they are precisely the circumstances that may lead to the discovery of subpopulation effects, if in fact they exist. As a major takeaway from these results is that music may have situational rather than universal benefits, and a possible future direction for this line of

study is to identify these situations, researchers may want to cast as wide a net as possible. In the spirit of Rosnow and Rosenthal's famous quip, "Surely, God loves the .06 nearly as much as the .05" (1989, p. 1277), an arbitrary threshold for statistical significance is by itself an insufficient standard for identifying potential effects in exploratory research with large populations. Given the variability inherent in large-scale studies incorporating a range of students, teachers, and schools, and their interrelationships, one would expect statistically nonsignificant but substantively important results to continue to emerge if researchers are looking for them.

Specificity and Breadth of Dependent Variables

The theory underlying many claims of extramusical benefits is that students develop skills like self-efficacy, critical thinking, or school engagement in music and then subsequently transfer these skills to other pursuits. However, this is mostly based upon assumptions. The extent to which music study actually does develop these abilities for students in school music programs has not been empirically studied on a large scale. It would be valuable for this line of research, for example, to first ascertain whether music study increases music self-efficacy before speculating that music self-efficacy transfers to (i.e., mediates music's relationship with) other school subjects or life situations. Available large-scale datasets do not have music-specific variables.

Available datasets also do not have information about other core benefits of the arts articulated by authors such as Reimer (2003), Eisner (2002), Rabkin (2004), and others: creativity, self-expression, enjoyment, fulfillment, emotional awareness, meaning, self-discovery, and the like. Certainly, one reason that research on music's benefits has revolved around values espoused by those from outside of the field of music education—this study included—is that it relies on data collected by others. Individuals and institutions collect data on

their priorities, and when we analyze others' data, we are pigeonholed into adopting their priorities as our own. In doing so, we may "miss the forest while counting the trees" (Kohn, 2012, p. 36; see also Chapman, 1982, 2001; K. F. McCarthy, 2004).

The Unanswered Question

Elpus's (2013b) article was provocatively titled, "Is it the music, or is it selection bias?" Even if future research can establish that music has benefits that extend beyond selection bias, the question remains of what exactly the putative mechanism is by which "the music" leads to extramusical benefits. Previous scholarship in the psychological tradition (e.g., Costa-Giomi, 2015; Degé et al., 2011; Schellenberg & Peretz, 2008) points toward potential cognitive or neurophysiological mechanisms, while other scholarship in the music education tradition (e.g., Adderley et al., 2003; Bartolome, 2013; Catterall, 2009; Miksza, 2010) highlights social benefits. Still, potential mediators or mechanisms are far from being conclusively described. According to Pearl (2000, pp. 25–26), this undermines the study of extramusical benefits as an advocacy tool.

The element of stability (of mechanisms) is also at the heart of the so-called explanatory accounts of causality, according to which causal models need not encode behavior under intervention but instead aim primarily to provide an "explanation" or "understanding" of how data are generated. Regardless of what use is eventually made of our "understanding" of things, we surely would prefer an understanding in terms of durable relationships, transportable across situations, over those based on transitory relationships. . . It is for reasons of stability that we regard the falling barometer as predicting but not explaining the rain; those predictions are not transportable to situations where the pressure surrounding the barometer is controlled by artificial means. . . Accordingly, we may as well view our unsatiated quest for understanding "how data is generated" or "how

things work” as a quest for acquiring the ability to make predictions under a wider range of circumstances, including circumstances in which things are taken apart, reconfigured, or undergo spontaneous change.

At the heart of an advocacy argument based upon demonstrated extramusical benefits as a result of music enrollment is the implicit or explicit assumption that future music students would enjoy similar results. Pearl’s cautionary note about the necessity of understanding the mechanisms that underpin causal relationships in order to establish them as “stable” rather than “transitory” is yet another reminder of how far away we are from extramusical benefits being settled science.

Summary

Given the astronomical ratio of completed studies to blockbuster findings, it is reasonable to wonder if future examinations of extramusical benefits are worth pursuing at all. Twenty years after the Reviewing Education in the Arts Project and *Champions of Change*, it seems unlikely that future noncausal studies primarily seeking across-the-board main effects will have seismic effects on discourse or policy. However, further research using causal techniques, accounting for music-specific considerations, collecting longitudinal data beginning prior to the first years of school music and extending through or beyond high school, seeking demographically and contextually dependent effects, or investigating benefit-generating mechanisms still has the potential to advance the field. A comprehensive line of inquiry incorporating these elements would take years and require a significant investment of financial and human resources. However, unless and until such commitments are made, the music education community will be left where we are now: with more questions than answers.

Conclusions

1. Controlling for a comprehensive set of variables, high school instrumental music students have higher college entrance examination reading scores than otherwise-comparable peers who did not enroll in music.
2. This analysis does not support the conclusion that music always causes students to have higher *ceteris paribus* standardized test scores. Several groups of music students did have higher test scores than students in the control groups, but nearly all of these instances were due to pretreatment differences.
3. This analysis does not support the conclusion that music causes, or is even associated with, widespread benefits vis-à-vis social-emotional learning. However, ensemble music students exhibit higher baseline levels of school engagement and especially school belongingness, which may be related to middle school music participation.
4. This analysis shows that there is no single population of “music students” or even “choir students” or “band students” and that treating them as such may hamper future research by obscuring within-group differences. In particular, this analysis demonstrates that students who participate in elective music programs for 3-4 years have different profiles than students who participate in the same types of music for 1-2 years.
5. This analysis shows that high school elective music students are systematically different from students who do not elect music in ways that may not be reflected in a minimal set of demographic controls, and that regression adjustment alone is insufficient as a confounder control strategy.

6. This analysis reveals that even putative extramusical benefits may not be universal, but may indeed exist for certain subpopulations, and identification of such subpopulations could be a fruitful area for future research.

References

- Abril, C. R. (2017, May 15). Should we keep using the term “general” in general music? *Music in a Minuet, National Association for Music Education*. <https://nafme.org/should-we-keep-using-term-general-music/>
- Adderley, C., Kennedy, M., & Berz, W. (2003). “A home away from home”: The world of the high school music classroom. *Journal of Research in Music Education*, 51(3), 190–205. <https://doi.org/10.2307/3345373>
- Allison, P. D. (2001). *Missing data*. Sage.
- Allison, P. D. (2009). *Fixed effects regression models*. Sage.
- Allsup, R. E., & Benedict, C. (2008). The problems of band: An inquiry into the future of instrumental music education. *Philosophy of Music Education Review*, 16(2), 156–173.
- Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology*, 84(3), 261–271. <https://doi.org/10.1037/0022-0663.84.3.261>
- Ansari, A., & Pianta, R. C. (2018). The role of elementary school quality in the persistence of preschool effects. *Children and Youth Services Review*, 86, 120–127. <https://doi.org/10.1016/j.childyouth.2018.01.025>
- Apple, M. W. (1979). On analyzing hegemony. In *Knowledge, power, and education: The selected works of Michael W. Apple* (pp. 19–40). Routledge.
- Aronson, J. (Ed.). (2002). *Improving academic achievement: Impact of psychological factors on education*. Academic Press.
- Ashmead, R. D. (2014). *Propensity score methods for estimating causal effects from complex survey data* [Doctoral dissertation, Ohio State University]. http://rave.ohiolink.edu/etdc/view?acc_num=osu1417616653

- Austin, P. C. (2009). Some methods of propensity-score matching had superior performance to others: Results of an empirical investigation and Monte Carlo simulations. *Biometrical Journal*, 51(1), 171–184. <https://doi.org/10.1002/bimj.200810488>
- Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine*, 26(4), 734–753. <https://doi.org/10.1002/sim.2580>
- Austin, P. C., Jembere, N., & Chiu, M. (2018). Propensity score matching and complex surveys. *Statistical Methods in Medical Research*, 27(4), 1240–1257. <https://doi.org/10.1177/0962280216658920>
- Austin, P. C., & Mamdani, M. M. (2006). A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine*, 25(12), 2084–2106. <https://doi.org/10.1002/sim.2328>
- Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, 34(28), 3661–3679. <https://doi.org/10.1002/sim.6607>
- Babo, G. D. (2004). The relationship between instrumental music participation and standardized assessment achievement of middle school students. *Research Studies in Music Education*, 22(1), 14–27. <https://doi.org/10.1177/1321103X040220010301>
- Baker, M., & Johnston, P. (2010). The impact of socioeconomic status on high stakes testing reexamined. *Journal of Instructional Psychology*, 37(3), 193–199.

- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. W. H. Freeman.
- Bandura, A. (2012). On the functional properties of perceived self-efficacy revisited. *Journal of Management*, 38(1), 9–44. <https://doi.org/10.1177/0149206311410606>
- Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4), 962–972. <https://doi.org/10.1111/j.1541-0420.2005.00377.x>
- Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin*, 128(4), 612–637. <https://doi.org/10.1037/0033-2909.128.4.612>
- Bartolome, S. J. (2013). “It’s like a whole bunch of me ”: The perceived values and benefits of the Seattle Girls’ Choir experience. *Journal of Research in Music Education*, 60(4), 395–418. <https://doi.org/10.1177/0022429412464054>
- Bates, V. C. (2012). Social class and school music. *Music Educators Journal*, 98(4), 33–37. <https://doi.org/10.1177/0027432112442944>
- Baum, S., Owen, S., & Oreck, B. (1997). Transferring individual self-regulation processes from arts to academics. *Arts Education Policy Review*, 98(4), 32–39.
- Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117(3), 497–529. <https://doi.org/10.1037/0033-2909.117.3.497>
- Berk, R. A. (2006). An introduction to ensemble methods for data analysis. *Sociological Methods & Research*, 34(3), 263–295. <https://doi.org/10.1177/0049124105283119>

- Berk, R. A., Li, A., & Hickman, L. J. (2005). Statistical difficulties in determining the role of race in capital cases: A re-analysis of data from the State of Maryland. *Journal of Quantitative Criminology*, 21(4), 365–390. <https://doi.org/10.1007/s10940-005-7354-7>
- Bifulco, R. (2012). Can nonexperimental estimates replicate estimates based on random assignment in evaluations of school choice? A within-study comparison. *Journal of Policy Analysis and Management*, 31(3), 729–751.
- Blakey, E., Matthews, D., Cragg, L., Buck, J., Cameron, D., Higgins, B., Pepper, L., Ridley, E., Sullivan, E., & Carroll, D. J. (2020). The role of executive functions in socioeconomic attainment gaps: Results from a randomized controlled trial. *Child Development, Advance online publication*. <https://doi.org/10.1111/cdev.13358>
- Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin*, 117(2), 187–215. <https://doi.org/10.1037//0033-2909.117.2.187>
- Bourdieu, P., & Passeron, J. C. (1990). *Reproduction in education, society, and culture* (R. Nice, Trans.; 2nd Ed). Sage.
- Bradby, D., Pedroso, R., & Rogers, A. (2007). *Secondary School Course Classification System: School Codes for the Exchange of Data (SCED)* (NCES 2007-341). National Center for Education Statistics. <https://nces.ed.gov/pubs2007/2007341.pdf>
- Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, 24(1), 61–100. <https://doi.org/10.3102/0091732X024001061>
- Brenner, B. (2010). Reflecting on the rationales for string study in schools. *Philosophy of Music Education Review*, 18(1), 45–64.

- Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163(12), 1149–1156. <https://doi.org/10.1093/aje/kwj149>
- Brown, N. C. M. (2001). The meaning of transfer in the practices of arts education. *Studies in Art Education*, 43(1), 83–102. <https://doi.org/10.2307/1320994>
- Brumback, B. A., Hernán, M. A., Haneuse, S. J. P. A., & Robins, J. M. (2004). Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in Medicine*, 23(5), 749–767. <https://doi.org/10.1002/sim.1657>
- Bruner, J. S. (1977). *The process of education*. Harvard University Press.
- Bugaj, K., & Brenner, B. (2011). The effects of music instruction on cognitive development and reading skills: An overview. *Bulletin of the Council for Research in Music Education*, 189, 89–104. <https://doi.org/10.5406/bulcouresmusedu.189.0089>
- Burgette, L. F., McCaffrey, D. F., & Griffin, B. A. (2015). Propensity score estimation with boosted regression. In W. Pan & H. Bai (Eds.), *Propensity score analysis: Fundamentals and developments* (pp. 49–73). Guilford Press.
- Burton, J. M., Horowitz, R., & Abeles, H. (1999). Learning in and through the arts: Curriculum implications. In E. B. Fiske (Ed.), *Champions of change: The impact of the arts on learning* (pp. 35–46). President’s Committee on the Arts and the Humanities.
- Burton, J. M., Horowitz, R., & Abeles, H. (2000). Learning in and through the arts: The question of transfer. *Studies in Art Education*, 41(3), 228–257.
- Butzlaff, R. (2000). Can music be used to teach reading? *Journal of Aesthetic Education*, 34(3/4), 167–178. <https://doi.org/10.2307/3333642>

- Cabanac, A., Perlovsky, L., Bonniot-Cabanac, M.-C., & Cabanac, M. (2013). Music and academic performance. *Behavioural Brain Research*, 256, 257–260.
<https://doi.org/10.1016/j.bbr.2013.08.023>
- Catterall, J. S. (1997). *Involvement in the arts and success in secondary school*. Americans for the Arts.
- Catterall, J. S. (1999). The arts and human development: The status of research. In M. McCarthy (Ed.), *Enlightened advocacy: Implications of research for arts education policy and practice* (pp. 31–42). University of Maryland.
- Catterall, J. S. (2002). The arts and the transfer of learning. In R. J. Deasy (Ed.), *Critical links: Learning in the arts and student academic and social development* (pp. 151–157). Arts Education Partnership.
- Catterall, J. S. (2009). *Doing well and doing good by doing art: A 12-year national study of education in the visual and performing arts*. Imagination Group/I-Group Books.
- Catterall, J. S., Chapleau, R., & Iwanaga, J. (1999). Involvement in the arts and human development: General involvement and intensive involvement in music and theater arts. In E. B. Fiske (Ed.), *Champions of change: The impact of the arts on learning* (pp. 1–18). President's Committee on the Arts and the Humanities.
- Catterall, J. S., Dumais, S. A., & Hampden-Thompson, G. (2012). *The arts and achievement in at-risk youth: Findings from four longitudinal studies* (Research Report #55). National Endowment for the Arts. <https://www.arts.gov/sites/default/files/Arts-At-Risk-Youth.pdf>
- Center for Academic, Social, and Emotional Learning. (2017). *Social and emotional learning competencies*. Author. <https://casel.org/what-is-sel/>
- Chabris, C. F. (1999). Prelude or requiem for the “Mozart effect”? *Nature*, 400, 826–827.

- Chapman, L. H. (1982). *Instant art, instant culture: The unspoken policy for American schools*. Teachers College Press.
- Chapman, L. H. (2001). Can the arts win hearts and minds? *Arts Education Policy Review*, 102(5), 21–23.
- Chen, C.-S., Sable, J., Mitchell, L., & Liu, F. (2012). *Documentation to the NCES Common Core of Data public elementary/secondary school universe survey: School year 2009–10* (NCES 2011-348rev). National Center for Education Statistics, US Department of Education. <http://nces.ed.gov/pubsearch/pubs.info.asp?pubid=2011348rev>
- Chorus America. (2009). *The chorus impact study: How children, adults, and communities benefit from choruses*. Author.
http://www.chorusamerica.org/sites/default/files/resources/ImpactStudy09_Report.pdf
- Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24, 295–313.
- Cohen, R. I. (2017, February 14). *Top 10 reasons to support the arts in 2017*. ARTS Blog.
<http://blog.americansforthearts.org/2017/02/14/top-10-reasons-to-support-the-arts-in-2017>
- Cole, S. R., & Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6), 656–664.
<https://doi.org/10.1093/aje/kwn164>
- Cole, S. R., Hernán, M. A., Margolick, J. B., Cohen, M. H., & Robins, J. M. (2005). Marginal structural models for estimating the effect of highly active antiretroviral therapy initiation on CD4 cell count. *American Journal of Epidemiology*, 162(5), 471–478.
<https://doi.org/10.1093/aje/kwi216>

- Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis*, 24(3), 175–199.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Houghton Mifflin.
- Cook, T. D., & Steiner, P. M. (2010). Case matching and the reduction of selection bias in quasi-experiments: The relative importance of pretest measures of outcome, of unreliable measurement, and of mode of data analysis. *Psychological Methods*, 15(1), 56–68.
<https://doi.org/10.1037/a0018536>
- Cook, T. D., Steiner, P. M., & Pohl, S. (2009). How bias reduction is affected by covariate choice, unreliability, and mode of data analysis: Results from two types of within-study comparisons. *Multivariate Behavioral Research*, 44(6), 828–847.
<https://doi.org/10.1080/00273170903333673>
- Corrigall, K. A., & Schellenberg, E. G. (2015). Predicting who takes music lessons: Parent and child characteristics. *Frontiers in Psychology*, 6, 1–8.
<https://doi.org/10.3389/fpsyg.2015.00282>
- Corrigall, K. A., Schellenberg, E. G., & Misura, N. M. (2013). Music training, cognition, and personality. *Frontiers in Psychology*, 4(222), 1–10.
<https://doi.org/10.3389/fpsyg.2013.00222>
- Costa-Giomi, E. (1999). The effects of three years of piano instruction on children's cognitive development. *Journal of Research in Music Education*, 47(3), 198–212.
<https://doi.org/10.2307/3345779>

- Costa-Giomi, E. (2004). Effects of three years of piano instruction on children's academic achievement, school performance and self-esteem. *Psychology of Music*, 32(2), 139–152. <https://doi.org/10.1177/0305735604041491>
- Costa-Giomi, E. (2008). Characteristics of elementary music programs in urban schools: What money can buy. *Bulletin of the Council for Research in Music Education*, 177, 19–28.
- Costa-Giomi, E. (2012). Music instruction and children's intellectual development: The educational context of music participation. In R. MacDonald, G. Kreutz, & L. Mitchell (Eds.), *Music, health, and wellbeing* (pp. 339–356). Oxford University Press.
- Costa-Giomi, E. (2015). The long-term effects of childhood music instruction on intelligence and general cognitive abilities. *Update: Applications of Research in Music Education*, 33(2), 20–26. <https://doi.org/10.1177/8755123314540661>
- Costa-Giomi, E., & Chappell, E. (2007). Characteristics of band programs in a large urban school district: Diversity or inequality? *Journal of Band Research*, 42(2), 1–18.
- Cury, F., Elliot, A. J., Da Fonseca, D., & Moller, A. C. (2006). The social-cognitive model of achievement motivation and the 2×2 achievement goal framework. *Journal of Personality and Social Psychology*, 90(4), 666–679. <https://doi.org/10.1037/0022-3514.90.4.666>
- Dagaz, M. C. (2012). Learning from the band: Trust, acceptance, and self-confidence. *Journal of Contemporary Ethnography*, 41(4), 432–461. <https://doi.org/10.1177/0891241612447813>
- Davidson, J. W., Howe, M. J. A., Moore, D. G., & Sloboda, J. A. (1996). The role of parental influences in the development of musical performance. *British Journal of Developmental Psychology*, 14(4), 399–412. <https://doi.org/10.1111/j.2044-835X.1996.tb00714.x>

- Deci, E. L., Koestner, R., & Ryan, R. M. (2001). Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of Educational Research*, 71(1), 1–27.
<https://doi.org/10.3102/00346543071001001>
- Degé, F., Kubicek, C., & Schwarzer, G. (2011). Music lessons and intelligence: A relation mediated by executive functions. *Music Perception*, 29(2), 195–201.
<https://doi.org/10.1525/mp.2011.29.2.195>
- Degé, F., & Schwarzer, G. (2018). The influence of an extended music curriculum at school on academic self-concept in 9- to 11-year-old children. *Musicae Scientiae*, 22(3), 305–321.
<https://doi.org/10.1177/1029864916688508>
- Degé, F., Wehrum, S., Stark, R., & Schwarzer, G. (2014). Music lessons and academic self-concept in 12- to 14-year-old children. *Musicae Scientiae*, 18(2), 203–215.
<https://doi.org/10.1177/1029864914523283>
- Demorest, S. M., & Morrison, S. J. (2000). Does music make you smarter? *Music Educators Journal*, 87(2), 33–58.
- Detterman, D. K. (1993). The case for the prosecution: Transfer as an epiphenomenon. In D. K. Detterman & R. J. Sternberg (Eds.), *Transfer on trial: Intelligence, cognition, and instruction*. (pp. 99–167). Ablex Publishing.
- Devroop, K. (2012). The social-emotional impact of instrumental music performance on economically disadvantaged South African students. *Music Education Research*, 14(4), 407–416. <https://doi.org/10.1080/14613808.2012.685456>
- Dewey, J. (1934). *Art as experience*. Capricorn Books.

- Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics & Statistics*, 95(3), 932–945. https://doi.org/10.1162/REST_a_00318
- Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, 41(1), 417. <https://doi.org/10.1146/annurev.ps.41.020190.002221>
- 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. <https://doi.org/10.1111/1475-6773.12090>
- Duprey, M. A., Pratt, D. J., Jewell, D. M., Cominole, M. B., Fritch, L. B., Ritchie, E. A., Rogers, J. E., Wescott, J. D., & Wilson, D. H. (2018). *High School Longitudinal Study of 2009 (HSLS:09) Base-year to second follow-up* (NCES 2018-140). National Center for Education Statistics. <https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2018140>
- Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, 82(1), 405–432. <https://doi.org/10.1111/j.1467-8624.2010.01564.x>
- Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist*, 41(10), 1040–1048. <https://doi.org/10.1037/0003-066X.41.10.1040>
- Dweck, C. S. (2002). Messages that motivate: How praise molds students' beliefs, motivation, and performance (in surprising ways). In J. Aronson (Ed.), *Improving academic achievement: Impact of psychological factors on education* (pp. 37–60). Academic Press.
- Dweck, C. S. (2006). *Mindset: The new psychology of success*. Random House.

- Dweck, C. S., Davidson, W., Nelson, S., & Enna, B. (1978). Sex differences in learned helplessness. *Developmental Psychology*, 14(3), 268–276. <https://doi.org/10.1037/0012-1649.14.3.268>
- Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, 95(2), 256–273.
- Eccles, J. (2009). Who am I and what am I going to do with my life? Personal and collective identities as motivators of action. *Educational Psychologist*, 44(2), 78–89. <https://doi.org/10.1080/00461520902832368>
- Edgar, S. N. (2017). *Music education and social emotional learning: The heart of teaching music*. GIA Publications.
- Eisner, E. W. (1998). What do the arts teach? *Improving Schools*, 1(3), 32–36. <https://doi.org/10.1177/136548029800100313>
- Eisner, E. W. (1999a). Does experience in the arts boost academic achievement? *Clearing House*, 72(3), 143–149.
- Eisner, E. W. (1999b). What justifies arts education: What research does not say. In M. McCarthy (Ed.), *Enlightened advocacy: Implications of research for arts education policy and practice* (pp. 17–30). University of Maryland.
- Eisner, E. W. (2002). *Arts and the creation of mind*. Yale University Press.
- Elliot, A. J., & McGregor, H. A. (2001). A 2x2 achievement goal framework. *Journal of Personality and Social Psychology*, 80(3), 501–519. <https://doi.org/10.1037/0022-3514.80.3.501>
- Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.

- Elpus, K. (2011). *Justifying music education: Econometric analyses of issues in music education policy* (AAT 3456550) [Doctoral dissertation, Northwestern University]. ProQuest Dissertations and Theses.
- Elpus, K. (2013a). *Arts education and positive youth development: Cognitive, behavioral, and social outcomes of adolescents who study the arts*. National Endowment for the Arts. <https://www.arts.gov/sites/default/files/Research-Art-Works-Maryland.pdf>
- Elpus, K. (2013b). Is it the music or is it selection bias? A nationwide analysis of music and non-music students' SAT scores. *Journal of Research in Music Education*, 61(2), 175–194. <https://doi.org/10.1177/0022429413485601>
- Elpus, K., & Abril, C. R. (2011). High school music ensemble students in the United States: A demographic profile. *Journal of Research in Music Education*, 59(2), 128–145. <https://doi.org/10.1177/0022429411405207>
- Elpus, K., & Abril, C. R. (2019). Who enrolls in high school music? A national profile of U.S. students, 2009–2013. *Journal of Research in Music Education*, 002242941986283. <https://doi.org/10.1177/0022429419862837>
- Enders, C. K., Keller, B. T., & Levy, R. (2018). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods*, 23(2), 298–317. <https://doi.org/10.1037/met0000148>
- Engle, R. A., Lam, D. P., Meyer, X. S., & Nix, S. E. (2012). How does expansive framing promote transfer? Several proposed explanations and a research agenda for investigating them. *Educational Psychologist*, 47(3), 215–231. <https://doi.org/10.1080/00461520.2012.695678>

- Evans, P. (2015). Self-determination theory: An approach to motivation in music education. *Musicae Scientiae*, 19(1), 65–83. <https://doi.org/10.1177/1029864914568044>
- Figlio, D. N., & Ladd, H. F. (2015). School accountability and student achievement. In H. F. Ladd & M. E. Goertz (Eds.), *Handbook of research in education finance and policy* (2nd Ed., pp. 194–210). Routledge.
- Fiske, E. B. (1999). *Champions of change: The impact of the arts on learning*. President's Committee on the Arts and the Humanities.
- Fitzpatrick, K. R. (2006). The effect of instrumental music participation and socioeconomic status on Ohio fourth-, sixth-, and ninth-grade proficiency test performance. *Journal of Research in Music Education*, 54(1), 73–84. <https://doi.org/10.2307/3653456>
- Fitzpatrick, K. R. (2011). A mixed methods portrait of urban instrumental music teaching. *Journal of Research in Music Education*, 59(3), 229–256. <https://doi.org/10.1177/0022429411414912>
- Foster, E. M., & Jenkins, J. V. (2017). Does participation in music and performing arts influence child development? *American Educational Research Journal*, 54(3), 399–443. <https://doi.org/10.3102/0002831217701830>
- Furrer, C., & Skinner, E. (2003). Sense of relatedness as a factor in children's academic engagement and performance. *Journal of Educational Psychology*, 95(1), 148–162. <https://doi.org/10.1037/0022-0663.95.1.148>
- Gardner, H. (1990). *Art education and human development*. The Getty Education Institute for the Arts.

- Gehlbach, H. (2010). The social side of school: Why teachers need social psychology. *Educational Psychology Review*, 22(3), 349–362. <https://doi.org/10.1007/s10648-010-9138-3>
- Gillen-O’Neel, C., & Fuligni, A. (2013). A longitudinal study of school belonging and academic motivation across high school. *Child Development*, 84(2), 678–692. <https://doi.org/10.1111/j.1467-8624.2012.01862.x>
- Gillespie, R. (1992). The elementary pull-out crisis: Using research effectively. *American String Teacher*, 42(2), 79–81. <https://doi.org/10.1177/000313139204200225>
- Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment*. Falmer Press.
- Gipps, C. V. (1995). What do we mean by equity in relation to assessment? *Assessment in Education: Principles, Policy & Practice*, 2(3), 271–281. <https://doi.org/10.1080/0969595950020303>
- Gipps, C. V. (1999). Socio-cultural aspects of assessment. *Review of Research in Education*, 24(1), 355–392. <https://doi.org/10.3102/0091732X024001355>
- Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, 48(1), 26–34. <https://doi.org/10.1037//0003-066X.48.1.26>
- Goodenow, C. (1992). Strengthening the links between educational psychology and the study of social contexts. *Educational Psychologist*, 27(2), 177–196. https://doi.org/10.1207/s15326985ep2702_4
- Goodenow, C. (1993). Classroom belonging among early adolescent students: Relationships to motivation and achievement. *The Journal of Early Adolescence*, 13(1), 21–43.

- Gouzouasis, P., Guhn, M., & Kishor, N. (2007). The predictive relationship between achievement and participation in music and achievement in core grade 12 academic subjects. *Music Education Research*, 9(1), 81–92.
<https://doi.org/10.1080/14613800601127569>
- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60, 549–576. <https://doi.org/10.1146/annurev.psych.58.110405.085530>
- Graham, S. E., & Kurlaender, M. (2011). Using propensity scores in educational research: General principles and practical applications. *Journal of Educational Research*, 104(5), 340–353. <https://doi.org/10.1080/00220671.2010.486082>
- Greenberg, M. (1970). Musical achievement and the self-concept. *Journal of Research in Music Education*, 18(1), 57–64. <https://doi.org/10.2307/3344358>
- Greenberg, M. T., Weissberg, R. P., O'Brien, M. U., Zins, J. E., Fredericks, L., Resnik, H., & Elias, M. J. (2003). Enhancing school-based prevention and youth development through coordinated social, emotional, and academic learning. *The American Psychologist*, 58(6–7). <https://doi.org/10.1037/0003-066X.58.6-7.466>
- Greenland, S., & Robins, J. M. (2009). Identifiability, exchangeability and confounding revisited. *Epidemiologic Perspectives & Innovations*, 6(4), 1–9.
<https://doi.org/10.1186/1742-5573-6-4>
- Groff, F. H. (1963). *Effect of academic achievement of excusing elementary school pupils from classes to study instrumental music* [Doctoral dissertation]. University of Connecticut.
- Guhn, M., Emerson, S. D., & Gouzouasis, P. (2020). A population-level analysis of associations between school music participation and academic achievement. *Journal of Educational Psychology*, 112(2), 308–328. <https://doi.org/10.1037/edu0000376>

- Guo, S., & Fraser, M. W. (2015). *Propensity score analysis: Statistical methods and applications* (2nd ed.). Sage.
- Hahs-Vaughn, D. L. (2015). Propensity score analysis with complex survey samples. In W. Pan & H. Bai (Eds.), *Propensity score analysis: Fundamentals and developments* (pp. 236–264). Guilford Press.
- Hahs-Vaughn, D. L., & Onwuegbuzie, A. J. (2006). Estimating and using propensity score analysis with complex samples. *The Journal of Experimental Education*, 75(1), 31–65. JSTOR.
- Hallam, S. (2006). *Music psychology in education*. Institute of Education, University of London.
- Hallam, S. (2010). The power of music: Its impact on the intellectual, social and personal development of children and young people. *International Journal of Music Education*, 28(3), 269–289. <https://doi.org/10.1177/0255761410370658>
- Hallberg, K., Cook, T. D., Steiner, P. M., & Clark, M. H. (2018). Pretest measures of the study outcome and the elimination of selection bias: Evidence from three within study comparisons. *Prevention Science*, 19, 274–283. <https://doi.org/10.1007/s11121-016-0732-6>
- Hanshumaker, J. (1980). The effects of arts education on intellectual and social development: A review of selected research. *Bulletin of the Council for Research in Music Education*, 61, 10–28.
- Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, 15(3), 234–249. <https://doi.org/10.1037/a0019623>

- Harris Interactive. (2006). *Understanding the linkages between music education and educational outcomes*. Harris Interactive.
<http://advocacy.nafme.org/files/2012/07/harrisprincipalsreportjuly2006.pdf>
- Harwell, M., & LeBeau, B. (2010). Student eligibility for a free lunch as an SES measure in education research. *Educational Researcher*, 39(2), 120–131.
<https://doi.org/10.3102/0013189X1036>
- Hash, P. M. (2004). Literature review: Pullout lessons in instrumental music education. *Bulletin of the Council for Research in Music Education*, 159, 1–10.
- Hash, P. M. (2011). Effect of pullout lessons on the academic achievement of eighth-grade band students. *Update: Applications of Research in Music Education*, 30(1), 16–22.
<https://doi.org/10.1177/8755123311418482>
- Haviland, A., Nagin, D. S., & Rosenbaum, P. R. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods*, 12(3), 247–267. <https://doi.org/10.1037/1082-989X.12.3.247>
- Heckman, J. J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, 66(5), 1017–1098. <https://doi.org/10.2307/2999630>
- Heckman, J. J., Ichimura, H., & Todd, P. (1998). Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65(2), 261–294.
- Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2), 107–128.
<https://doi.org/10.3102/10769986006002107>
- Heeringa, S. G., West, B. T., & Berglund, P. A. (2017). *Applied survey data analysis* (2nd Ed). Chapman and Hall/CRC.

- Helmrich, B. H. (2010). Window of opportunity? Adolescence, music, and algebra. *Journal of Adolescent Research*, 25(4), 557–577. <https://doi.org/10.1177/0743558410366594>
- Hendricks, K. S. (2014). Changes in self-efficacy beliefs over time: Contextual influences of gender, rank-based placement, and social support in a competitive orchestra environment. *Psychology of Music*, 42(3), 347–365. <https://doi.org/10.1177/0305735612471238>
- Hernán, M. A., Hernández-Díaz, S., & Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, 15(5), 615–625.
- Hetland, L., & Winner, E. (2001). The arts and academic achievement: What the evidence shows. *Arts Education Policy Review*, 102(5), 3.
- Hetland, L., Winner, E., Veenema, S., & Sheridan, K. M. (2013). *Studio thinking 2: The real benefits of visual arts education*. Teachers College Press.
- Hille, A., & Schupp, J. (2015). How learning a musical instrument affects the development of skills. *Economics of Education Review*, 44, 56–82.
<https://doi.org/10.1016/j.econedurev.2014.10.007>
- Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, 2(3), 259–278.
<https://doi.org/10.1023/A:1020371312283>
- Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4), 1161–1189.
<https://doi.org/10.1111/1468-0262.00442>

- Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3), 199–236. <https://doi.org/10.1093/pan/mpi013>
- Hogrebe, N., & Strietholt, R. (2016). Does non-participation in preschool affect children's reading achievement? International evidence from propensity score analyses. *Large-Scale Assessments in Education*, 4(1), 1–22. <https://doi.org/10.1186/s40536-016-0017-3>
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945–960. JSTOR. <https://doi.org/10.2307/2289064>
- Holochwost, S. J., Wolf, D. P., Fisher, K. R., & O'Grady, K. (2016). *The socioemotional benefits of the arts: A new mandate for arts education*. William Penn Foundation. http://williampennfoundation.org/sites/default/files/reports/Socioemotional%20Benefits%20of%20the%20Arts_Summary.pdf
- Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663–685. <https://doi.org/10.2307/2280784>
- Howe, M. J. A., Davidson, J. W., & Sloboda, J. A. (1998). Innate talents: Reality or myth? *Behavioral and Brain Sciences*, 21(3), 399–442. <https://doi.org/10.1017/S0140525X9800123X>
- Huang, C. (2011). Self-concept and academic achievement: A meta-analysis of longitudinal relations. *Journal of School Psychology*, 49(5), 505–528. <https://doi.org/10.1016/j.jsp.2011.07.001>
- Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86(1), 4–29.

- Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1), 5–86.
- Ingels, S. J., Pratt, D. J., Herget, D. R., Bryan, M., Fritch, L. B., Ottem, R., Rogers, J. E., & Wilson, D. (2015). *High School Longitudinal Study of 2009 (HSL:09) 2013 update and high school transcript* (NCES 2015-036). Institute of Education Sciences, U.S. Department of Education. <https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2015036>
- Ingels, S. J., Pratt, D. J., Herget, D. R., Burns, L. J., Dever, J. A., Ottem, R., Jin, Y., & Leinwand, S. (2011). *High School Longitudinal Study of 2009 (HSL:09): Base-year data file documentation* (NCES 2011-328). National Center for Education Statistics, US Department of Education. <https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2011328>
- Ingels, S. J., Pratt, D. J., Herget, D. R., Dever, J. A., Fritch, L. B., Ottem, R., Rogers, J. E., Kitmitto, S., & Leinwand, S. (2013). *High School Longitudinal Study of 2009 (HSL:09) Base year to first follow-up data file documentation* (NCES 2014-361; p. 190). National Center for Education Statistics, US Department of Education. <https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2014361>
- Jacobi, B. S. (2012). Opportunities for socioemotional learning in music classrooms. *Music Educators Journal*, 99(2), 68–74. <https://doi.org/10.1177/0027432112459082>
- Jenkins, J. S. (2001). The Mozart effect. *Journal of the Royal Society of Medicine*, 94(4), 170–172.
- Johnson, C. M., & Memmott, J. E. (2006). Examination of relationships between participation in school music programs of differing quality and standardized test results. *Journal of Research in Music Education*, 54(4), 293–307. <https://doi.org/10.1177/002242940605400403>

- Jorgensen, E. R. (1994). Justifying music instruction in American public schools: An historical perspective. *Bulletin of the Council for Research in Music Education*, 120, 17–31.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4), 523–539. <https://doi.org/10.1214/07-STS227>
- Kaschub, M. (2002). Defining emotional intelligence in music education. *Arts Education Policy Review*, 103(5), 9–15.
- Kelcey, B. (2011). Covariate selection in propensity scores using outcome proxies. *Multivariate Behavioral Research*, 46(3), 453–476. <https://doi.org/10.1080/00273171.2011.570164>
- King, G., & Nielsen, R. (2018). *Why propensity scores should not be used for matching [working paper]*. <https://gking.harvard.edu/files/gking/files/psnot.pdf>
- Kinney, D. W. (2008). Selected demographic variables, school music participation, and achievement test scores of urban middle school students. *Journal of Research in Music Education*, 56(2), 145–161. <https://doi.org/10.1177/0022429408322530>
- Kinney, D. W. (2010). Selected nonmusic predictors of urban students' decisions to enroll and persist in middle school band programs. *Journal of Research in Music Education*, 57(4), 334–350. <https://doi.org/10.1177/0022429409350086>
- Kinney, D. W. (2018). Selected nonmusic predictors of urban students' decisions to enroll and persist in middle and high school music ensemble electives. *Journal of Research in Music Education*, advance online publication. <https://doi.org/10.1177/0022429418809972>

- Kinney, D. W., & Forsythe, J. L. (2005). The effects of the Arts IMPACT curriculum upon student performance on the Ohio fourth-grade proficiency test. *Bulletin of the Council for Research in Music Education*, 164, 35–48.
- Klein, D. (2016, June 10). *Marginal effects in multiply imputed datasets*. 14th German Stata Users Group Meeting, Cologne, Germany.
https://www.stata.com/meeting/germany16/slides/de16_klein.pdf
- Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences*. American Psychological Association.
- Klinedinst, R. E. (1991). Predicting performance achievement and retention of fifth-grade instrumental students. *Journal of Research in Music Education*, 39(3), 225–238.
<https://doi.org/10.2307/3344722>
- Klinedinst, R. E. (1992). Predicting performance achievement of beginning band students: Second year results. *Contributions to Music Education*, 19, 46–59.
- Kohn, A. (2012, September 19). Schooling beyond measure. *Education Week*, 32(4), 36.
- Kohn, A. (2013). The roots of grades-and-tests. In J. Bower & P. L. Thomas (Eds.), *De-testing and de-grading schools: Authentic alternatives to accountability and standardization* (pp. 1–6). Peter Lang.
- Kohn, A. (2015, September 30). Do this and you'll get that: A bad way to defend good programs. *Education Week*. <https://www.edweek.org/ew/articles/2015/09/30/do-this-and-youll-get-that-a.html>
- Kokotsaki, D., & Hallam, S. (2007). Higher education music students' perceptions of the benefits of participative music making. *Music Education Research*, 9(1), 93–109.
<https://doi.org/10.1080/14613800601127577>

- Koza, J. E. (2006). "Save the music"? Toward culturally relevant, joyful, and sustainable school music. *Philosophy of Music Education Review*, 14(1), 23–38.
- Kraus, N., & Strait, D. L. (2015). Emergence of biological markers of musicianship with school-based music instruction. *Annals of the New York Academy of Sciences*, 1337(1), 163–169. <https://doi.org/10.1111/nyas.12631>
- Kvet, E. J. (1985). Excusing elementary school students from regular classroom activities for the study of instrumental music: The effect on sixth-grade reading, language, and mathematics achievement. *Journal of Research in Music Education*, 33(1), 45–54. <https://doi.org/10.2307/3344757>
- Lazarsfeld, P. F. (1959). Problems in methodology. In R. K. Merton, L. Broom, & L. S. Cottrell (Eds.), *Sociology today: Problems and prospects* (Vol. 1, pp. 39–72). Basic Books.
- Leary, M. R. (1999). Making sense of self-esteem. *Current Directions in Psychological Science*, 8(1), 32–35.
- Leary, M. R., Haupt, A. L., Strausser, K. S., & Chokel, J. T. (1998). Calibrating the sociometer: The relationship between interpersonal appraisals and state self-esteem. *Journal of Personality and Social Psychology*, 74(5), 1290–1299. <https://doi.org/10.1037/0022-3514.74.5.1290>
- Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PLoS ONE*, 6(3), e18174. <https://doi.org/10.1371/journal.pone.0018174>
- Lee, R. M., & Robbins, S. B. (1995). Measuring belongingness: The social connectedness and the social assurance scales. *Journal of Counseling Psychology*, 42(2), 232–241. <https://doi.org/10.1037/0022-0167.42.2.232>

- Leite, W. L., Aydin, B., & Gurel, S. (2018). A comparison of propensity score weighting methods for evaluating the effects of programs with multiple versions. *The Journal of Experimental Education*, 0(0), 1–14. <https://doi.org/10.1080/00220973.2017.1409179>
- Lenis, D., Ackerman, B., & Stuart, E. A. (2018). Measuring model misspecification: Application to propensity score methods with complex survey data. *Computational Statistics & Data Analysis*, 128, 48–57. <https://doi.org/10.1016/j.csda.2018.05.003>
- Lenis, D., Nguyen, T. Q., Dong, N., & Stuart, E. A. (2017). It's all about balance: Propensity score matching in the context of complex survey data. *Biostatistics*, Advance online publication. <https://doi.org/10.1093/biostatistics/kxx063>
- Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1), 14–23. <https://doi.org/10.1002/widm.8>
- Lohr, S. L. (1999). *Sampling: Design and analysis*. Duxbury Press.
- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Sage.
- Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity scores: An introduction and experimental test. *Evaluation Review*, 29(6), 530–558. <https://doi.org/10.1177/0193841X05275596>
- Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23(19), 2937–2960. <https://doi.org/10.1002/sim.1903>
- MacDonald, R. A. (2002). What are musical identities, and why are they important? In R. A. MacDonald, D. J. Hargreaves, & D. Miell (Eds.), *Musical identities*. Oxford University Press.

