


January 1st, 2016

A Meta-Analysis of Middle School Science Engagement

Leanna B. Aker
Seattle Pacific University

Follow this and additional works at: https://digitalcommons.spu.edu/soe_etd

 Part of the [Educational Methods Commons](#), and the [Science and Mathematics Education Commons](#)

Recommended Citation

Aker, Leanna B., "A Meta-Analysis of Middle School Science Engagement" (2016). *Education Dissertations*. 10.
https://digitalcommons.spu.edu/soe_etd/10

This Dissertation is brought to you for free and open access by the Education, School of at Digital Commons @ SPU. It has been accepted for inclusion in Education Dissertations by an authorized administrator of Digital Commons @ SPU.

A Meta-Analysis of Middle School Science Engagement

by

Leanna B. Aker

Dissertation presented to the
Faculty of the Graduate School of Education at
Seattle Pacific University
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy in Education

April 2016

A Meta-Analysis of Middle School Science Engagement

By LEANNA B. AKER

A dissertation submitted in partial fulfillment

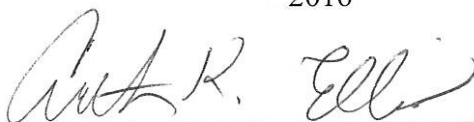
Of the requirements for the degree of

Doctor of Education


Seattle Pacific University

2016


Approved by



(ARTHUR K. ELLIS, EdD, Chairperson of the Dissertation Committee)



(STAMATIS VOKOS, PhD)



(RICK EIGENBROOD, PhD)

Program Authorized to Offer Degree

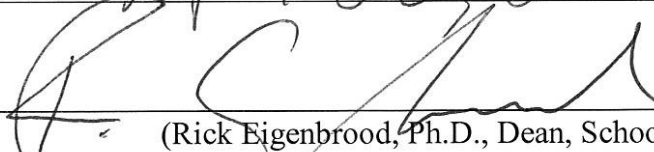


School of Education

Date



April 28, 2016



(Rick Eigenbrood, Ph.D., Dean, School of Education)

Copyright Page

In presenting this dissertation in partial fulfillment of the requirements for the Doctor of Philosophy in Education degree at Seattle Pacific University, I agree that the library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with "Fair Use" as prescribed in U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to University Microfilms, 1490 Eisenhower Place, PO Box 975, Ann Arbor, MI 48106, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microfilm and/or (b) printed copies of the manuscript from microfilm."

Signature SeanroB. Allen

Date 5/27/14

Table of Contents

| | |
|---|-----------|
| Chapter 1: Introduction | 2 |
| Background..... | 2 |
| Purpose of the Study | 4 |
| Significance of the Study | 4 |
| Definitions..... | 7 |
| Engagement..... | 7 |
| Behavioral Engagement..... | 7 |
| Affective Engagement..... | 7 |
| Cognitive Engagement..... | 8 |
| Research Questions and Null Hypotheses | 9 |
| Research Questions..... | 9 |
| Null Hypotheses..... | 10 |
| Content of the Following Chapters | 10 |
| Chapter 2: Literature Review..... | 11 |
| Engagement..... | 11 |
| Origins and Historical Evolution of the Construct | 12 |
| Accepted Model of Engagement..... | 17 |
| Related Constructs | 19 |
| Flow..... | 20 |
| Motivation..... | 21 |
| Behavioral Intent..... | 23 |
| Situational Interest..... | 24 |

| | |
|--|-----------|
| Theoretical Frameworks | 25 |
| Self-Determination Theory | 26 |
| Stage-Environment Fit Theory | 28 |
| Measurement of Engagement | 30 |
| Psychometric Instruments | 30 |
| Validity and Reliability | 33 |
| Grain Size | 37 |
| Review of Engagement Research | 38 |
| Meta-Analyses | 39 |
| Kumar Study | 39 |
| Links Between Engagement and Achievement | 41 |
| Chang et al. Study | 41 |
| Singh et al. Study | 43 |
| Predictors of Engagement | 44 |
| Uekawa et al. Study | 45 |
| Lau and Roeser Study | 47 |
| Assor et al. Study | 49 |
| Qualitative Research | 52 |
| Olitsky Study | 52 |
| Raphael et al. Study | 54 |
| Logan and Skamp Study | 56 |
| Summary | 57 |
| Chapter 3: Research Methods | 59 |

| | |
|--|----|
| Criticisms and Limitations of Meta-Analysis..... | 60 |
| Literature Search Methods..... | 64 |
| Inclusion/Exclusion Criteria | 67 |
| Source Characteristics..... | 67 |
| Study Characteristics. | 68 |
| Content Area and Age Range. | 70 |
| Instrumentation | 70 |
| Methodology and Experimental Design. | 70 |
| Statistical Considerations..... | 72 |
| Study Characteristics and Coding..... | 73 |
| Source Characteristics..... | 73 |
| Study Characteristics. | 73 |
| Predictor Classification: Type..... | 73 |
| Predictor Classification: Self-determination theory..... | 74 |
| Engagement Conceptualization. | 74 |
| Instrumentation Reliability and Validity | 75 |
| Methodology and Experimental Design. | 76 |
| Research Synthesis Methods..... | 76 |
| Calculations for Individual Studies..... | 77 |
| Effect Size Calculation | 77 |
| Converting Between Effect Sizes. | 79 |
| Variance..... | 80 |
| Confidence Intervals | 80 |