- MacIntyre, P. D., Potter, G. K., & Burns, J. N. (2012). The socio-educational model of music motivation. *Journal of Research in Music Education*, 60(2), 129–144.
<https://doi.org/10.1177/0022429412444609>
- Markus, H., & Nurius, P. (1986). Possible selves. *American Psychologist*, 41(9), 954–969.
<https://doi.org/10.1037/0003-066X.41.9.954>
- Marsh, H. W., Pekrun, R., Murayama, K., Arens, A. K., Parker, P. D., Guo, J., & Dicke, T. (2018). An integrated model of academic self-concept development: Academic self-concept, grades, test scores, and tracking over 6 years. *Developmental Psychology*, 54(2), 263–280. <https://doi.org/10.1037/dev0000393>
- McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4), 403–425. <https://doi.org/10.1037/1082-989X.9.4.403>
- McCarthy, K. F. (2004). *Gifts of the muse: Reframing the debate about the benefits of the arts*. RAND Research in the Arts.
- McCormick, J., & McPherson, G. E. (2007). Expectancy-value motivation in the context of a music performance examination. *Musicae Scientiae*, 11(2), 37–52.
- McCrae, R. R., & John, O. P. (1992). An Introduction to the five-factor model and its applications. *Journal of Personality*, 60(2), 175–215. <https://doi.org/10.1111/j.1467-6494.1992.tb00970.x>
- McPherson, G. E. (1995). Redefining the teaching of musical performance. *Quarterly Journal of Music Teaching and Learning*, 6, 56–61.

- McPherson, G. E. (2005). From child to musician: Skill development during the beginning stages of learning an instrument. *Psychology of Music*, 33(1), 5–35.
<https://doi.org/10.1177/0305735605048012>
- McPherson, G. E. (2009). The role of parents in children's musical development. *Psychology of Music*, 37(1), 91–110. <https://doi.org/10.1177/0305735607086049>
- McPherson, G. E., Davidson, J. W., & Faulkner, R. (2012). *Music in our lives: Rethinking musical ability, development and identity*. Oxford University Press.
- McPherson, G. E., & Hendricks, K. S. (2010). Students' motivation to study music: The United States of America. *Research Studies in Music Education*, 32(2), 201–213.
<https://doi.org/10.1177/1321103X10384200>
- Michel, D. E. (1971). Self esteem and academic achievement in black junior high school students: Effects of automated guitar instruction. *Bulletin of the Council for Research in Music Education*, 24, 15–23.
- Michel, D. E., & Farrell, D. M. (1973). Music and self-esteem: Disadvantaged problem boys in an all-black elementary school. *Journal of Research in Music Education*, 21(1), 80–84.
<https://doi.org/10.2307/3343983>
- Miksza, P. (2007). Music participation and socioeconomic status as correlates of change: A longitudinal analysis of academic achievement. *Bulletin of the Council for Research in Music Education*, 172, 41–58.
- Miksza, P. (2009). Relationships among impulsivity, achievement goal motivation, and the music practice of high school wind players. *Bulletin of the Council for Research in Music Education*, 180, 9–27.

- Miksza, P. (2010). Investigating relationships between participation in high school music ensembles and extra-musical outcomes: An analysis of the Education Longitudinal Study of 2002 using a bioecological development model. *Bulletin of the Council for Research in Music Education*, 186, 7–25.
- Miksza, P. (2011). A review of research on practicing: Summary and synthesis of the extant research with implications for a new theoretical orientation. *Bulletin of the Council for Research in Music Education*, 190, 51–92. Academic OneFile.
- Miksza, P. (2013). Arts education advocacy: The relative effects of school-level influences on resources for arts education. *Arts Education Policy Review*, 114(1), 25–32.
<https://doi.org/10.1080/10632913.2013.744245>
- Miksza, P., & Gault, B. M. (2014). Classroom music experiences of U.S. elementary school children: An analysis of the Early Childhood Longitudinal Study of 1998–1999. *Journal of Research in Music Education*, 62(1), 4–17. <https://doi.org/10.1177/0022429413519822>
- Moreno, S., Bialystok, E., Barac, R., Schellenberg, E. G., Cepeda, N. J., & Chau, T. (2011). Short-term music training enhances verbal intelligence and executive function. *Psychological Science*, 22(11), 1425–1433. <https://doi.org/10.1177/0956797611416999>
- Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge University Press.
- Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Causal inference in educational and social science research*. Oxford University Press.
- National Association for Music Education. (2017). *Advocacy resources*. NAFME.
<http://nafme.org/advocacy/advocacy-resources/>

- Newman, L. A., Madaus, J. W., & Javitz, H. S. (2016). Effect of transition planning on postsecondary support receipt by students with disabilities. *Exceptional Children*, 82(4), 497–514. <https://doi.org/10.1177/0014402915615884>
- O’Connell, A. A., & Gray, D. L. (2011). Cause and event: Supporting causal claims through logistic models. *Educational Psychology Review*, 23(2), 245–261. <https://doi.org/10.1007/s10648-011-9158-7>
- Osterman, K. F. (2000). Students’ need for belonging in the school community. *Review of Educational Research*, 70(3), 323–367.
- Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research*, 66(4), 543–578. <https://doi.org/10.2307/1170653>
- Pajares, F. (2003). Self-efficacy beliefs, motivation, and achievement in writing: A review of the literature. *Reading & Writing Quarterly*, 19(2), 139–158. <https://doi.org/10.1080/10573560308222>
- Pajares, F., & Miller, M. D. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal of Educational Psychology*, 86(2), 193–203. <https://doi.org/10.1037//0022-0663.86.2.193>
- Pajares, F., & Schunk, D. H. (2002). Self and self-belief in psychology and education: A historical perspective. In J. Aronson (Ed.), *Improving academic achievement: Impact of psychological factors on education* (pp. 5–21). Academic Press.
- Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4), 669–688. JSTOR. <https://doi.org/10.2307/2337329>
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

- Pitts, S. E. (2007). Anything goes: A case study of extra-curricular musical participation in an English secondary school. *Music Education Research*, 9(1), 145–165.
<https://doi.org/10.1080/14613800601127627>
- Pitts, S. E., Davidson, J. W., & McPherson, G. E. (2000). Models of success and failure in instrumental learning: Case studies of young players in the first 20 months of learning. *Bulletin of the Council for Research in Music Education*, 146, 51–69.
- Pont, G. (2004). Philosophy and science of music in ancient Greece: The predecessors of Pythagoras and their contribution. *Nexus Network Journal*, 6(1), 17–29.
<https://doi.org/10.1007/s00004-004-0003-x>
- Portowitz, A., Lichtenstein, O., Egorova, L., & Brand, E. (2009). Underlying mechanisms linking music education and cognitive modifiability. *Research Studies in Music Education*, 31(2), 107–128. <https://doi.org/10.1177/1321103X09344378>
- R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rabkin, N. (2004). Learning and the arts. In N. Rabkin & R. Redmond (Eds.), *Putting the arts in the picture: Reframing education in the 21st century* (pp. 5–15). Center for Arts Policy at Columbia College Chicago.
- Raghunathan, T., Berglund, P. A., & Solenberger, P. (2018). *Multiple imputation in practice: With examples using IVEware*. CRC Press.
<http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1855546>

- Raghunathan, T., Solenberger, P., Berglund, P. A., & van Hoewyk, J. (2016). *IVEware: Imputation and variance estimation software* (Version 0.3) [Computer software]. Survey Research Center, Institute for Social Research, University of Michigan.
<https://www.src.isr.umich.edu/software/>
- Rauscher, F. H. (2002). Mozart and the mind: Factual and fictional effects of musical enrichment. In J. Aronson (Ed.), *Improving academic achievement: Impact of psychological factors on education* (pp. 267–278). Academic Press.
- Rauscher, F. H., & Hinton, S. C. (2011). Music instruction and its diverse extra-musical benefits. *Music Perception: An Interdisciplinary Journal*, 29(2), 215–226.
<https://doi.org/10.1525/mp.2011.29.2.215>
- Rauscher, F. H., Shaw, G. L., & Ky, K. N. (1993). Music and spatial task performance. *Nature*, 365, 611. <https://doi.org/10.1038/365611a0>
- Rauscher, F. H., Shaw, G. L., & Ky, K. N. (1995). Listening to Mozart enhances spatial-temporal reasoning: Towards a neurophysiological basis. *Neuroscience Letters*, 185(1), 44–47. [https://doi.org/10.1016/0304-3940\(94\)11221-4](https://doi.org/10.1016/0304-3940(94)11221-4)
- Regelski, T. A. (1998). The Aristotelian bases of praxis for music and music education as praxis. *Philosophy of Music Education Review*, 6(1), 22–59.
- Reimer, B. (1999). Facing the risks of the “Mozart effect.” *Music Educators Journal*, 86(1), 37–43. <https://doi.org/10.2307/3399576>
- Reimer, B. (2003). *A philosophy of music education: Advancing the vision* (3rd ed.). Prentice Hall.
- Reimer, B. (2005). The danger of music education advocacy. *International Journal of Music Education*, 23(2), 139–142. <https://doi.org/10.1177/0255761405052409>

- Rickard, N. S., Appelman, P., James, R., Murphy, F., Gill, A., & Bambrick, C. (2013). Orchestrating life skills: The effect of increased school-based music classes on children's social competence and self-esteem. *International Journal of Music Education*, 31(3), 292–309. <https://doi.org/10.1177/0255761411434824>
- Rickard, N. S., Bambrick, C. J., & Gill, A. (2012). Absence of widespread psychosocial and cognitive effects of school-based music instruction in 10–13-year-old students. *International Journal of Music Education*, 30(1), 57–78. <https://doi.org/10.1177/0255761411431399>
- Ridgeway, G. (2006). Assessing the effect of race bias in post-traffic stop outcomes using propensity scores. *Journal of Quantitative Criminology*, 22(1), 1–29. JSTOR. <https://doi.org/10.1007/s10940-005-9000-9>
- Ridgeway, G., Kovalchik, S. A., Griffin, B. A., & Kabeto, M. U. (2015). Propensity score analysis with survey weighted data. *Journal of Causal Inference*, 3(2), 237–249. <https://doi.org/10.1515/jci-2014-0039>
- Ridgeway, G., & McCaffrey, D. F. (2007). Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4), 540–543. <https://doi.org/10.1214/07-STS227C>
- Ridgeway, G., McCaffrey, D., Morral, A., Burgette, L., & Griffin, B. A. (2017). *Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the twang package*. RAND Corporation. <ftp://xyz.csail.mit.edu/pub/CRAN/web/packages/twang/vignettes/twang.pdf>
- Rigdon, J., Baiocchi, M., & Basu, S. (2018). Near-far matching in R: The nearfar package. *Journal of Statistical Software*, 86(Code Snippet 5). <https://doi.org/10.18637/jss.v086.c05>

- Robins, J. M. (1999). Association, causation, and marginal structural models. *Synthese*, 121(1/2), 151–179.
- Robins, J. M., Hernán, M. Á., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5), 550–560.
- Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429), 122–129. <https://doi.org/10.2307/2291135>
- Robins, J. M., Sued, M., Lei-Gomez, Q., & Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when “inverse probability” weights are highly variable. *Statistical Science*, 22(4), 544–559. <https://doi.org/10.1214/07-STS22>
- Robitaille, J. P., & O’Neal, S. (1981). Why instrumental music in elementary schools? *Phi Delta Kappan*, 63(3), 213.
- Roscigno, V. J., & Ainsworth-Darnell, J. W. (1999). Race, cultural capital, and educational resources: Persistent inequalities and achievement returns. *Sociology of Education*, 72(3), 158–178. <https://doi.org/10.2307/2673227>
- Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society. Series A (General)*, 147(5), 656–666. <https://doi.org/10.2307/2981697>
- Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398), 387–394. <https://doi.org/10.2307/2289440>
- Rosenbaum, P. R. (1993). Hodges-Lehmann point estimates of treatment effect in observational studies. *Journal of the American Statistical Association*, 88(424), 1250–1253. <https://doi.org/10.2307/2291264>

- Rosenbaum, P. R. (2010). *Design of observational studies*. Springer.
- Rosenbaum, P. R., & Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society*, 45(2), 212–218.
- Rosenbaum, P. R., & Rubin, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.
<https://doi.org/10.2307/2335942>
- Rosenbaum, P. R., & Rubin, D. B. (1985). The bias due to incomplete matching. *Biometrics*, 41(1), 103–116. <https://doi.org/10.2307/2530647>
- Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44(10), 1276–1284.
<https://doi.org/10.1037/0003-066X.44.10.1276>
- Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, 80, 1–28.
<https://doi.org/10.1037/h0092976>
- Royer, J. M., Mestre, J. P., & Dufresne, R. J. (2005). Introduction: Framing the transfer problem. In J. P. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. vii–xxvi). Information Age.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688–701.
<https://doi.org/10.1037/h0037350>
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592. JSTOR.
<https://doi.org/10.2307/2335739>

- Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1), 34–58. JSTOR.
- Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366), 318–328. <https://doi.org/10.2307/2286330>
- Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396), 961–962.
- Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127(8), 757–763. https://doi.org/10.7326/0003-4819-127-8_Part_2-199710151-00064
- Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3–4), 169–188. <https://doi.org/10.1023/A:1020363010465>
- Rubin, D. B. (2006). *Matched sampling for causal effects*. Cambridge University Press.
- Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, 26(1), 20–36. <https://doi.org/10.1002/sim.2739>
- Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3), 808–840.
- Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, 52(1), 249–264.

- Ryan, R. M., Stiller, J. D., & Lynch, J. H. (1994). Representations of relationships to teachers, parents, and friends as predictors of academic motivation and self-esteem. *Journal of Early Adolescence*, 14(2), 226–249.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147–177. <https://doi.org/10.1037//1082-989X.7.2.147>
- Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, 13(4), 279–313. <https://doi.org/10.1037/a0014268>
- Schellenberg, E. G. (2004). Music lessons enhance IQ. *Psychological Science*, 15(8), 511–514. <https://doi.org/10.1111/j.0956-7976.2004.00711.x>
- Schellenberg, E. G. (2006). Long-term positive associations between music lessons and IQ. *Journal of Educational Psychology*, 98(2), 457–468. <https://doi.org/10.1037/0022-0663.98.2.457>
- Schellenberg, E. G. (2011). Examining the association between music lessons and intelligence. *British Journal of Psychology*, 102(3), 283–302. <https://doi.org/10.1111/j.2044-8295.2010.02000.x>
- Schellenberg, E. G., & Mankarious, M. (2012). Music training and emotion comprehension in childhood. *Emotion*, 12(5), 887–891. <https://doi.org/10.1037/a0027971>
- Schellenberg, E. G., & Peretz, I. (2008). Music, language and cognition: Unresolved issues. *Trends in Cognitive Sciences*, 12(2), 45–46. <https://doi.org/10.1016/j.tics.2007.11.005>
- Schellenberg, E. G., & Weiss, M. W. (2013). Music and cognitive abilities. In D. Deutsch (Ed.), *The Psychology of Music* (3rd Ed, pp. 499–550). Academic Press. <https://doi.org/10.1016/B978-0-12-381460-9.00012-2>

- Schmidt, C. P. (2007). Intrinsic-mastery motivation in instrumental music: Extension of a higher order construct. *Bulletin of the Council for Research in Music Education*, 173, 7–23.
- Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs*. American Educational Research Association.
- Schneider, T. W., & Klotz, J. (2000). *The impact of music education and athletic participation on academic achievement* (No. ED448186).
<https://files.eric.ed.gov/fulltext/ED448186.pdf>
- Schoen-Nazzaro, M. B. (1978). Plato and Aristotle on the ends of music. *Laval Théologique et Philosophique Québec*, 34(3), 261–274.
- Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist*, 26(3/4), 207–231.
- Schwartz, D. L., Bransford, J. D., & Sears, D. (2005). Efficiency and innovation in transfer. In J. P. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 1–52). Information Age.
- Scott, L. (1992). Attention and perseverance behaviors of preschool children enrolled in Suzuki violin lessons and other activities. *Journal of Research in Music Education*, 40(3), 225–235. <https://doi.org/10.2307/3345684>
- Scripp, L. (2002). An overview of research on music and learning. In R. J. Deasy (Ed.), *Critical links: Learning in the arts and student academic and social development* (pp. 132–136). Arts Education Partnership.
- See, B. H., & Kokotsaki, D. (2016). Impact of arts education on children’s learning and wider outcomes. *Review of Education*, 4(3), 234–262. <https://doi.org/10.1002/rev3.3070>

- Shadish, W. R. (2010). Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings. *Psychological Methods*, 15(1), 3–17.
<https://doi.org/10.1037/a0015916>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth.
- Shadish, W. R., & Steiner, P. M. (2010). A primer on propensity score analysis. *Newborn and Infant Nursing Reviews*, 10(1), 19–26. <https://doi.org/10.1053/j.nainr.2009.12.010>
- Shaw, B. P. (2018). *Music assessment for better ensembles*. Oxford University Press.
- Shen, C., Li, X., Li, L., & Were, M. C. (2011). Sensitivity analysis for causal inference using inverse probability weighting. *Biometrical Journal*, 53(5), 822–837.
<https://doi.org/10.1002/bimj.201100042>
- Shields, C. (2001). Music education and mentoring as intervention for at-risk urban adolescents: Their self-perceptions, opinions, and attitudes. *Journal of Research in Music Education*, 49(3), 273–286. <https://doi.org/10.2307/3345712>
- Shin, J. (2011). An investigation of participation in weekly music workshops and its relationship to academic self-concept and self-esteem of middle school students in low-income communities. *Contributions to Music Education*, 38(2), 29–42.
- Sichivitsa, V. O. (2007). The influences of parents, teachers, peers and other factors on students' motivation in music. *Research Studies in Music Education*, 29(1), 55–68.
<https://doi.org/10.1177/1321103X07087568>
- Smith, J. A., & Todd, P. E. (2001). Reconciling conflicting evidence on the performance of propensity-score matching methods. *The American Economic Review*, 91(2), 112–118.

- Southgate, D. E., & Roscigno, V. J. (2009). The impact of music on childhood and adolescent achievement. *Social Science Quarterly*, 90(1), 4–21. <https://doi.org/10.1111/j.1540-6237.2009.00598.x>
- Stapleton, L. M., & Thomas, S. L. (2008). The use of national datasets for teaching and research. In A. A. O’Connell & D. B. McCoach, *Multilevel modeling of educational data* (pp. 11–58). Information Age Publishing.
- StataCorp. (2015). *Stata statistical software: Release 14*. StataCorp.
- Steiner, P. M., & Cook, D. (2013). Matching and propensity scores. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods* (Vol. 1, pp. 237–259). Oxford University Press.
- Steiner, P. M., Cook, T. D., & Shadish, W. R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, 36(2), 213–236. <https://doi.org/10.3102/1076998610375835>
- 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), 250–267. <https://doi.org/10.1037/a0018719>
- Stevenson, L. M., & Deasy, R. (2005). *Third space: When learning matters*. Arts Education Partnership.
- Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14(4), 323–348. <https://doi.org/10.1037/a0016973>

- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1), 1–21. <https://doi.org/10.1214/09-STS313>
- Taylor, R. D., Oberle, E., Durlak, J. A., & Weissberg, R. P. (2017). Promoting positive youth development through school-based social and emotional learning interventions: A meta-analysis of follow-up effects. *Child Development*, 88(4), 1156–1171. <https://doi.org/10.1111/cdev.12864>
- Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research*, 46(1), 90–118. <https://doi.org/10.1080/00273171.2011.540475>
- Thompson, W. F., Schellenberg, E. G., & Husain, G. (2001). Arousal, mood, and the Mozart effect. *Psychological Science*, 12(3), 248–251. <https://doi.org/10.1111/1467-9280.00345>
- Tierney, A., & Kraus, N. (2013). Music training for the development of reading skills. *Progress in Brain Research*, 207, 209–241. <https://doi.org/10.1016/B978-0-444-63327-9.00008-4>
- Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, 37(1), 1–2. <https://doi.org/10.1080/01973533.2015.1012991>
- Tuckman, B. W., & Monetti, D. M. (2010). *Educational psychology*. Wadsworth.
- Usher, E. L., & Pajares, F. (2008a). Self-efficacy for self-regulated learning a validation study. *Educational and Psychological Measurement*, 68(3), 443–463. <https://doi.org/10.1177/0013164407308475>
- Usher, E. L., & Pajares, F. (2008b). Sources of self-efficacy in school: Critical review of the literature and future directions. *Review of Educational Research*, 78(4), 751–796. <https://doi.org/10.3102/0034654308321456>

- Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist*, 39(2), 111–133. https://doi.org/10.1207/s15326985ep3902_3
- Varela, W., Abrami, P. C., & Uptis, R. (2016). Self-regulation and music learning: A systematic review. *Psychology of Music*, 44(1), 55–74. <https://doi.org/10.1177/0305735614554639>
- Vaughn, K. (2000). Music and mathematics: Modest support for the oft-claimed relationship. *Journal of Aesthetic Education*, 34(3/4), 149–166. <https://doi.org/10.2307/3333641>
- Vaughn, K., & Winner, E. (2000). SAT scores of students who study the arts: What we can and cannot conclude about the association. *Journal of Aesthetic Education*, 34(3–4), 77–89.
- Vitale, J. L. (2011). Music makes you smarter: A new paradigm for music education? Perceptions and perspectives from four groups of elementary education stakeholders. *Canadian Journal of Education*, 34(3), 317–343.
- Wagner, M. M., Newman, L. A., & Javitz, H. S. (2016). The benefits of high school career and technical education (CTE) for youth with learning disabilities. *Journal of Learning Disabilities*, 49(6), 658–670. <https://doi.org/10.1177/0022219415574774>
- Wagner, M. M., Newman, L. A., & Javitz, H. S. (2017). Vocational education course taking and post-high school employment of youth with emotional disturbances. *Career Development and Transition for Exceptional Individuals*, 40(3), 132–143. <https://doi.org/10.1177/2165143415626399>
- Walberg, H. J., Zins, J. E., & Weissberg, R. P. (2004). Recommendations and conclusions: Implications for practice, training, research, and policy. In J. E. Zins, R. P. Weissberg, M. C. Wang, & H. J. Walberg (Eds.), *Building academic success on social and emotional learning: What does the research say?* (pp. 209–218). Teachers College Press.

- Wallick, M. D. (1998). A comparison study of the Ohio proficiency test results between fourth-grade string pullout students and those of matched ability. *Journal of Research in Music Education*, 46(2), 239–247. <https://doi.org/10.2307/3345626>
- Wasserstein, R. L. (Ed.). (2016). ASA statement on statistical significance and p-values. *The American Statistician*, 70(2), 131–133. <https://doi.org/10.1080/00031305.2016.1154108>
- West, C., & Clauhs, M. (2015). Strengthening music programs while avoiding advocacy pitfalls. *Arts Education Policy Review*, 116(2), 57–62. <https://doi.org/10.1080/10632913.2015.1007831>
- West, S. G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D. C., Holtgrave, D., Szapocznik, J., Fishbein, M., Rapkin, B., Clatts, M., & Mullen, P. D. (2008). Alternatives to the randomized controlled trial. *American Journal of Public Health*, 98(8), 1359–1366. <https://doi.org/10.2105/AJPH.2007.124446>
- Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: Neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, 63(8), 826–833. <https://doi.org/10.1016/j.jclinepi.2009.11.020>
- Wetter, O. E., Koerner, F., & Schwaninger, A. (2009). Does musical training improve school performance? *Instructional Science*, 37(4), 365–374. <https://doi.org/10.1007/s11251-008-9052-y>
- What Works Clearinghouse. (2014). *Procedures and standards handbook, version 3.0*. Institute of Education Sciences. http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf

- Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, 25(1), 68–81.
<https://doi.org/10.1006/ceps.1999.1015>
- Wiggins, G. P., & McTighe, J. (2005). *Understanding by design* (2nd Ed.). ASCD.
- William, D. (2000). Review essay. *British Educational Research Journal*, 26(5), 661–663.
<https://doi.org/10.1080/01411920020024626>
- Willie, C. V. (2001). The contextual effects of socioeconomic status on student achievement test scores by race. *Urban Education*, 36(4), 461–478.
<https://doi.org/10.1177/0042085901364002>
- Winner, E., & Cooper, M. (2000). Mute those claims: No evidence (yet) for a causal link between arts study and academic achievement. *Journal of Aesthetic Education*, 34(3/4), 11–75.
- Winner, E., Goldstein, T. R., & Vincent-Lancrin, S. (2013). *Art for art's sake? The impact of arts education*. OECD Publishing.
- Winner, E., & Hetland, L. (2000). The arts in education: Evaluating the evidence for a causal link. *Journal of Aesthetic Education*, 34(3/4), 3–10.
- Wolters, C. A., Yu, S. L., & Pintrich, P. R. (1996). The relation between goal orientation and students' motivational beliefs and self-regulated learning. *Learning and Individual Differences*, 8(3), 211–238. [https://doi.org/10.1016/S1041-6080\(96\)90015-1](https://doi.org/10.1016/S1041-6080(96)90015-1)
- Wong, V. C., Valentine, J. C., & Miller-Bains, K. (2017). Empirical performance of covariates in education observational studies. *Journal of Research on Educational Effectiveness*, 10(1), 207–236. <https://doi.org/10.1080/19345747.2016.1164781>
- Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. MIT Press.

- Yang, P. (2015). The impact of music on educational attainment. *Journal of Cultural Economics*, 39(4), 369–396. <https://doi.org/10.1007/s10824-015-9240-y>
- Young, W. T. (1971). The role of musical aptitude, intelligence, and academic achievement in predicting the musical attainment of elementary instrumental music students. *Journal of Research in Music Education*, 19(4), 385–398. <https://doi.org/10.2307/3344291>
- Zanutto, E. L. (2006). A comparison of propensity score and linear regression analysis of complex survey data. *Journal of Data Science*, 4(1), 67–91.
- Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. University of Michigan Press.
- Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology*, 81(3), 329–339.
- Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, 25(1), 82–91. <https://doi.org/10.1006/ceps.1999.1016>
- Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal*, 45(1), 166–183. <https://doi.org/10.3102/0002831207312909>
- Zimmerman, B. J., & Bandura, A. (1994). Impact of self-regulatory influences on writing course attainment. *American Educational Research Journal*, 31(4), 845–862.
- Zimmerman, B. J., Bandura, A., & Martinez-Pons, M. (1992). Self-motivation for academic attainment: The role of self-efficacy beliefs and personal goal setting. *American Educational Research Journal*, 29(3), 663–676. <https://doi.org/10.2307/1163261>

- Zins, J. E., Bloodworth, M. R., Weissberg, R. P., & Walberg, H. J. (2004). The scientific base linking social and emotional learning to school success. In J. E. Zins, R. P. Weissberg, M. C. Wang, & H. J. Walberg (Eds.), *Building academic success on social and emotional learning: What does the research say?* (pp. 3–22). Teachers College Press.
- Zins, J. E., Weissberg, R. P., Wang, M. C., & Walberg, H. J. (Eds.). (2004). *Building academic success on social and emotional learning: What does the research say?* Teachers College Press.

Appendix A: Variables Used for Propensity Score Estimation

Table A.1

Variables Used for Propensity Score Estimation: Demographics

Description	HSLs:09 Variable Name
Sex	X1SEX
Race	X1RACE
Specific Asian Origin	X1ASIAN TYPE
Specific Hispanic Origin	X1HISPTYPE
Bilingual	X1DUALLANG
Native Language Spoken (If not English)	X1NATIVELANG
Speaks Other Language With Mother	S1LANGMOM
Speaks Other Language With Friends	S1LANGFRIEND
Parent Has Language Difficulty Joining in School Events	P1DIFSCHLNG
Born in United States	P1USBORN9
Parent 1 Born in U.S.	P1USBORN
Parent 2 Born in U.S.	P1USBORN2
Parent 1 Race	X1PAR1RACE
Parent 2 Race	X1PAR2RACE
Parents' Relationship Pattern	X1PARPATTERN
Number in Household	X1HHNUMBER

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), "Base-Year to Second Follow-Up, 2018."

Table A.2

Variables Used for Propensity Score Estimation: Socioeconomic Status

Description	HSLs:09 Variable Name
Parent 1 Educational Attainment	X1PAR1EDU
Parent 2 Educational Attainment	X1PAR2EDU
Parent 1 Employment Status	X1PAR1EMP
Parent 2 Employment Status	X1PAR2EMP
Parent 1 Occupation	X1PAR1OCC2
Parent 2 Occupation	X1PAR2OCC2
Family Below Poverty Level	X1POVERTY
Family At 130% of Poverty Level	X1POVERTY130
Family At 185% of Poverty Level	X1POVERTY185
Household Income	X1FAMINCOME
Socioeconomic Status Scale Score	X1SES
Socioeconomic Status With Urbanicity	X1SES_U
Own Home, Rent Home, or Other Arrangement	P1OWNHOME
Number of Times Changing Schools	P1CHANGESCH

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Table A.3

Variables Used for Propensity Score Estimation: Time Use

Description	HSLs:09 Variable Name
Hours Per Week Spent on Math Homework	S1HRMHOMWK
Hours Per Week Spent on Science Homework	S1HRSHOMWK
Hours Per Week Spent on Homework	S1HROTHHOMWK
Hours Per Week Spent on School Activities	S1HRACTIVITY
Hours Per Week Spent Working	S1HRWORK
Hours Per Week Spent With Family	S1HRFAMILY
Hours Per Week Spent With Friends	S1HRFRIENDS
Hours Per Week Watching TV	S1HRTV
Hours Per Week Playing Video Games	S1HRVIDEO
Hours Per Week Spent Online	S1HRONLINE

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Table A.4

Variables Used for Propensity Score Estimation: Activities

Description	HSLs:09 Variable Name
In the Past Year:	
Participated in a Math Club	S1MCLUB
Participated in a Math Competition	S1MCOMPETE
Attended a Math Camp	S1MCAMP
Participated in Math Tutoring or Study Group	S1MTUTOR
Participated in a Science Club	S1SCLUB
Participated in a Science Competition	S1SCOMPETE
Attended a Science Camp	S1SCAMP
Participated in Science Tutoring or Study Group	S1STUTOR
Did Not Participate in Listed Math or Science Activities	S1NOMSACT
Reads Science Books or Magazines	S1SBOOKS
Uses Internet for Technology Information	S1WEBINFO
Visited Science Museum or Planetarium	S1SMUSEUM
In the Last Year, Participated in Extracurricular:	
Visual or Performing Arts	P1ARTS
Organized Sports	P1SPORTS
Religious Group	P1RELIGGRP
Scouting or Other Group or Club	P1CLUB
Academic Instruction or Tutoring	P1ACADEMIC
Math or Science Camp	P1CAMPMS
Another Camp	P1CAMPOTH
None of These Activities	P1NOOUTSCH
In the Last Year, Parent and Student:	
Visited A Science or Engineering Museum	P1MUSEUM
Worked or Played on a Computer	P1COMPUTER
Built or Fixed Something	P1FIXED
Visited a Library	P1LIBRARY
Went to a Concert, Play, or Live Show	P1SHOW
None of These Activities	P1NOACT

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), "Base-Year to Second Follow-Up, 2018."

Table A.5

Variables Used for Propensity Score Estimation: School Attitudes

Description	HSLs:09 Variable Name
Student Feels Safe at School	S1SAFE
Student Is Proud to Be Part of His/Her School	S1PROUD
Student Has Teacher/Adult in School He/She Can Talk to About Problems	S1TALKPROB
Student Feels That School Is a Waste of Time	S1SCHWASTE
Getting Good Grades Is Important to the Student	S1GOODGRADES
How Often Student Goes to Class Without Their Homework Done	S1NOHWDN
How Often Student Goes to Class Without Pencil or Paper	S1NOPAPER
How Often Student Goes to Class Without Books	S1NOBOOKS
How Often Student Goes to Class Late	S1LATE
Student Thinks Studying in School Rarely Pays Off Later With Good Job	S1PAYOFF
Student Thinks Even if He/She Studies He/She Won't Get Into College	S1TEFRNDS
Student Believes Time/Effort in Math/Science Means Not Enough Time for Extracurriculars	S1TEACTIV
Student Believes Time/Effort in Math/Science Means He/She Won't Be Popular	S1TEPOPULAR
Student Believes Time/Effort in Math/Science Means People Will Make Fun of Him/Her	S1TEMAKEFUN
Student Has Teacher/Adult in School He/She Can Talk to About Problems	S1TALKPROB
Student Talked to Mother About Personal Problems	S1MOMTALKPRB
Student Talked to Father About Personal Problems	S1DADTALKPRB
Student Talked to a Friend About Personal Problems	S1FRNDTLKPRB
Student Talked to a Teacher About Personal Problems	S1TCHTALKPRB
Student Talked to School Counselor About Personal Problems	S1CNSLTLKPRB
Student Didn't Talk to These People About Personal Problems	S1NOTALKPRB

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), "Base-Year to Second Follow-Up, 2018."

Table A.6

Variables Used for Propensity Score Estimation: School Behaviors

Description	HSLs:09 Variable Name
Difficulty Learning or Paying Attention	P1LEARN
Difficulty Speaking or Communicating	P1SPEAK
Difficulty With Anxiety or Depression	P1MOOD
Difficulty With Behavior Problems	P1ACTOUT
Difficulty Making or Keeping Friends	P1FRIEND
Suspension or Expulsion	P1SUSPEND
Parent Contacts for Problem Behavior	P1BEHAVE
Parent Contacts for Attendance	P1ATTEND
Parent Contacts for Academic Performance	P1PERFORM

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), "Base-Year to Second Follow-Up, 2018."

Table A.7

Variables Used for Propensity Score Estimation: School and Postsecondary Expectations

Description	HSLs:09 Variable Name
How Far Student Expects to Go in School	X1STUEDEXPCT
How Far Parent Expects Student to Go in School	X1PAREDEXPCT
In First Year After High School:	
Likely to Pursue an Associate's	S1FYAA
Likely to Pursue a Bachelor's	S1FYBA
Likely to Obtain a License or Certificate	S1FYLICENSE
Likely to Pursue an Apprenticeship	S1FYAPPR
Likely to Enter the Military	S1FYMILITARY
Likely to Enter the Workforce	S1FYJOB
Likely to Start a Family	S1FYFAMILY
Likely to Travel	S1FYTRAVEL
Likely to Volunteer	S1FYVOLUN
Unsure of Plans After High School	S1FYNOTSURE

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), "Base-Year to Second Follow-Up, 2018."

Table A.8

Variables Used for Propensity Score Estimation: Prior Academic Achievement

Description	HSLs:09 Variable Name
Ever Repeated a Grade	P1REPEATGRD
Ever Skipped a Grade	P1SKIPGRD
8 th Grade Math Course Taken	S1M8
8 th Grade Math Course Grade	S1M8GRADE
8 th Grade Science Course Taken	S1S8
8 th Grade Science Course Grade	S1S8GRADE

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), "Base-Year to Second Follow-Up, 2018."

Table A.9

Variables Used for Propensity Score Estimation: School-Level Variables

Description	HSLs:09 Variable Name
Locale	X1LOCALE
Region of the Country	X1REGION
Census Division	X1CENDIV
State	X1STATE
School Size	A2HSSIZE
Percent of Students Repeating 9 th Grade	X1REPEAT9TH
Percent American Indian Students	X1SCHAMIND
Percent Asian Students	X1SCHASIAN
Percent Black Students	X1SCHBLACK
Percent Hispanic Students	X1SCHHISP
Percent White Students	X1SCHWHITE
Average Daily Attendance Rate	A1ADA
School Is in Need of Improvement After Not Making AYP	A1AYP

Continued

Table A.9 Continued

Description	HSLs:09 Variable Name
Scale of School Climate	X1SCHOOLCLI
Scale of School Problems	X2PROBLEM
Whether School Offers IB Courses	A1IB
Whether School Offers AP Courses	A1OFFERAP
Percent Receiving Free or Reduced-Price Lunch	A1FREELUNCH
Percent ELL Students	A1ELL
Percent in Special Education	A1SPECIALED
Percent in Alternative Programs	A1ALTPROG
Percent in Dropout Prevention Programs	A1DROPOUTPRV
Percent in AP Courses	A1AP
Percent Returning To Same School	A1RETURN09
After Graduation, Percent Who:	
Pursued Bachelor's Degree	A14YRDEGREE
Pursued Associate's Degree	A12YRDEGREE
Entered the Work Force	A1WORK
Entered the Military	A1MILITARY
Did Something Else	A1DIDOTHER
Number of Full-Time Teachers	A1FTTCHRS
Number of Part-Time Teachers	A1PTTCHRS
Difficulty Filling Math Vacancies	A1FILLMTH
Difficulty Filling Science Vacancies	A1FILLSCI
Average Percent Teachers Absent	A1ABSENTTCHR

SOURCE: *Public Elementary/Secondary School Universe Survey, 2009–10*, Common Core of Data (CCD), Version Provisional 2a, U.S. Department of Education, National Center for Education Statistics; and U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Appendix B: Missingness Rates

Table B.1

Missingness Rates: Student-Level Variables

HSLs:09 Variable Name	Initial Percent Missing	Percent Missing After Imputing from Other Data Sources
X1SEX	0.23%	0.00%
X1RACE	7.83%	0.00%
X1ASIAN TYPE	5.47%	3.64%
X1HISP TYPE	10.17%	6.49%
X1DUALLANG	15.02%	9.58%
X1NATIVELANG	13.02%	8.17%
S1LANGMOM	15.08%	9.60%
S1LANGFRIEND	15.13%	9.68%
P1DIFSCHLNG	37.13%	31.02%
P1USBORN9	35.82%	29.57%
P1USBORN	35.81%	29.56%
P1USBORN2	34.89%	28.75%
X1PAR1RACE	36.18%	29.92%
X1PAR2RACE	35.11%	29.04%
X1PARPATTERN	32.70%	1.33%
X1HHNUMBER	32.77%	26.49%
X1PAR1EDU	32.60%	26.31%
X1PAR2EDU	32.61%	26.32%
X1PAR1EMP	32.75%	26.44%
X1PAR2EMP	32.70%	26.41%
X1PAR1OCC2	32.66%	26.35%
X1PAR2OCC2	32.62%	26.34%
X1POVERTY	32.93%	26.58%
X1POVERTY130	32.93%	26.58%
X1POVERTY185	32.93%	26.58%
X1FAMINCOME	32.76%	26.42%

Continued

Table B.1 Continued

HSL:09 Variable Name	Initial Percent Missing	Percent Missing After Imputing from Other Data Sources
X1SES	12.75%	8.21%
X1SES_U	12.75%	8.21%
P1OWNHOME	37.04%	22.91%
P1CHANGESCH	36.86%	30.52%
S1HRMHOMWK	17.12%	11.53%
S1HRSHOMWK	17.24%	11.68%
S1HROTHHOMWK	17.94%	12.25%
S1HRACTIVITY	18.28%	12.69%
S1HRWORK	20.66%	15.03%
S1HRFAMILY	18.10%	12.41%
S1HRFRIENDS	18.11%	12.42%
S1HRTV	18.04%	12.27%
S1HRVIDEO	19.82%	13.97%
S1HRONLINE	18.78%	13.07%
S1MCLUB	16.38%	10.90%
S1MCOMPETE	16.38%	10.90%
S1MCAMP	16.38%	10.90%
S1MTUTOR	16.38%	10.90%
S1SCLUB	16.38%	10.90%
S1SCOMPETE	16.38%	10.90%
S1SCAMP	16.38%	10.90%
S1STUTOR	16.38%	10.90%
S1NOMSACT	16.38%	10.90%
S1SBOOKS	15.95%	10.55%
S1WEBINFO	16.24%	10.63%
S1SMUSEUM	16.04%	10.53%

Continued

Table B.1 Continued

HSLs:09 Variable Name	Initial Percent Missing	Percent Missing After Imputing from Other Data Sources
P1ARTS	38.34%	32.13%
P1SPORTS	38.34%	32.13%
P1RELIGGRP	38.34%	32.13%
P1CLUB	38.34%	32.13%
P1ACADEMIC	38.34%	32.13%
P1CAMPMS	38.34%	32.13%
P1CAMPOTH	38.34%	32.13%
P1NOOUTSCH	38.34%	32.13%
P1MUSEUM	38.03%	31.77%
P1COMPUTER	38.03%	31.77%
P1FIXED	38.03%	31.77%
P1LIBRARY	38.03%	31.77%
P1SHOW	38.03%	31.77%
P1NOACT	38.03%	31.77%
S1SAFE	16.18%	10.54%
S1PROUD	16.63%	11.00%
S1TALKPROB	16.58%	11.00%
S1SCHWASTE	16.68%	11.03%
S1GOODGRADES	16.44%	10.77%
S1NOHWDN	16.21%	10.60%
S1NOPAPER	16.42%	10.78%
S1NOBOOKS	16.52%	10.87%
S1LATE	16.51%	10.85%
S1PAYOFF	16.62%	11.00%
S1TEFRNDS	17.05%	11.51%
S1TEACTIV	17.30%	11.63%
S1TEPOPULAR	17.39%	11.82%
S1TEMAKEFUN	17.62%	12.10%

Continued

Table B.1 Continued

HSLs:09 Variable Name	Initial Percent Missing	Percent Missing After Imputing from Other Data Sources
S1TALKPROB	16.58%	11.00%
S1MOMTALKPRB	17.11%	11.56%
S1DADTALKPRB	16.93%	11.46%
S1FRNDTLKPRB	17.15%	11.60%
S1TCHTALKPRB	17.15%	11.60%
S1CNSLTLKPRB	17.15%	11.60%
S1NOTALKPRB	17.15%	11.60%
P1LEARN	37.69%	31.41%
P1SPEAK	37.67%	31.43%
P1MOOD	37.89%	31.62%
P1ACTOUT	37.67%	31.41%
P1FRIEND	37.73%	31.50%
P1SUSPEND	36.72%	30.41%
P1BEHAVE	37.70%	31.45%
P1ATTEND	37.86%	31.66%
P1PERFORM	37.96%	31.71%
X1STUEDEXPCT	14.93%	9.46%
X1PAREDEXPCT	32.79%	26.44%
S1FYAA	16.60%	11.06%
S1FYBA	16.60%	11.06%
S1FYLICENSE	16.60%	11.06%
S1FYAPPR	16.60%	11.06%
S1FYMILITARY	16.60%	11.06%
S1FYJOB	16.60%	11.06%
S1FYFAMILY	16.60%	11.06%
S1FYTRAVEL	16.60%	11.06%
S1FYVOLUN	16.60%	11.06%
S1FYNOTSURE	16.60%	11.06%

Continued

Table B.1 Continued

HSLs:09 Variable Name	Initial Percent Missing	Percent Missing After Imputing from Other Data Sources
P1REPEATGRD	36.56%	30.27%
P1SKIPGRD	36.68%	30.37%
S1M8	15.57%	10.07%
S1M8GRADE	16.22%	10.67%
S1S8	16.53%	10.98%
S1S8GRADE	17.28%	11.74%
X1TXMTH	14.93%	9.46%
X2TXMTH	18.30%	4.06%
X1MTHID	16.06%	10.54%
X1MTHUTI	17.02%	11.40%
X1MTHEFF	17.87%	11.60%
X1MTHINT	18.64%	12.84%
X2MTHID	21.14%	7.30%
X2MTHUTI	20.78%	6.79%
X2MTHEFF	21.59%	7.31%
X2MTHINT	23.43%	10.12%
X1SCIID	16.25%	10.72%
X1SCIUTI	17.03%	11.45%
X1SCIEFF	17.22%	11.53%
X1SCIINT	20.30%	12.73%
X2SCIID	20.90%	7.02%
X2SCIUTI	21.19%	7.27%
X2SCIEFF	22.42%	8.32%
X2SCIINT	22.56%	9.08%
X3TXSATMATH	55.74%	51.33%
X3TXSATREAD	55.80%	51.39%
X3TXSATCOMP	55.82%	51.41%
X1SCHOOLBEL	18.25%	12.25%
X1SCHOOLLENG	17.98%	11.43%
X2BEHAVEIN	20.85%	6.77%

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), "Base-Year to Second Follow-Up, 2018."

Table B.2

Missingness Rates: School-Level Covariates

HSLs:09 Variable Name	Initial Percent Missing	Percent Missing After Imputing from Other Data Sources
X1LOCALE	0.00%	0.00%
X1CENDIV	0.00%	0.00%
X1STATE	0.00%	0.00%
A2HSSIZE	3.71%	0.11%
X1REPEAT9TH	8.72%	9.55%
X1SCHAMIND	8.46%	0.34%
X1SCHASIAN	8.18%	0.34%
X1SCHBLACK	8.05%	0.34%
X1SCHHISP	7.89%	0.34%
X1SCHWHITE	8.02%	0.34%
A1ADA	18.76%	3.38%
A1AYP	11.46%	13.26%
X1SCHOOLCLI	21.58%	21.57%
X2PROBLEM	24.29%	7.90%
A1IB	14.81%	15.01%
A1OFFERAP	6.04%	6.64%
A1FREELUNCH	8.44%	0.11%
A1ELL	8.10%	0.27%
A1SPECIALED	8.53%	0.17%
A1ALTPROG	9.73%	10.40%
A1DROPOUTPRV	14.56%	17.04%
A1AP	10.53%	11.04%
A1RETURN09	19.37%	20.80%
A1FTTCHRS	13.40%	0.11%
A1PTTCHRS	8.69%	0.44%
A1FILLMTH	14.75%	2.18%
A1FILLSCI	15.09%	2.08%
A1ABSENTTCHR	17.48%	1.99%

SOURCE: *Public Elementary/Secondary School Universe Survey, 2009–10*, Common Core of Data (CCD), Version Provisional 2a, U.S. Department of Education, National Center for Education Statistics; and U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Appendix C: Stata Programs for Processing HSLS:09 Transcript Data

qrec

qrec stands for “quickly recode” and recodes the HSLS:09’s numerical missing-value information with Stata’s missing-value indicators.

```
program qrec
if "`": value label `1'''' != "Cont" {
    recode `1'      (-1=.) /// Don't know
                    (-3=.c) /// Carry through missing
                    (-4=. ) /// Item not administered: abbreviated interview
                    (-6=.n) /// Component not applicable
                    (-7=.n) /// NA
                    (-8=. ) /// Unit non-response
                    (-9=. ) /// Missing
    }
else {
    qui summ `1'
    if `r(min)' <= 4 {
        recode `1'      (-6=.n) /// Component not applicable
                        (-7=.n) /// NA
                        (-8=. ) /// Unit non-response
                        (-9=. ) /// Missing
    }
    else {
        recode `1'      (-4=. ) /// Item not administered: abbreviated interview
                        (-6=.n) /// Component not applicable
                        (-7=.n) /// NA
                        (-8=. ) /// Unit non-response
                        (-9=. ) /// Missing
    }
}
end
```

x12

x12 displays either crosstabs or side-by-side histograms of the BY and F1 values of a variable. The program is called with the full variable name after the X1 or X2 designation; for example, to investigate X1SES and X2SES, the call is x12 SES.

```
program x12
capture drop P1 P2

if "`": value label X1`1'"" != "Cont" {
    egen P1 = anymatch(X1`1'), v(-1/-9)
    qui summ P1
    scalar R1 = `r(max)'
    egen P2 = anymatch(X2`1'), v(-1/-9)
    qui summ P2
    scalar R2 = `r(max)'
}
else {
    egen P1 = anymatch(X1`1'), v(-4/-9)
    qui summ P1
    scalar R1 = `r(max)'
    egen P2 = anymatch(X2`1'), v(-4/-9)
    qui summ P2
    scalar R2 = `r(max)'
}

if R1 == 1 {
    qui preserve
    qui qrec      X1`1'
}

if R2 == 1 & R1 == 0 {
    qui preserve
}

if R2 == 1 {
    qui qrec      X2`1'
}

di _n _n _n "*****      " "x12 FOR VARIABLE:  `1'" " *****"

if "`": value label X1`1'"" != "Cont" {
    tab      X1`1' X2`1', m
}
else {
    qui gen      x1 = X1`1'
    qui gen      x2 = X2`1'
    qui hist      x1, nodraw xti(" ") yti(" ") name(g1, replace)
    qui hist      x2, nodraw xti(" ") yti(" ") name(g2, replace)
    gr combine     g1 g2, xsize(11) ti("`1'") subti(Missing values excluded)
    drop          x1 x2
    gr drop        g1 g2
}

if R1 == 1 | R2 == 1 {
    qui restore
}

capture drop P1 P2
capture scalar drop R1 R2
end
```

cors

cors stands for “correlations” and displays the correlation coefficient and Cronbach’s α between the BY and F1 values of a variable in order to help determine whether it is appropriate to aggregate their values. The program is called with the variable name after the X1 or X2 prefix.

```
program cors
capture drop P1 P2

if ``: value label X1`1'"" != "Cont" {
    egen P1 = anymatch(X1`1'), v(-1/-9)
    qui summ P1
    scalar R1 = `r(max)'
    egen P2 = anymatch(X2`1'), v(-1/-9)
    qui summ P2
    scalar R2 = `r(max)'
}
else {
    egen P1 = anymatch(X1`1'), v(-4/-9)
    qui summ P1
    scalar R1 = `r(max)'
    egen P2 = anymatch(X2`1'), v(-4/-9)
    qui summ P2
    scalar R2 = `r(max)'
}

if R1 == 1 {
    qui preserve
    qui qrec      X1`1'
}
if R2 == 1 & R1 == 0 {
    qui preserve
}
if R2 == 1 {
    qui qrec      X2`1'
}

qui gen x1 = X1`1'
qui gen x2 = X2`1'

di _n _n _n "***** " "cors FOR VARIABLE: `1'" " *****"

qui alpha x1 x2
di "ALPHA: " r(alpha)
qui cor x1 x2
di "CORRELATION: " r(rho)
drop x1 x2

if R1 == 1 | R2 == 1 {
    qui restore
}
capture drop P1 P2
capture scalar drop R1 R2
end
```

fix12

fix12 sets the value of the BY variable *X* to either the BY value or the F1 value, according to the options specified by the analyst. The program begins by choosing the observed value of either variable if one is missing. The analyst can specify the “highest” or “lowest” option to specify that the higher or lower of the BY and F1 values should be chosen, or the “use1” or “use2” options to denote that the first or second wave’s value should be assigned to the BY variable for use in analysis. Variable names are handled similarly to `cors` and `x12`.

```
program fix12
capture drop P1 P2

if ``': value label X1`1'' != "Cont" {
    egen P1 = anymatch(X1`1'), v(-1/-9)
    qui summ P1
    scalar R1 = `r(max)'
    egen P2 = anymatch(X2`1'), v(-1/-9)
    qui summ P2
    scalar R2 = `r(max)'
}
else {
    egen P1 = anymatch(X1`1'), v(-4/-9)
    qui summ P1
    scalar R1 = `r(max)'
    egen P2 = anymatch(X2`1'), v(-4/-9)
    qui summ P2
    scalar R2 = `r(max)'
}

if R1 == 1 {
    qui preserve
    qui qrec      X1`1'
}
if R2 == 1 & R1 == 0 {
    qui preserve
}
if R2 == 1 {
    qui qrec      X2`1'
}

di _n _n _n "***** " "fix12 FOR VARIABLE: `1'" " *****"

di _n "Replacing X1 with X2 if X1 is missing and X2 is not missing"
replace      X1`1'=X2`1' if missing(X1`1') & !missing(X2`1')

di _n "Replacing X2 with X1 if X2 is missing and X1 is not missing"
replace      X2`1'=X1`1' if missing(X2`1') & !missing(X1`1')
```

```

if "`2'" == "highest" {
    di _n "Replacing X1 with X2 only if X2 is higher and not missing"
    replace      X1`1' = X2`1' if X2`1' > X1`1' & !missing(X2`1')
}

else if "`2'" == "lowest" {
    di _n "Replacing X1 with X2 only if X2 is lower and not missing"
    replace      X1`1' = X2`1' if X2`1' < X1`1' & !missing(X2`1')
}

else if "`2'" == "use1" {
    di _n "Replacing X2 with X1 if they are different and X1 is not missing"
    replace      X2`1' = X1`1' if X2`1' != X1`1' & !missing(X1`1')
}

else if "`2'" == "use2" {
    di _n "Replacing X1 with X2 if they are different and X2 is not missing"
    replace      X1`1' = X2`1' if X1`1' != X2`1' & !missing(X2`1')
}

if ``': value label X1`1' != "Cont" {
    di _n "And here's where we ended up"
    tab      X1`1' X2`1', m
}

else {
    qui gen      x1 = X1`1'
    qui gen      x2 = X2`1'
    qui hist      x1, nodraw xti(" ") yti(" ") name(g1, replace)
    qui hist      x2, nodraw xti(" ") yti(" ") name(g2, replace)
    gr combine    g1 g2, xsize(11) ti("`1'") subti(Missing values excluded)
    drop          x1 x2
    gr drop      g1 g2
}

if R1 == 1 | R2 == 1{
    qui restore
}

capture drop P1 P2
capture scalar drop R1 R2

end

```

swap

swap is a standalone program written as a DO file. The analyst updates the swap.do file with individual class sequence numbers (T3SCRSE_SEQ in the HSLS:09 transcript file), the incorrect music classification, and the correct music classification. Then, interactively or in a separate DO file, the analyst calls `do swap {subject}` where {subject} is the name of the indicator variable of any of the music classifications used to update the music classifications of

all of the courses listed in that subject. (This is to avoid having to wait while the entire program re-runs changes in other subjects.) In this study, the music subject indicator variables were band, choir, orch, jazz, inst, guitar, piano, comp, mtha, biz, tech, and other. The analyst can also call `do swap orphan` to process courses originally not classified as music courses. In the `swap.do` file, the analyst replaces the “#####” with the appropriate value of T3SCRSE_ID identifier, the incorrect subject, and the correct subject. `swap` can be run as often as desired.

```
set output error
args prev

capture program drop swap
program swap
  args clnum old new
  replace `old' = 0 if T3SCRSE_SEQ == `clnum' & `old'==1
  if "`new'" != "" {
    replace `new' = 1 if T3SCRSE_SEQ == `clnum'
  }
end

if strpos("`prev'", "orphan") > 0 {
  set output proc
  swap          ##### music choir           // keep "music" as a placeholder
  swap          ##### music other
  set output error
}

else if strpos("`prev'", "band") > 0 {
  set output proc
  swap          ##### band choir
  swap          ##### band orch
  set output error
}

else if strpos("`prev'", "orch") > 0 {
  set output proc
  swap          ##### orch guitar
  swap          ##### orch
  set output error
}

else if strpos("`prev'", "choir") > 0 {
  set output proc
  swap          ##### choir mtha
  swap          ##### choir band
  set output error
}

else if strpos("`prev'", "jazz") > 0 {
  set output proc
  swap          ##### jazz guitar
  set output error
}
```

```

else if strpos("`prev'", "guitar") > 0 {
set output proc
swap          ##### guitar piano
set output error
}

else if strpos("`prev'", "piano") > 0 {
set output proc
swap          ##### piano jazz
set output error
}

else if strpos("`prev'", "comp") > 0 {
set output proc
swap          ##### comp mtha
set output error
}

else if strpos("`prev'", "mtha") > 0 {
set output proc
swap          ##### mtha comp
set output error
}

else if strpos("`prev'", "inst") > 0 {
set output proc
swap          ##### inst
swap          ##### inst choir
set output error
}

else if strpos("`prev'", "tech") > 0 {
set output proc
swap          ##### tech comp
set output error
}

else if strpos("`prev'", "biz") > 0 {
set output proc
swap          ##### biz tech
set output error
}

else if strpos("`prev'", "other") > 0 {
set output proc
swap          ##### other mtha
swap          ##### other jazz
set output error
}

else {
set output proc
di as error "Warning! Not recognized!"
}

set output proc

```


orphan

`orphan` (and the related program `noswitch`) perform batch classifications of courses initially classified as non-music to the appropriate music classification using key words from the course titles. For example, if the analyst calls `orphan wind band`, the program classifies all non-music courses with the word “wind” in the title as band courses. Quotes can be used for multi-word phrases; for example, a call of `orphan “voc mus” choir` would classify courses with “voc mus” in the title as choir classes. Appropriate key words can be found by reviewing results from the “list” option, which displays the course titles of all non-music courses containing the keyword from the call. For example, `orphan list “music theory”` would return a list of all courses not classified as music containing the string “music theory” in the title. Afterward, batch changes can be made using `orphan` or individual changes can be made using `swap`.

```
program orphan
capture confirm variable nosw
if _rc {
    gen          nosw = 0
}
if "`1'" == "list" {
    list          T3SCRSE_SEQ T3SCRSNAM if music==0 & ///
                  strpos(lower(T3SCRSNAM), lower("`2'")) > 0
}
else {
    capture confirm    variable `2'
    if !_rc {
        replace `2' = 1 if music==0 & `2' == 0 & nosw==0 & ///
                  strpos(lower(T3SCRSNAM), lower("`1'")) > 0
        replace music = 1 if music==0 & nosw==0 & ///
                  strpos(lower(T3SCRSNAM), "`1'") > 0
    }
    else {
        di as error "`2' is not a variable. SORRY!!!"
    }
}
capture drop nosw
end
```

noswitch

When `noswitch` is called prior to calling `orphan`, it excludes the listed words or phrases from being changed. For example, the code `noswitch garage before orphan band band` would classify all courses with “band” in the title as band classes except those that also contained the word “garage” in the title.

```
program noswitch
capture drop nosw
gen                nosw = 0

forvalues i = 1/11 {
    if "`i'" != "" {
        di      "`i'"
        replace      nosw = 1 if strpos(lower(T3SCRSNAM), "`i'") > 0
    }
}
end
```

wordlist

`wordlist` returns a list of courses with titles that do not contain up to 14 key words (or double-quoted phrases) in order to help clean up existing classifications. For example, `wordlist band concert march wind` would return a list of all of the courses classified as band courses that do not contain the words “concert,” “march,” or “wind.”

```
program wordlist
capture drop crsnam
gen                crsnam = strpos(lower(T3SCRSNAM), "`1'") > 0

forvalues i = 2/15 {
    if "`i'" != "" {
        di      "`i'"
        replace      crsnam = 1 if strpos(lower(T3SCRSNAM), "`i'") > 0
    }
}

list                T3SCRSE_SEQ T3SCRSNAM if `1'==1 & crsnam == 0
drop                crsnam
end
```

cleanup

While orphan assisted the analyst with identifying music courses initially mis-coded as non-music classes, cleanup automates the “cleanup” of the music course classifications that are already done; i.e., ensuring that they are properly classified by music subject. Calling `cleanup list {subject} keyword` returns the titles of the course currently coded as the specified subject that contain the keyword. For example, `cleanup list band string` lists the courses currently classified as band courses that contain the word “string” (as with the other commands, double-quoted phrases are accepted). Calling cleanup without the list argument automates the reclassification of courses. For example, `cleanup band string orch` reclassifies band courses with “string” in the title as orchestra courses.

```
program cleanup
global          stus "bandstu orchstu choirstu"
if "`1'" == "list" {
    list          T3SCRSE_SEQ T3SCRSNAM $stus if `2'==1 & ///
                  strpos(lower(T3SCRSNAM), "`3'") > 0
}
else {
    capture confirm variable `2'
    if !_rc {
        replace    `2' = 1 if `1'==1 & strpos(lower(T3SCRSNAM), "`2'") > 0
        replace    `1' = 0 if `2'==1 & `1'==1
    }
    else {
        replace    `3' = 1 if `1'==1 & strpos(lower(T3SCRSNAM), "`2'") > 0
        replace    `1' = 0 if `3'==1 & `1'==1
    }
}
end
```

sturedo

sturedo re-classifies students as band students, orchestra students, or choir students for current listing with cleanup and sked.

```
program sturedo
capture drop $stus

bysort STU_ID:      egen bandstu = max(band)
lab var            bandstu "Ever took band"
bysort STU_ID:      egen orchstu = max(orch)
lab var            orchstu "Ever took orchestra"
bysort STU_ID:      egen choirstu = max(choir)
lab var            choirstu "Ever took choir"

sort               STU_ID T3SCRSE_SEQ
end
```

sked

In order to help classify a course with an ambiguous title, sked displays a student's complete coded transcript history and whether each music course listed was classified as band, orchestra, jazz, choir, or instrumental.

```
program sked
global            mus1 "band orch jazz choir inst"
qui sum           STU_ID if T3SCRSE_SEQ == `1'
list              T3SCRSNAM $mus1 if STU_ID==r(mean), sepby(T3STERM)
end
```

Appendix D: Stata Programs for Calculating the Effect Size

getsvysd

This program returns the pooled weighted standard deviation necessary to calculate the effect size statistic Hedges's g with multiply imputed complex survey data. `getsvysd` does not return a value, but stores the standard deviation in a scalar named `sdstar`. This program should be run once per dependent variable to keep `sdstar` current. Miguel Dorta from StataCorp suggested the `intreg` workaround to estimate the standard deviation correctly. In the code for both of these programs, `$tx` is a macro for the treatment indicator variable.

```
*data should already be svyset or mi svyset
program getsvysd
args var

if "`var'" == "" {
    di as error "You need to include the DV!"
    error 100
}

preserve
qui mi passive: gen double `var'_temp = `var'
qui mi estimate, cmdok: svy, subpop(if $tx == 0): intreg `var' `var'_temp //control grp
scalar sd_c = exp(el(e(b_mi),1,2))
qui mi estimate, cmdok: svy, subpop(if $tx == 1): intreg `var' `var'_temp //tx grp
scalar sd_t = exp(el(e(b_mi),1,2))

qui count if $tx == 0 & _mi_m == 1
scalar n_c = `r(N)'
qui count if $tx == 1 & _mi_m == 1
scalar n_t = `r(N)'

scalar sdstar = sqrt(((n_c-1)*(sd_c^2)) + ((n_t-1)*(sd_t^2)))/(n_c + n_t - 2))

restore

di "The pooled weighted standard deviation is: " `=sdstar'

end
```

getsvyg

This program uses the pooled weighted standard deviation from `getsvysd` in order to calculate values of Hedges's g . It is currently configured to be run after a regression equation using `mi estimate: svy: regress` with arguments for main effects (with no options) or with options to run after the `pwcompare` command, such as:

```
mimrgns $tx
pwcompare $tx, over(X1MALE) cmdmargins pwcompare(pveffects)
getsvyg male

*data should already be svyset or mi svyset
*for new applications, the particular elements of matrix rbvs can be changed to
*calculate glass's g for any of the tests that appear in your -pwcompare- results

program getsvyg
args whichmargins

if "`whichmargins'" == "" {
    matrix rbvs = r(b_vs)
    scalar diff = rbvs[1,1]
    di as result "Glass's g: " round(`=diff' / `=sdstar', .01) _n _n
    matrix drop rbvs
}
else if "`whichmargins'" == "lowgr" {
    matrix rbvs = r(b_vs)
    scalar diff = rbvs[1,6]
    di as result "Low grade Glass's g: " round(`=diff' / `=sdstar', .01) _n _n
    scalar diff = rbvs[1,1]
    di as result "Not low grade Glass's g: " round(`=diff' / `=sdstar', .01) _n _n
    matrix drop rbvs
}
else if "`whichmargins'" == "male" {
    matrix rbvs = r(b_vs)
    scalar diff = rbvs[1,6]
    di as result "Male Glass's g: " round(`=diff' / `=sdstar', .01) _n _n
    scalar diff = rbvs[1,1]
    di as result "Female Glass's g: " round(`=diff' / `=sdstar', .01) _n _n
    matrix drop rbvs
}
else if "`whichmargins'" == "white" {
    matrix rbvs = r(b_vs)
    scalar diff = rbvs[1,6]
    di as result "White Glass's g: " round(`=diff' / `=sdstar', .01) _n _n
    scalar diff = rbvs[1,1]
    di as result "Not white Glass's g: " round(`=diff' / `=sdstar', .01) _n _n
    matrix drop rbvs
}
else {
    di as error "Incorrect option specified."
    error 111
}
end
```

Appendix E: Regression Tables

Regression Tables for Academic Outcomes

Choral Music

Table E.1

Regression Results for College Examination Mathematics Scores for 1-2 Years of Choral Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	-16.495 (8.724)	0.829 (7.199)	-13.666 (13.221)	-2.171 (7.896)	2.376 (6.015)	-12.076 (10.895)
Male		22.322*** (3.772)	18.134*** (3.935)		17.549** (6.630)	19.123*** (4.590)
Male × Music			-10.768 (12.718)			-11.484 (12.923)
White		17.973*** (4.791)	16.791*** (4.886)		25.709** (8.048)	17.003** (5.650)
White × Music			19.353 (13.338)			20.571 (11.925)
SES		20.642*** (3.136)	19.243*** (3.165)		27.637*** (5.224)	19.254*** (3.490)
SES × Music			14.906 (8.793)			14.297 (8.697)
School Engagement (9 th Grade)		-0.074 (2.299)	-0.051 (2.178)		1.272 (4.410)	1.910 (3.540)
C, D, or F in Math (8 th Grade)		-23.882*** (5.532)	-27.327*** (5.846)		-19.307* (9.413)	-34.027*** (7.893)
C, D, or F in Math × Music			45.215** (15.833)			42.927* (17.634)
Math Reasoning Score (9 th Grade)		68.282*** (3.110)	59.995*** (4.123)		57.383*** (6.128)	57.393*** (4.779)
Square Term			13.063*** (2.118)			16.409*** (3.252)
9 th Grade Math Reasoning × Male			8.233 (4.320)			10.394 (6.017)
9 th Grade Math Reasoning × Music			-9.370 (8.150)			-7.843 (7.094)
Intercept	496.182*** (3.853)	457.030*** (4.838)	450.838*** (5.312)	485.939*** (3.739)	452.765*** (7.410)	446.672*** (6.059)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), "Base-Year to Second Follow-Up, 2018."

Table E.2

Regression Results for College Examination Mathematics Scores for 3-4 Years of Choral Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	4.661 (9.992)	6.908 (6.575)	-3.338 (16.627)	3.409 (8.825)	4.241 (6.094)	-3.805 (14.810)
Male		23.955*** (3.724)	18.116*** (4.009)		29.941** (9.233)	7.010 (6.439)
Male × Music			13.731 (16.103)			19.078 (15.836)
White		16.998*** (4.670)	16.913*** (4.863)		14.295 (8.496)	16.696* (6.771)
White × Music			-0.127 (14.901)			-5.265 (13.910)
SES		19.012*** (2.972)	19.328*** (3.161)		17.983** (5.700)	23.773*** (3.732)
SES × Music			-5.005 (9.432)			-7.497 (9.359)
School Engagement (9 th Grade)		-0.750 (2.221)	-0.861 (2.164)		-1.647 (4.028)	-1.790 (3.936)
C, D, or F in Math (8 th Grade)		-25.999*** (5.590)	-27.587*** (5.854)		-21.655* (9.567)	-34.871*** (6.846)
C, D, or F in Math × Music			14.559 (15.217)			19.644 (15.401)
Math Reasoning Score (9 th Grade)		71.331*** (2.772)	60.344*** (4.160)		79.091*** (4.777)	58.610*** (6.323)
Square Term			11.692*** (2.265)			6.330 (4.344)
9 th Grade Math Reasoning × Male			8.796 (4.551)			19.539* (9.514)
9 th Grade Math Reasoning × Music			17.206* (8.384)			17.993* (8.342)
Intercept	496.182*** (3.853)	456.534*** (4.757)	451.868*** (5.263)	502.106*** (4.369)	458.304*** (8.246)	462.159*** (5.758)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table E.3

Regression Results for College Examination Reading Scores for 1-2 Years of Choral Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	-4.572 (8.690)	0.859 (8.082)	-11.029 (14.177)	1.830 (7.825)	2.221 (6.907)	-5.018 (11.471)
Male		10.696 (7.582)	11.430 (8.092)		4.126 (8.371)	7.153 (6.223)
Male × Music			-10.627 (20.136)			-5.776 (18.947)
White		21.373** (7.818)	19.437* (8.579)		27.393*** (7.944)	21.864** (7.739)
White × Music			16.925 (15.898)			10.395 (13.859)
SES		43.968*** (5.201)	44.054*** (5.425)		47.543*** (6.336)	46.703*** (5.223)
Square Term		-0.875 (3.790)	-1.074 (3.778)		-5.295 (5.013)	-5.310 (5.077)
SES × Music			4.157 (9.542)			2.265 (9.477)
School Engagement (9 th Grade)		2.698 (2.690)	2.818 (2.705)		4.622 (3.696)	4.863 (3.672)
C, D, or F in Science (8 th Grade)		-69.204*** (15.198)	-72.758*** (16.454)		-45.671*** (10.246)	-54.578*** (11.538)
C, D, or F in Science × Music			36.967 (23.711)			17.430 (21.153)
Intercept	486.178*** (4.182)	473.660*** (8.939)	474.999*** (9.323)	483.223*** (3.667)	470.460*** (7.294)	474.132*** (7.333)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Table E.4

Regression Results for College Examination Reading Scores for 3-4 Years of Choral Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	21.102* (10.069)	17.893 (10.347)	18.183 (19.903)	8.092 (9.341)	10.589 (9.743)	16.120 (18.311)
Male		11.954 (7.824)	11.212 (8.120)		12.636 (12.250)	-2.793 (5.987)
Male × Music			16.922 (22.772)			26.990 (22.518)
White		19.530* (8.157)	19.696* (8.540)		12.519 (10.994)	18.099* (8.329)
White × Music			-4.938 (21.563)			-10.003 (20.765)
SES		43.769*** (5.367)	44.062*** (5.435)		47.694*** (10.014)	50.293*** (7.041)
Square Term		-0.712 (3.685)	-0.788 (3.707)		-5.693 (6.865)	-5.491 (6.951)
SES × Music			-3.633 (13.533)			-5.148 (11.935)
School Engagement (9 th Grade)		1.438 (2.762)	1.484 (2.756)		-7.088 (6.014)	-6.965 (6.077)
C, D, or F in Science (8 th Grade)		-72.526*** (15.580)	-73.196*** (16.422)		-64.595*** (14.103)	-52.959*** (10.065)
C, D, or F in Science × Music			7.308 (25.768)			-19.053 (20.876)
Intercept	486.178*** (4.182)	474.742*** (9.032)	475.082*** (9.309)	501.464*** (4.133)	492.018*** (9.512)	489.137*** (8.360)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HLS:09), “Base-Year to Second Follow-Up, 2018.”

Table E.5

Regression Results for Mathematics Reasoning Scores for 1-2 Years of Choral Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	-0.166* (0.078)	-0.060 (0.047)	-0.209** (0.079)	-0.072 (0.065)	-0.044 (0.044)	-0.195* (0.080)
Male		0.041 (0.034)	0.026 (0.033)		0.064 (0.059)	0.019 (0.034)
Male × Music			0.036 (0.108)			0.084 (0.104)
White		-0.034 (0.032)	-0.056 (0.034)		0.075 (0.056)	-0.020 (0.039)
White × Music			0.218* (0.103)			0.215* (0.098)
SES		0.187*** (0.023)	0.175*** (0.023)		0.203*** (0.037)	0.167*** (0.026)
SES × Music			0.020 (0.070)			0.034 (0.067)
School Engagement (9 th Grade)		-0.006 (0.019)	-0.002 (0.019)		0.046 (0.032)	0.049 (0.028)
C, D, or F in Math (8 th Grade)		-0.237*** (0.044)	-0.250*** (0.045)		-0.206** (0.071)	-0.227*** (0.048)
C, D, or F in Math × Music			0.141 (0.158)			0.057 (0.136)
Math Reasoning Score (9 th Grade)		0.739*** (0.020)	0.751*** (0.018)		0.732*** (0.040)	0.755*** (0.022)
Square Term			0.105*** (0.012)			0.138*** (0.024)
9 th Grade Math Reasoning × Music			0.074 (0.076)			0.026 (0.068)
Intercept	0.761*** (0.030)	0.779*** (0.037)	0.706*** (0.041)	0.720*** (0.030)	0.674*** (0.047)	0.627*** (0.047)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table E.6

Regression Results for Mathematics Reasoning Scores for 3-4 Years of Choral Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	0.171* (0.086)	0.037 (0.059)	-0.075 (0.124)	0.057 (0.078)	0.023 (0.051)	-0.069 (0.109)
Male		0.045 (0.033)	0.026 (0.033)		0.089 (0.099)	-0.058 (0.068)
Male × Music			0.111 (0.191)			0.225 (0.192)
White		-0.046 (0.032)	-0.055 (0.034)		0.049 (0.065)	0.024 (0.046)
White × Music			0.101 (0.139)			0.048 (0.118)
SES		0.183*** (0.023)	0.176*** (0.023)		0.168*** (0.042)	0.198*** (0.038)
SES × Music			-0.023 (0.079)			-0.042 (0.072)
School Engagement (9 th Grade)		-0.013 (0.020)	-0.010 (0.019)		0.000 (0.043)	0.003 (0.042)
C, D, or F in Math (8 th Grade)		-0.252*** (0.044)	-0.253*** (0.045)		-0.201* (0.085)	-0.183** (0.057)
C, D, or F in Math × Music			0.036 (0.150)			-0.027 (0.136)
Math Reasoning Score (9 th Grade)		0.744*** (0.018)	0.751*** (0.018)		0.813*** (0.041)	0.745*** (0.023)
Square Term			0.098*** (0.012)			0.060* (0.030)
9 th Grade Math Reasoning × Music			0.074 (0.083)			0.077 (0.070)
Intercept	0.761*** (0.030)	0.786*** (0.037)	0.713*** (0.040)	0.889*** (0.036)	0.696*** (0.071)	0.706*** (0.055)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Instrumental Music

Table E.7

Regression Results for College Examination Mathematics Scores for 1-2 Years of Instrumental Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	22.119 (11.523)	14.208 (8.322)	3.491 (15.329)	19.848* (8.273)	15.377* (6.438)	12.132 (11.388)
Male		25.709*** (3.729)	20.993*** (4.183)		14.251* (6.698)	21.768*** (5.512)
Male × Music			-7.297 (13.948)			-17.592 (11.860)
White		14.899*** (4.208)	14.965*** (4.137)		14.289 (7.610)	15.242** (5.096)
White × Music			9.408 (12.550)			8.191 (11.688)
SES		20.051*** (3.169)	18.877*** (2.853)		15.940** (5.768)	22.157*** (3.555)
SES × Music			1.545 (15.284)			-15.944 (9.971)
School Engagement (9 th Grade)		-2.288 (2.477)	-1.904 (2.365)		-3.732 (5.494)	-1.028 (4.750)
C, D, or F in Math (8 th Grade)		-25.259*** (5.619)	-27.947*** (5.533)		-16.281 (13.841)	-30.010*** (6.247)
C, D, or F in Math × Music			39.399 (25.746)			29.116 (25.079)
Math Reasoning Score (9 th Grade)		71.866*** (2.735)	61.265*** (3.853)		70.977*** (5.419)	61.170*** (5.266)
Square Term			13.674*** (2.534)			15.234*** (3.720)
9 th Grade Math Reasoning × Male			10.922** (4.100)			8.676 (7.332)
9 th Grade Math Reasoning × Music			-0.325 (8.131)			3.120 (7.323)
Intercept	497.769*** (3.921)	457.470*** (4.223)	450.251*** (4.760)	504.822*** (3.990)	466.011*** (6.307)	450.487*** (5.630)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Table E.8

Regression Results for College Examination Mathematics Scores for 3-4 Years of Instrumental Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	38.006*** (8.370)	4.269 (5.858)	10.803 (10.911)	18.405* (7.731)	3.329 (5.506)	9.194 (10.004)
Male		26.191*** (3.861)	21.842*** (4.109)		26.223*** (5.839)	21.804*** (5.184)
Male × Music			-5.946 (9.835)			-2.570 (9.810)
White		14.555*** (3.855)	14.730*** (4.125)		11.976 (6.263)	11.643* (5.593)
White × Music			3.027 (11.959)			2.574 (11.597)
SES		19.068*** (2.675)	18.953*** (2.848)		16.950*** (4.203)	22.586*** (3.967)
SES × Music			-7.561 (6.814)			-12.302 (6.644)
School Engagement (9 th Grade)		0.115 (2.309)	-0.097 (2.286)		4.522 (3.550)	3.674 (3.480)
C, D, or F in Math (8 th Grade)		-28.763*** (5.706)	-27.247*** (5.523)		-38.644** (13.932)	-27.553*** (7.204)
C, D, or F in Math × Music			-25.818 (31.195)			-22.365 (26.129)
Math Reasoning Score (9 th Grade)		72.363*** (2.576)	62.192*** (3.896)		72.638*** (3.787)	59.908*** (4.864)
Square Term			12.587*** (2.421)			10.159*** (2.126)
9 th Grade Math Reasoning × Male			9.583* (4.009)			8.136 (5.035)
9 th Grade Math Reasoning × Music			-5.957 (5.947)			-2.249 (5.298)
Intercept	497.769*** (3.921)	457.444*** (4.047)	450.355*** (4.659)	521.463*** (4.268)	461.609*** (5.381)	455.731*** (5.438)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HLS:09), “Base-Year to Second Follow-Up, 2018.”

Table E.9

Regression Results for College Examination Reading Scores for 1-2 Years of Instrumental Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	36.943*** (10.143)	30.063*** (8.645)	41.720* (17.141)	30.214** (9.353)	23.912** (8.421)	30.629 (17.932)
Male		14.804* (7.000)	16.154* (7.485)		2.410 (9.318)	13.619* (5.840)
Male × Music			-18.631 (21.247)			-20.729 (18.576)
White		22.832** (8.019)	23.335** (8.188)		19.265 (10.599)	17.872** (6.782)
White × Music			-9.514 (19.556)			1.862 (18.161)
SES		43.595*** (4.934)	43.230*** (5.139)		53.863*** (7.190)	53.729*** (5.183)
Square Term		0.914 (3.741)	0.782 (3.702)		-5.528 (5.163)	-5.810 (5.251)
SES × Music			7.478 (10.654)			1.465 (11.733)
School Engagement (9 th Grade)		4.666 (2.762)	4.670 (2.764)		15.390** (5.763)	15.580** (5.683)
C, D, or F in Science (8 th Grade)		-70.637*** (15.843)	-71.143*** (16.050)		-50.069** (16.876)	-57.029*** (9.710)
C, D, or F in Science × Music			23.401 (46.372)			20.357 (43.475)
Intercept	485.328*** (4.121)	466.947*** (8.522)	466.167*** (8.781)	493.598*** (3.832)	474.826*** (8.817)	471.235*** (6.502)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), "Base-Year to Second Follow-Up, 2018."

Table E.10

Regression Results for College Examination Reading Scores for 3-4 Years of Instrumental Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	49.027*** (9.519)	35.228*** (8.202)	53.192** (16.143)	24.271** (8.863)	22.180** (7.716)	24.897 (13.958)
Male		13.517 (6.967)	16.091* (7.494)		3.511 (8.237)	10.695 (6.690)
Male × Music			-16.822 (14.250)			-12.547 (13.496)
White		22.199** (7.119)	23.633** (8.075)		14.463 (8.972)	7.917 (7.633)
White × Music			-12.280 (18.263)			10.621 (16.388)
SES		43.360*** (4.720)	42.629*** (5.102)		48.227*** (7.011)	52.665*** (6.572)
Square Term		2.009 (3.669)	2.147 (3.600)		3.224 (5.705)	3.217 (5.555)
SES × Music			3.580 (9.713)			-7.157 (10.179)
School Engagement (9 th Grade)		4.538 (2.732)	4.355 (2.704)		8.736 (4.502)	8.570 (4.457)
C, D, or F in Science (8 th Grade)		-73.891*** (15.155)	-71.264*** (16.005)		-78.778*** (16.406)	-59.723*** (10.666)
C, D, or F in Science × Music			-42.311 (31.568)			-38.173 (32.700)
Intercept	485.328*** (4.121)	467.840*** (7.999)	465.475*** (8.625)	512.134*** (4.196)	487.078*** (9.082)	485.726*** (6.991)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), "Base-Year to Second Follow-Up, 2018."

Table E.11

Regression Results for Mathematics Reasoning Scores for 1-2 Years of Instrumental Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	0.187* (0.082)	0.056 (0.058)	0.056 (0.108)	0.086 (0.067)	0.012 (0.049)	0.001 (0.092)
Male		0.039 (0.033)	0.041 (0.033)		-0.015 (0.056)	0.071 (0.039)
Male × Music			-0.170 (0.119)			-0.162 (0.105)
White		-0.035 (0.034)	-0.041 (0.034)		0.027 (0.063)	0.004 (0.041)
White × Music			0.116 (0.104)			0.126 (0.098)
SES		0.161*** (0.026)	0.163*** (0.026)		0.110* (0.045)	0.163*** (0.033)
SES × Music			-0.208* (0.099)			-0.163* (0.073)
School Engagement (9 th Grade)		0.002 (0.019)	0.006 (0.018)		0.003 (0.036)	0.008 (0.034)
C, D, or F in Math (8 th Grade)		-0.219*** (0.042)	-0.221*** (0.042)		-0.219** (0.079)	-0.224*** (0.047)
C, D, or F in Math × Music			-0.006 (0.144)			0.010 (0.134)
Math Reasoning Score (9 th Grade)		0.738*** (0.017)	0.752*** (0.017)		0.750*** (0.034)	0.719*** (0.024)
Square Term			0.114*** (0.012)			0.150*** (0.019)
9 th Grade Math Reasoning × Music			0.057 (0.060)			0.059 (0.057)
Intercept	0.749*** (0.029)	0.765*** (0.036)	0.663*** (0.039)	0.841*** (0.032)	0.770*** (0.056)	0.604*** (0.050)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HLS:09), “Base-Year to Second Follow-Up, 2018.”

Table E.12

Regression Results for Mathematics Reasoning Scores for 3-4 Years of Instrumental Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	0.599*** (0.057)	0.168*** (0.039)	0.070 (0.090)	0.318*** (0.056)	0.132*** (0.038)	0.068 (0.090)
Male		0.081* (0.032)	0.042 (0.033)		0.214*** (0.040)	0.058 (0.040)
Male × Music			0.223** (0.077)			0.255** (0.077)
White		-0.038 (0.033)	-0.041 (0.034)		-0.026 (0.052)	0.003 (0.048)
White × Music			0.023 (0.094)			-0.036 (0.094)
SES		0.173*** (0.023)	0.164*** (0.026)		0.151*** (0.032)	0.143*** (0.030)
SES × Music			-0.049 (0.065)			-0.017 (0.059)
School Engagement (9 th Grade)		0.004 (0.019)	0.008 (0.018)		0.038 (0.024)	0.037 (0.024)
C, D, or F in Math (8 th Grade)		-0.217*** (0.043)	-0.220*** (0.042)		-0.220*** (0.065)	-0.230*** (0.048)
C, D, or F in Math × Music			-0.012 (0.105)			0.006 (0.100)
Math Reasoning Score (9 th Grade)		0.736*** (0.018)	0.750*** (0.017)		0.762*** (0.030)	0.735*** (0.023)
Square Term			0.107*** (0.013)			0.090*** (0.019)
9 th Grade Math Reasoning × Music			-0.071 (0.057)			-0.059 (0.053)
Intercept	0.749*** (0.029)	0.744*** (0.036)	0.669*** (0.039)	1.058*** (0.035)	0.695*** (0.047)	0.671*** (0.053)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HLS:09), “Base-Year to Second Follow-Up, 2018.”

General Music

Table E.13

Regression Results for College Examination Mathematics Scores for 1-2 Years of General Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	-0.832 (7.940)	5.153 (5.335)	19.989* (9.695)	0.949 (7.049)	3.320 (5.006)	18.846 (10.179)
Male		21.758*** (4.221)	20.134*** (4.918)		18.271*** (5.244)	23.594*** (5.384)
Male × Music			-16.448 (8.940)			-20.108* (9.091)
White		11.426* (4.901)	14.127** (5.335)		10.381 (5.820)	13.689** (5.216)
White × Music			-8.035 (10.590)			-4.374 (10.185)
SES		15.919*** (3.700)	14.038*** (4.119)		18.358*** (4.359)	15.450*** (3.427)
SES × Music			4.439 (8.528)			4.610 (7.522)
School Engagement (9 th Grade)		-0.153 (2.296)	-0.041 (2.331)		0.335 (3.065)	0.700 (3.020)
C, D, or F in Math (8 th Grade)		-24.855*** (6.049)	-23.199*** (6.714)		-26.657*** (6.949)	-22.028** (7.251)
C, D, or F in Math × Music			-14.007 (11.763)			-12.288 (11.727)
Math Reasoning Score (9 th Grade)		75.083*** (2.908)	65.574*** (3.917)		74.643*** (4.058)	64.291*** (4.615)
Square Term			11.073*** (2.471)			9.588*** (2.289)
9 th Grade Math Reasoning × Male			9.480* (4.704)			11.549 (6.374)
9 th Grade Math Reasoning × Music			1.752 (8.468)			-0.674 (7.520)
Intercept	495.201*** (3.929)	461.173*** (4.645)	452.423*** (5.612)	496.235*** (3.809)	464.054*** (5.287)	452.075*** (5.099)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HLS:09), “Base-Year to Second Follow-Up, 2018.”

Table E.14

Regression Results for College Examination Reading Scores for 1-2 Years of General Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	-1.394 (6.762)	4.707 (5.841)	16.572 (13.415)	-0.981 (6.140)	2.498 (5.374)	12.434 (12.122)
Male		5.515 (10.543)	10.797 (12.506)		1.285 (10.521)	16.063 (15.111)
Male × Music			-30.168 (15.613)			-29.090 (15.484)
White		23.481*** (5.200)	23.603*** (6.229)		25.199*** (5.935)	24.014** (7.512)
White × Music			-0.551 (15.423)			2.165 (13.216)
SES		36.980*** (4.488)	37.617*** (4.574)		38.906*** (4.982)	40.157*** (4.840)
Square Term		10.880*** (2.921)	10.798*** (2.873)		11.677** (3.765)	11.725** (3.732)
SES × Music			1.795 (7.248)			0.113 (6.900)
School Engagement (9 th Grade)		2.645 (3.362)	2.501 (3.251)		2.145 (3.283)	2.032 (3.145)
C, D, or F in Science (8 th Grade)		-66.939*** (18.677)	-72.762*** (21.075)		-52.912*** (12.740)	-69.579*** (18.708)
C, D, or F in Science × Music			36.742 (24.994)			33.109 (24.036)
Intercept	484.472*** (3.711)	466.412*** (6.453)	464.588*** (7.828)	485.077*** (3.537)	466.774*** (7.157)	462.182*** (9.805)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), "Base-Year to Second Follow-Up, 2018."

Table E.15

Regression Results for Mathematics Reasoning Scores for 1-2 Years of General Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	0.003 (0.071)	0.032 (0.049)	0.084 (0.120)	0.050 (0.065)	0.038 (0.036)	0.139 (0.085)
Male		0.070 (0.052)	0.042 (0.044)		0.061 (0.055)	0.074 (0.047)
Male × Music			0.019 (0.117)			-0.062 (0.076)
White		-0.014 (0.040)	-0.001 (0.050)		-0.030 (0.043)	0.005 (0.051)
White × Music			-0.015 (0.089)			-0.032 (0.081)
SES		0.176*** (0.029)	0.149*** (0.032)		0.182*** (0.030)	0.152*** (0.031)
SES × Music			0.050 (0.063)			0.026 (0.053)
School Engagement (9 th Grade)		-0.002 (0.024)	0.006 (0.023)		0.007 (0.025)	0.014 (0.024)
C, D, or F in Math (8 th Grade)		-0.203*** (0.049)	-0.158** (0.052)		-0.278*** (0.068)	-0.160** (0.051)
C, D, or F in Math × Music			-0.242 (0.153)			-0.228 (0.125)
Math Reasoning Score (9 th Grade)		0.775*** (0.023)	0.786*** (0.025)		0.781*** (0.028)	0.773*** (0.028)
Square Term			0.104*** (0.016)			0.092*** (0.017)
9 th Grade Math Reasoning × Music			0.050 (0.069)			0.043 (0.056)
Intercept	0.714*** (0.043)	0.710*** (0.042)	0.608*** (0.054)	0.737*** (0.050)	0.749*** (0.039)	0.609*** (0.050)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Regression Tables for Social-Emotional Outcomes

Choral Music

Table E.16

Regression Results for School Motivation for 1-2 Years of Choral Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	0.049 (0.063)	-0.031 (0.059)	-0.103 (0.110)	0.009 (0.049)	-0.003 (0.048)	-0.002 (0.093)
Male		-0.158*** (0.037)	-0.161*** (0.040)		-0.168** (0.051)	-0.181*** (0.043)
Male × Music			0.040 (0.117)			0.025 (0.102)
White		-0.066 (0.039)	-0.078 (0.042)		-0.015 (0.052)	-0.028 (0.044)
White × Music			0.111 (0.111)			0.023 (0.098)
SES		-0.002 (0.025)	0.005 (0.027)		-0.038 (0.032)	-0.012 (0.027)
SES × Music			-0.064 (0.067)			-0.048 (0.065)
School Engagement (9 th Grade)		0.275*** (0.028)	0.275*** (0.028)		0.267*** (0.033)	0.266*** (0.033)
C, D, or F in Science (8 th Grade)		-0.234*** (0.059)	-0.232*** (0.062)		-0.240*** (0.066)	-0.171** (0.052)
C, D, or F in Science × Music			-0.036 (0.137)			-0.136 (0.131)
Intercept	0.082*** (0.024)	0.234*** (0.038)	0.242*** (0.040)	0.134*** (0.022)	0.201*** (0.050)	0.201*** (0.041)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table E.17

Regression Results for School Motivation for 3-4 Years of Choral Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	0.117 (0.072)	0.038 (0.070)	0.072 (0.124)	0.032 (0.060)	0.048 (0.061)	0.080 (0.104)
Male		-0.165*** (0.039)	-0.162*** (0.040)		-0.136 (0.085)	-0.150** (0.057)
Male × Music			-0.052 (0.152)			0.028 (0.115)
White		-0.075 (0.041)	-0.076 (0.042)		-0.060 (0.072)	-0.048 (0.048)
White × Music			0.010 (0.135)			-0.026 (0.115)
SES		0.003 (0.026)	0.005 (0.027)		-0.007 (0.041)	0.031 (0.040)
SES × Music			-0.041 (0.073)			-0.066 (0.068)
School Engagement (9 th Grade)		0.280*** (0.028)	0.280*** (0.028)		0.353*** (0.043)	0.353*** (0.043)
C, D, or F in Science (8 th Grade)		-0.239*** (0.059)	-0.231*** (0.063)		-0.196* (0.096)	-0.158* (0.065)
C, D, or F in Science × Music			-0.131 (0.203)			-0.070 (0.157)
Intercept	0.082*** (0.024)	0.244*** (0.039)	0.242*** (0.040)	0.179*** (0.027)	0.197** (0.064)	0.182*** (0.043)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Table E.18

Regression Results for Mathematics Self-Efficacy for 1-2 Years of Choral Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	0.004 (0.087)	0.046 (0.076)	0.086 (0.140)	0.033 (0.063)	0.037 (0.059)	0.085 (0.113)
Male		0.173*** (0.046)	0.167*** (0.049)		0.176** (0.062)	0.195*** (0.043)
Male × Music			0.040 (0.147)			-0.045 (0.135)
White		-0.139** (0.044)	-0.135** (0.045)		-0.185** (0.070)	-0.187*** (0.051)
White × Music			-0.053 (0.145)			0.003 (0.123)
SES		0.021 (0.029)	0.037 (0.031)		-0.020 (0.041)	0.042 (0.031)
SES × Music			-0.165 (0.087)			-0.105 (0.075)
School Engagement (9 th Grade)		0.057* (0.023)	0.056* (0.023)		0.140*** (0.033)	0.140*** (0.032)
C, D, or F in Math (8 th Grade)		-0.263*** (0.060)	-0.244*** (0.058)		-0.316*** (0.092)	-0.216*** (0.058)
C, D, or F in Math × Music			-0.182 (0.200)			-0.192 (0.175)
Math Self-Efficacy (9 th Grade)		0.309*** (0.021)	0.311*** (0.023)		0.270*** (0.038)	0.322*** (0.026)
9 th Grade Math Self- Efficacy × Music			-0.004 (0.083)			-0.092 (0.072)
Intercept	0.021 (0.022)	0.047 (0.049)	0.044 (0.051)	-0.021 (0.023)	0.067 (0.058)	0.042 (0.048)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table E.19

Regression Results for Mathematics Self-Efficacy for 3-4 Years of Choral Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	-0.133 (0.079)	-0.053 (0.073)	0.054 (0.141)	-0.071 (0.068)	-0.038 (0.063)	0.128 (0.140)
Male		0.150** (0.049)	0.162** (0.050)		0.089 (0.091)	0.188** (0.066)
Male $\times$ Music			-0.243 (0.185)			-0.187 (0.171)
White		-0.134** (0.044)	-0.130** (0.045)		-0.161 (0.089)	-0.070 (0.064)
White $\times$ Music			-0.061 (0.160)			-0.159 (0.155)
SES		0.046 (0.030)	0.040 (0.031)		0.122* (0.053)	0.073 (0.039)
SES $\times$ Music			0.076 (0.090)			0.076 (0.092)
School Engagement (9 th Grade)		0.037 (0.023)	0.036 (0.023)		0.026 (0.048)	0.024 (0.048)
C, D, or F in Math (8 th Grade)		-0.252*** (0.058)	-0.247*** (0.059)		-0.223 (0.132)	-0.181* (0.090)
C, D, or F in Math $\times$ Music			-0.068 (0.235)			-0.070 (0.222)
Math Self-Efficacy (9 th Grade)		0.327*** (0.022)	0.319*** (0.023)		0.396*** (0.044)	0.356*** (0.039)
9 th Grade Math Self- Efficacy $\times$ Music			0.117 (0.079)			0.074 (0.083)
Intercept	0.020 (0.023)	0.054 (0.050)	0.045 (0.051)	-0.035 (0.037)	0.065 (0.077)	-0.027 (0.062)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), "Base-Year to Second Follow-Up, 2018."

Table E.20

Regression Results for Science Self-Efficacy for 1-2 Years of Choral Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	0.065 (0.076)	0.120 (0.080)	0.237* (0.118)	0.062 (0.060)	0.069 (0.060)	0.167 (0.100)
Male		0.187*** (0.053)	0.175** (0.055)		0.183* (0.071)	0.098* (0.046)
Male × Music			0.088 (0.180)			0.146 (0.129)
White		0.021 (0.052)	0.047 (0.052)		-0.070 (0.074)	0.046 (0.059)
Male × Music			-0.277* (0.127)			-0.218 (0.117)
SES		0.061 (0.031)	0.059 (0.033)		0.087 (0.047)	0.049 (0.032)
SES × Music			0.019 (0.101)			0.066 (0.086)
School Engagement (9 th Grade)		0.031 (0.027)	0.029 (0.026)		0.050 (0.039)	0.047 (0.036)
C, D, or F in Science (8 th Grade)		-0.197* (0.083)	-0.190* (0.089)		-0.146 (0.086)	-0.095 (0.065)
C, D, or F in Science × Music			-0.053 (0.177)			-0.108 (0.161)
Science Self-Efficacy (9 th Grade)		0.224*** (0.027)	0.247*** (0.026)		0.132** (0.040)	0.228*** (0.026)
9 th Grade Science Self-Efficacy × Music			-0.202** (0.073)			-0.176** (0.062)
Intercept	-0.029 (0.025)	-0.095* (0.047)	-0.104* (0.047)	-0.046 (0.027)	-0.037 (0.060)	-0.084 (0.052)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table E.21

Regression Results for Science Self-Efficacy for 3-4 Years of Choral Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	-0.069 (0.065)	-0.042 (0.060)	0.080 (0.158)	-0.065 (0.063)	-0.074 (0.057)	0.063 (0.146)
Male		0.174** (0.053)	0.175** (0.055)		0.154 (0.090)	0.112 (0.057)
Male × Music			-0.014 (0.163)			0.084 (0.159)
White		0.040 (0.050)	0.052 (0.052)		-0.105 (0.088)	-0.007 (0.063)
White × Music			-0.146 (0.166)			-0.152 (0.157)
SES		0.066* (0.032)	0.060 (0.033)		0.130* (0.057)	0.113** (0.037)
SES × Music			0.047 (0.100)			0.016 (0.099)
School Engagement (9 th Grade)		0.019 (0.026)	0.021 (0.026)		0.019 (0.043)	0.022 (0.044)
C, D, or F in Science (8 th Grade)		-0.199* (0.085)	-0.190* (0.089)		-0.205 (0.121)	-0.052 (0.081)
C, D, or F in Science × Music			-0.099 (0.218)			-0.240 (0.224)
Science Self-Efficacy (9 th Grade)		0.256*** (0.025)	0.248*** (0.025)		0.306*** (0.050)	0.221*** (0.041)
9 th Grade Science Self-Efficacy × Music			0.106 (0.089)			0.140 (0.096)
Intercept	-0.029 (0.025)	-0.097* (0.046)	-0.107* (0.047)	-0.031 (0.033)	0.041 (0.083)	-0.050 (0.059)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Instrumental Music

Table E.22

Regression Results for School Motivation for 1-2 Years of Instrumental Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	-0.031 (0.070)	-0.097 (0.070)	-0.103 (0.084)	-0.034 (0.055)	-0.047 (0.055)	-0.013 (0.083)
Male		-0.166*** (0.038)	-0.152*** (0.040)		-0.240*** (0.055)	-0.171*** (0.039)
Male × Music			-0.208 (0.141)			-0.132 (0.105)
White		-0.079* (0.037)	-0.092* (0.039)		-0.019 (0.050)	-0.031 (0.044)
White × Music			0.132 (0.110)			0.015 (0.103)
SES		0.038 (0.025)	0.032 (0.026)		0.055 (0.040)	0.048 (0.026)
SES × Music			0.085 (0.102)			0.020 (0.074)
School Engagement (9 th Grade)		0.270*** (0.023)	0.271*** (0.023)		0.243*** (0.031)	0.243*** (0.032)
C, D, Or F in Science (8 th Grade)		-0.168** (0.055)	-0.185*** (0.055)		-0.063 (0.101)	-0.145** (0.050)
C, D, Or F in Science × Music			0.352 (0.216)			0.177 (0.192)
Intercept	0.082*** (0.021)	0.227*** (0.035)	0.230*** (0.036)	0.125*** (0.020)	0.224*** (0.045)	0.209*** (0.037)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table E.23

Regression Results for School Motivation for 3-4 Years of Instrumental Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	0.175*** (0.043)	0.104** (0.037)	0.035 (0.059)	0.096* (0.039)	0.092* (0.036)	0.089 (0.067)
Male		-0.143*** (0.037)	-0.152*** (0.039)		-0.090* (0.042)	-0.154*** (0.042)
Male × Music			0.082 (0.081)			0.116 (0.071)
White		-0.084* (0.036)	-0.094* (0.039)		0.002 (0.047)	0.020 (0.052)
White × Music			0.089 (0.086)			-0.036 (0.085)
SES		0.031 (0.023)	0.031 (0.026)		0.008 (0.029)	0.015 (0.034)
SES × Music			-0.004 (0.053)			-0.013 (0.055)
School Engagement (9 th Grade)		0.280*** (0.022)	0.279*** (0.022)		0.298*** (0.028)	0.296*** (0.027)
C, D, or F in Science (8 th Grade)		-0.196*** (0.054)	-0.179** (0.056)		-0.330*** (0.088)	-0.186** (0.057)
C, D, or F in Science × Music			-0.318 (0.195)			-0.279 (0.192)
Intercept	0.082*** (0.021)	0.222*** (0.035)	0.229*** (0.037)	0.176*** (0.021)	0.177*** (0.040)	0.180*** (0.047)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table E.24

Regression Results for Mathematics Self-Efficacy for 1-2 Years of Instrumental Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	0.087 (0.068)	0.044 (0.063)	-0.151 (0.124)	0.061 (0.053)	0.039 (0.049)	-0.140 (0.101)
Male		0.234*** (0.045)	0.219*** (0.047)		0.364*** (0.057)	0.250*** (0.043)
Male × Music			0.195 (0.125)			0.212* (0.103)
White		-0.118** (0.038)	-0.121** (0.041)		-0.115* (0.058)	-0.150*** (0.039)
White × Music			0.057 (0.121)			0.076 (0.107)
SES		0.072** (0.026)	0.060* (0.028)		0.078 (0.043)	0.056* (0.027)
SES × Music			0.095 (0.097)			0.030 (0.081)
School Engagement (9 th Grade)		0.033 (0.023)	0.035 (0.023)		0.042 (0.035)	0.045 (0.035)
C, D, or F in Math (8 th Grade)		-0.236*** (0.051)	-0.257*** (0.053)		-0.166 (0.090)	-0.218*** (0.050)
C, D, or F in Math × Music			0.271 (0.202)			0.088 (0.168)
Math Self-Efficacy (9 th Grade)		0.325*** (0.021)	0.317*** (0.021)		0.370*** (0.031)	0.321*** (0.023)
9 th Grade Math Self- Efficacy × Music			0.087 (0.055)			0.088 (0.053)
Intercept	0.018 (0.022)	-0.005 (0.047)	0.009 (0.050)	0.063** (0.022)	-0.070 (0.049)	0.023 (0.043)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table E.25

Regression Results for Mathematics Self-Efficacy for 3-4 Years of Instrumental Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	0.070 (0.071)	-0.005 (0.060)	-0.247 (0.160)	0.021 (0.054)	-0.010 (0.048)	-0.166 (0.105)
Male		0.235*** (0.044)	0.213*** (0.047)		0.301*** (0.052)	0.239*** (0.048)
Male × Music			0.174 (0.104)			0.105 (0.102)
White		-0.100** (0.037)	-0.119** (0.041)		-0.052 (0.058)	-0.122* (0.050)
White × Music			0.163 (0.131)			0.124 (0.105)
SES		0.076** (0.025)	0.058* (0.029)		0.115*** (0.031)	0.085** (0.031)
SES × Music			0.127 (0.070)			0.051 (0.060)
School Engagement (9 th Grade)		0.037 (0.023)	0.038 (0.022)		0.051 (0.032)	0.052 (0.031)
C, D, or F in Math (8 th Grade)		-0.240*** (0.053)	-0.257*** (0.053)		-0.188** (0.070)	-0.247*** (0.065)
C, D, or F in Math × Music			0.171 (0.137)			0.103 (0.138)
Math Self-Efficacy (9 th Grade)		0.323*** (0.020)	0.314*** (0.021)		0.362*** (0.030)	0.362*** (0.028)
9 th Grade Math Self- Efficacy × Music			0.043 (0.055)			-0.006 (0.057)
Intercept	0.017 (0.023)	-0.015 (0.045)	0.011 (0.050)	0.125*** (0.027)	-0.058 (0.054)	0.031 (0.054)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table E.26

Regression Results for Science Self-Efficacy for 1-2 Years of Instrumental Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	0.137 (0.075)	0.076 (0.069)	-0.079 (0.152)	0.109 (0.062)	0.073 (0.060)	-0.098 (0.119)
Male		0.213*** (0.049)	0.197*** (0.052)		0.282*** (0.069)	0.164*** (0.044)
Male × Music			0.201 (0.137)			0.224 (0.120)
White		0.066 (0.046)	0.059 (0.049)		0.094 (0.069)	0.033 (0.046)
White × Music			0.075 (0.150)			0.114 (0.126)
SES		0.079** (0.030)	0.078* (0.031)		0.036 (0.049)	0.068* (0.029)
SES × Music			-0.003 (0.103)			-0.067 (0.087)
School Engagement (9 th Grade)		0.048 (0.024)	0.049* (0.024)		0.039 (0.038)	0.042 (0.038)
C, D, or F in Science (8 th Grade)		-0.153 (0.089)	-0.162 (0.092)		-0.061 (0.099)	-0.070 (0.069)
C, D, or F in Science × Music			0.156 (0.218)			-0.001 (0.204)
Science Self-Efficacy (9 th Grade)		0.238*** (0.024)	0.240*** (0.025)		0.234*** (0.038)	0.245*** (0.028)
9 th Grade Science Self-Efficacy × Music			-0.031 (0.080)			-0.018 (0.072)
Intercept	-0.016 (0.025)	-0.137** (0.046)	-0.124* (0.048)	0.028 (0.025)	-0.179** (0.057)	-0.087 (0.045)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HLS:09), “Base-Year to Second Follow-Up, 2018.”

Table E.27

Regression Results for Science Self-Efficacy for 3-4 Years of Instrumental Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	0.249*** (0.061)	0.139* (0.055)	0.101 (0.129)	0.167** (0.052)	0.119* (0.049)	0.083 (0.101)
Male		0.201*** (0.047)	0.195*** (0.052)		0.201*** (0.051)	0.169*** (0.049)
Male × Music			0.040 (0.111)			0.055 (0.100)
White		0.054 (0.047)	0.059 (0.049)		0.017 (0.059)	0.017 (0.053)
White × Music			-0.035 (0.114)			0.000 (0.101)
SES		0.087** (0.028)	0.078* (0.031)		0.116** (0.037)	0.105** (0.034)
SES × Music			0.058 (0.070)			0.018 (0.063)
School Engagement (9 th Grade)		0.047* (0.023)	0.048* (0.023)		0.038 (0.031)	0.038 (0.031)
C, D, or F in Science (8 th Grade)		-0.149 (0.089)	-0.163 (0.093)		-0.044 (0.097)	-0.046 (0.065)
C, D, or F in Science × Music			0.210 (0.205)			0.002 (0.181)
Science Self-Efficacy (9 th Grade)		0.243*** (0.022)	0.238*** (0.023)		0.268*** (0.031)	0.257*** (0.027)
9 th Grade Science Self-Efficacy × Music			0.046 (0.064)			0.022 (0.057)
Intercept	-0.016 (0.025)	-0.124** (0.045)	-0.121* (0.048)	0.078** (0.028)	-0.098 (0.057)	-0.078 (0.052)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

General Music

Table E.28

Regression Results for School Motivation for 1-2 Years of General Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	-0.015 (0.069)	-0.044 (0.060)	-0.122 (0.131)	-0.034 (0.056)	-0.033 (0.047)	-0.051 (0.094)
Male		-0.123** (0.037)	-0.155*** (0.046)		-0.097* (0.045)	-0.162** (0.051)
Male × Music			0.149 (0.113)			0.115 (0.094)
White		-0.060 (0.041)	-0.071 (0.045)		-0.023 (0.054)	-0.011 (0.048)
White × Music			0.033 (0.115)			-0.030 (0.103)
SES		0.047 (0.029)	0.022 (0.029)		0.068* (0.031)	0.009 (0.028)
SES × Music			0.111* (0.055)			0.103 (0.056)
School Engagement (9 th Grade)		0.278*** (0.022)	0.281*** (0.022)		0.294*** (0.031)	0.296*** (0.031)
C, D, or F in Science (8 th Grade)		-0.220** (0.071)	-0.216** (0.080)		-0.237** (0.081)	-0.173* (0.077)
C, D, or F in Science × Music			-0.046 (0.151)			-0.128 (0.152)
Intercept	0.085*** (0.023)	0.208*** (0.034)	0.229*** (0.040)	0.103*** (0.024)	0.178*** (0.042)	0.190*** (0.038)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table E.29

Regression Results for Mathematics Self-Efficacy for 1-2 Years of General Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	0.019 (0.061)	-0.017 (0.058)	-0.111 (0.121)	0.035 (0.058)	0.003 (0.059)	-0.029 (0.128)
Male		0.248*** (0.044)	0.226*** (0.051)		0.275*** (0.060)	0.269*** (0.062)
Male × Music			0.097 (0.109)			0.007 (0.112)
White		-0.083 (0.044)	-0.104* (0.051)		-0.049 (0.060)	-0.083 (0.048)
White × Music			0.095 (0.122)			0.055 (0.121)
SES		0.047 (0.033)	0.035 (0.037)		0.063 (0.040)	0.046 (0.034)
SES × Music			0.040 (0.075)			0.027 (0.075)
School Engagement (9 th Grade)		0.016 (0.029)	0.018 (0.029)		0.032 (0.039)	0.031 (0.039)
C, D, or F in Math (8 th Grade)		-0.288*** (0.064)	-0.288*** (0.067)		-0.275*** (0.067)	-0.294*** (0.053)
C, D, or F in Math × Music			0.009 (0.155)			0.036 (0.127)
Math Self-Efficacy (9 th Grade)		0.303*** (0.023)	0.312*** (0.026)		0.295*** (0.032)	0.329*** (0.024)
9 th Grade Math Self- Efficacy × Music			-0.039 (0.065)			-0.060 (0.065)
Intercept	0.013 (0.026)	-0.020 (0.049)	0.003 (0.056)	0.032 (0.025)	-0.057 (0.052)	-0.034 (0.054)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table E.30

Regression Results for Science Self-Efficacy for 1-2 Years of General Music

	Unadjusted Model 1	Unadjusted Model 2	Unadjusted Model 3	PS Adjusted Model 1	PS Adjusted Model 2	PS Adjusted Model 3
Music	0.099 (0.070)	0.081 (0.060)	0.097 (0.107)	0.186 (0.273)	0.158 (0.147)	0.086 (0.223)
Male		0.198*** (0.047)	0.204** (0.062)		0.353*** (0.088)	0.391* (0.186)
Male × Music			-0.028 (0.142)			-0.060 (0.331)
White		0.068 (0.047)	0.069 (0.047)		0.143 (0.090)	0.134 (0.089)
White × Music			-8.801 (13.096)			-4.511 (17.881)
SES		0.067 (0.038)	0.076 (0.040)		-0.106 (0.104)	-0.044 (0.111)
SES × Music			-0.044 (0.071)			-0.084 (0.146)
School Engagement (9 th Grade)		0.031 (0.031)	0.030 (0.031)		0.006 (0.095)	-0.001 (0.089)
C, D, or F in Science (8 th Grade)		-0.185 (0.109)	-0.180 (0.125)		-0.406 (0.363)	-0.600 (0.517)
C, D, or F in Science × Music			-0.029 (0.162)			0.374 (0.431)
Science Self-Efficacy (9 th Grade)		0.233*** (0.025)	0.234*** (0.029)		0.274*** (0.070)	0.215*** (0.056)
9 th Grade Science Self-Efficacy × Music			-0.010 (0.078)			0.109 (0.162)
Intercept	-0.031 (0.028)	-0.132** (0.044)	-0.137** (0.047)	-0.153 (0.106)	-0.287*** (0.072)	-0.241** (0.080)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Regression Tables for Base-Year Social-Emotional Variables

Table E.31

Regression Results for 9th Grade School Belongingness for 1-2 Years of Choral Music

	Model 1	Model 2	Model 3
Music	0.013 (0.076)	0.012 (0.076)	0.071 (0.140)
Male		-0.018 (0.050)	-0.005 (0.053)
Male × Music			-0.166 (0.176)
White		0.103 (0.061)	0.099 (0.067)
White × Music			0.042 (0.151)
SES		0.086** (0.030)	0.079* (0.031)
SES × Music			0.077 (0.114)
C, D, or F in Math (8 th Grade)		-0.264*** (0.053)	-0.256*** (0.056)
C, D, or F in Math × Music			-0.103 (0.177)
Intercept	-0.006 (0.027)	0.013 (0.071)	0.006 (0.076)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSLs:09), “Base-Year to Second Follow-Up, 2018.”

Table E.32

Regression Results for 9th Grade School Engagement for 1-2 Years of Choral Music

	Model 1	Model 2	Model 3
Music	0.129* (0.059)	0.098 (0.060)	0.101 (0.104)
Male		-0.146*** (0.043)	-0.158*** (0.045)
Male × Music			0.143 (0.137)
White		0.149** (0.046)	0.161** (0.049)
White × Music			-0.120 (0.118)
SES		0.102*** (0.028)	0.107*** (0.030)
SES × Music			-0.057 (0.096)
C, D, or F in Math (8 th Grade)		-0.396*** (0.069)	-0.398*** (0.074)
C, D, or F in Math × Music			0.049 (0.169)
Intercept	0.053 (0.030)	0.145** (0.049)	0.146** (0.052)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table E.33

Regression Results for 9th Grade School Belongingness for 3-4 Years of Choral Music

	Model 1	Model 2	Model 3
Music	0.235*** (0.060)	0.193** (0.060)	0.221 (0.150)
Male		-0.011 (0.050)	-0.005 (0.053)
Male × Music			-0.124 (0.167)
White		0.098 (0.064)	0.099 (0.067)
White × Music			-0.052 (0.156)
SES		0.077** (0.028)	0.079* (0.031)
SES × Music			-0.021 (0.093)
C, D, or F in Math (8 th Grade)		-0.243*** (0.052)	-0.256*** (0.056)
C, D, or F in Math × Music			0.194 (0.201)
Intercept	-0.006 (0.027)	0.006 (0.073)	0.006 (0.076)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table E.34

Regression Results for 9th Grade School Engagement for 3-4 Years of Choral Music

	Model 1	Model 2	Model 3
Music	0.125 (0.070)	0.012 (0.068)	-0.025 (0.152)
Male		-0.168*** (0.045)	-0.158*** (0.045)
Male × Music			-0.189 (0.189)
White		0.173*** (0.048)	0.161** (0.049)
White × Music			0.163 (0.163)
SES		0.100*** (0.028)	0.107*** (0.030)
SES × Music			-0.065 (0.094)
C, D, or F in Math (8 th Grade)		-0.406*** (0.074)	-0.398*** (0.074)
C, D, or F in Math × Music			-0.133 (0.193)
Intercept	0.053 (0.030)	0.146** (0.050)	0.146** (0.052)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table E.35

Regression Results for 9th Grade School Belongingness for 1-2 Years of Instrumental Music

	Model 1	Model 2	Model 3
Music	0.037 (0.056)	0.027 (0.057)	0.240* (0.114)
Male		-0.022 (0.049)	-0.011 (0.052)
Male × Music			-0.145 (0.126)
White		0.026 (0.062)	0.049 (0.065)
White × Music			-0.295* (0.123)
SES		0.078** (0.028)	0.079* (0.031)
SES × Music			0.003 (0.070)
C, D, or F in Math (8 th Grade)		-0.242*** (0.048)	-0.249*** (0.051)
C, D, or F in Math × Music			0.098 (0.138)
Intercept	0.003 (0.026)	0.058 (0.071)	0.041 (0.074)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table E.36

Regression Results for 9th Grade School Engagement for 1-2 Years of Instrumental Music

	Model 1	Model 2	Model 3
Music	0.162* (0.067)	0.142* (0.066)	0.376** (0.128)
Male		-0.185*** (0.041)	-0.168*** (0.042)
Male × Music			-0.227 (0.133)
White		0.095* (0.043)	0.111* (0.045)
White × Music			-0.225 (0.129)
SES		0.078** (0.028)	0.073* (0.030)
SES × Music			0.077 (0.095)
C, D, or F in Math (8 th Grade)		-0.404*** (0.063)	-0.406*** (0.067)
C, D, or F in Math × Music			-0.000 (0.176)
Intercept	0.072** (0.025)	0.213*** (0.045)	0.195*** (0.047)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table E.37

Regression Results for 9th Grade School Belongingness for 3-4 Years of Instrumental Music

	Model 1	Model 2	Model 3
Music	0.263*** (0.052)	0.207*** (0.052)	0.195 (0.103)
Male		-0.019 (0.047)	-0.011 (0.052)
Male × Music			-0.078 (0.097)
White		0.055 (0.059)	0.049 (0.065)
White × Music			0.074 (0.115)
SES		0.086** (0.028)	0.079* (0.031)
SES × Music			0.053 (0.059)
C, D, or F in Math (8 th Grade)		-0.251*** (0.048)	-0.249*** (0.051)
C, D, or F in Math × Music			-0.050 (0.164)
Intercept	0.003 (0.026)	0.043 (0.068)	0.041 (0.074)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table E.38

Regression Results for 9th Grade School Engagement for 3-4 Years of Instrumental Music

	Model 1	Model 2	Model 3
Music	0.171** (0.052)	0.087 (0.050)	0.069 (0.101)
Male		-0.170*** (0.039)	-0.168*** (0.042)
Male × Music			-0.014 (0.097)
White		0.118** (0.041)	0.111* (0.045)
White × Music			0.065 (0.106)
SES		0.072** (0.027)	0.073* (0.030)
SES × Music			-0.016 (0.061)
C, D, or F in Math (8 th Grade)		-0.414*** (0.064)	-0.406*** (0.067)
C, D, or F in Math × Music			-0.126 (0.208)
Intercept	0.072** (0.025)	0.194*** (0.043)	0.195*** (0.047)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table E.39

Regression Results for 9th Grade School Belongingness for 1-2 Years of General Music

	Model 1	Model 2	Model 3
Music	0.056 (0.078)	0.058 (0.072)	0.216 (0.188)
Male		0.006 (0.048)	0.029 (0.060)
Male × Music			-0.096 (0.128)
White		0.056 (0.066)	0.086 (0.087)
White × Music			-0.132 (0.156)
SES		0.085* (0.039)	0.117* (0.046)
SES × Music			-0.142 (0.085)
C, D, or F in Math (8 th Grade)		-0.289*** (0.053)	-0.240*** (0.059)
C, D, or F in Math × Music			-0.264 (0.160)
Intercept	0.002 (0.030)	0.035 (0.073)	-0.004 (0.092)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table E.40

Regression Results for 9th Grade School Engagement for 1-2 Years of General Music

	Model 1	Model 2	Model 3
Music	0.091 (0.058)	0.101 (0.065)	0.263* (0.103)
Male		-0.256*** (0.046)	-0.195*** (0.048)
Male × Music			-0.256** (0.094)
White		0.107* (0.046)	0.147** (0.054)
White × Music			-0.195 (0.126)
SES		0.091** (0.032)	0.112** (0.036)
SES × Music			-0.081 (0.071)
C, D, or F in Math (8 th Grade)		-0.415*** (0.095)	-0.474*** (0.086)
C, D, or F in Math × Music			0.325* (0.147)
Intercept	0.066* (0.029)	0.237*** (0.045)	0.197*** (0.051)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table E.41

Regression Results for 9th Grade School Engagement for 3-4 and 1-2 Years of Choral Music

	Model 1	Model 2	Model 3
3-4 Years (vs. 1-2 Years)	-0.004 (0.082)	-0.058 (0.084)	-0.127 (0.173)
Male		-0.140 (0.120)	-0.015 (0.131)
Male × Music			-0.332 (0.223)
White		0.144 (0.097)	0.040 (0.110)
White × Music			0.283 (0.189)
SES		0.038 (0.063)	0.050 (0.090)
SES × Music			-0.008 (0.133)
C, D, or F in Math (8 th Grade)		-0.427*** (0.120)	-0.349* (0.152)
C, D, or F in Math × Music			-0.183 (0.270)
Intercept	0.181*** (0.051)	0.244** (0.078)	0.247** (0.090)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table E.42

Regression Results for 9th Grade School Belongingness for 3-4 and 1-2 Years of Choral Music

	Model 1	Model 2	Model 3
3-4 Years (vs. 1-2 Years)	0.222* (0.096)	0.156 (0.080)	0.150 (0.185)
Male		-0.155 (0.124)	-0.171 (0.164)
Male × Music			0.042 (0.219)
White		0.115 (0.096)	0.142 (0.132)
White × Music			-0.094 (0.204)
SES		0.116 (0.087)	0.156 (0.111)
SES × Music			-0.098 (0.127)
C, D, or F in Math (8 th Grade)		-0.236 (0.138)	-0.359* (0.169)
C, D, or F in Math × Music			0.298 (0.231)
Intercept	0.007 (0.071)	0.055 (0.097)	0.077 (0.115)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table E.43

Regression Results for 9th Grade School Engagement for 3-4 and 1-2 Years of Instrumental Music

	Model 1	Model 2	Model 3
3-4 Years (vs. 1-2 Years)	0.009 (0.082)	-0.057 (0.081)	-0.306 (0.165)
Male		-0.264*** (0.075)	-0.396** (0.127)
Male × Music			0.214 (0.151)
White		0.062 (0.071)	-0.113 (0.122)
White × Music			0.290 (0.164)
SES		0.093 (0.050)	0.150 (0.088)
SES × Music			-0.093 (0.102)
C, D, or F in Math (8 th Grade)		-0.447*** (0.119)	-0.407* (0.166)
C, D, or F in Math × Music			-0.126 (0.274)
Intercept	0.234*** (0.066)	0.420*** (0.084)	0.571*** (0.122)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”

Table E.44

Regression Results for 9th Grade School Belongingness for 3-4 and 1-2 Years of Instrumental Music

	Model 1	Model 2	Model 3
3-4 Years (vs. 1-2 Years)	0.226** (0.071)	0.188* (0.075)	-0.045 (0.112)
Male		-0.114 (0.059)	-0.156 (0.112)
Male × Music			0.067 (0.150)
White		-0.032 (0.077)	-0.246* (0.108)
White × Music			0.368** (0.137)
SES		0.113** (0.036)	0.082 (0.062)
SES × Music			0.050 (0.091)
C, D, or F in Math (8 th Grade)		-0.209 (0.107)	-0.151 (0.125)
C, D, or F in Math × Music			-0.148 (0.187)
Intercept	0.040 (0.051)	0.156* (0.066)	0.281** (0.092)

* $p < .05$; ** $p < .01$; *** $p < .001$

SOURCE: U.S. Department of Education, National Center for Education Statistics, High School Longitudinal Study of 2009 (HSL:09), “Base-Year to Second Follow-Up, 2018.”