| | |
|--|-----------|
| Complex Data Structures..... | 80 |
| Independent Sub-Groups..... | 81 |
| Non-independent Subgroups..... | 82 |
| Synthesis of Multiple Studies..... | 84 |
| Statistical Model..... | 84 |
| Summary statistics..... | 85 |
| Issues of Precision and Variance..... | 87 |
| Meta-regression..... | 90 |
| Publication Bias Analysis..... | 91 |
| Limitations and Delimitations of the Study..... | 92 |
| Summary..... | 94 |
| Chapter 4: Results..... | 95 |
| Descriptive Statistics..... | 95 |
| Source Characteristics..... | 95 |
| Study Characteristics..... | 95 |
| Summary Effect Size..... | 98 |
| Effect Sizes for Individual Studies..... | 99 |
| Research Questions..... | 100 |
| Research Question 1: Moderators of Engagement..... | 100 |
| Publication Status..... | 100 |
| Peer Review Status..... | 102 |
| School Structure..... | 104 |
| School Type..... | 105 |

| | |
|---|-----|
| School Setting | 106 |
| Socioeconomic Status | 110 |
| Geographic Location..... | 110 |
| Instrument Validity..... | 114 |
| Instrument Reliability..... | 115 |
| Repeat Authors..... | 117 |
| Summary..... | 119 |
| Research Question 2: Practically Significant Predictors of Engagement | 120 |
| Summary..... | 122 |
| Research Question 3: Commonalities in Practically Significant Predictors..... | 124 |
| Descriptive Statistics..... | 124 |
| Inferential Statistics | 125 |
| Predictor Classification: Type..... | 125 |
| Predictor Classification: Self-Determination Theory | 127 |
| Combined Predictor/Moderator Models..... | 129 |
| Summary..... | 130 |
| Research Questions 4 and 5: Predictors of Engagement Types..... | 133 |
| Affective Engagement Predictors..... | 134 |
| Cognitive Engagement Predictors..... | 144 |
| Behavioral Engagement Predictors..... | 150 |
| Summary..... | 153 |
| Research Questions 6 & 7: Underrepresented Engagement Predictors/Types ... | 155 |
| Publication Bias Analysis | 157 |

| | |
|--|------------|
| Chapter 5: Summary of Findings..... | 160 |
| Research Question 1: Moderators of Engagement..... | 160 |
| Statistically Significant Moderators..... | 160 |
| Publication Status..... | 160 |
| Geographic Location..... | 162 |
| School Setting..... | 163 |
| Instrument Reliability..... | 164 |
| Statistically Nonsignificant Moderators..... | 166 |
| Peer-review Status..... | 166 |
| Study Methodology..... | 167 |
| Repeat Authors..... | 168 |
| Moderators Not Analyzed via Meta-regression..... | 168 |
| School Structure..... | 168 |
| School Type..... | 169 |
| Instrument Validity..... | 169 |
| Socioeconomic Status..... | 171 |
| Research Question 2 & 3: Commonalities in Engagement Predictors..... | 172 |
| Research Question 4 & 5: Commonalities in Affective Engagement Predictors..... | 178 |
| Affective Engagement..... | 178 |
| Cognitive Engagement..... | 180 |
| Behavioral Engagement..... | 182 |
| Summary..... | 183 |
| Research Questions 6 & 7: Underrepresented Engagement Predictors/Types ... | 183 |

| | |
|-----------------------------|------------|
| Engagement Types..... | 183 |
| Engagement Predictors. | 186 |
| Recommendations..... | 187 |
| Conclusion | 190 |
| References..... | 194 |
| Appendices..... | 218 |

List of Figures

| | |
|--|-----|
| Figure 1. Stem and leaf plot of sample sizes from 76 studies..... | 96 |
| Figure 2. Stem and leaf plot of 158 point estimates from 79 studies..... | 99 |
| Figure 3. Regression of point estimates on publication status..... | 102 |
| Figure 4. Regression of point estimates on peer review status..... | 104 |
| Figure 5. Regression of point estimates on school location (without rural)..... | 108 |
| Figure 6. Regression of point estimates on school setting..... | 109 |
| Figure 7. Regression of point estimates on geographic location | 112 |
| Figure 8. Regression of point estimates on study methodology | 114 |
| Figure 9. Regression of point estimates on instrument reliability | 117 |
| Figure 10. Regression of point estimates on author..... | 119 |
| Figure 11. Forest plot of 51 engagement effect sizes (ES) with Hedges' $g > .41$ | 123 |
| Figure 12. Regression of point estimates on predictor classification: type | 127 |
| Figure 13. Regression of point estimates on predictor classification: SDT..... | 129 |
| Figure 14. Forest plot of 28 affective engagement ES with Hedges' $g >$ than $.41$ | 136 |
| Figure 15. Regression of affective point estimates on predictor classification: type | 139 |
| Figure 16. Regression of affective point estimates on predictor classification: SDT..... | 142 |
| Figure 17. Forest plot of 12 cognitive engagement ES with Hedges' $g >$ than $.41$ | 146 |
| Figure 18. Regression of cognitive point estimates on predictor classification: type | 148 |
| Figure 19. Regression of cognitive point estimates on predictor classification: SDT.... | 150 |
| Figure 20. Forest plot of 10 behavioral engagement effect sizes | 152 |
| Figure 21. Funnel plot of 79 studies | 158 |
| Figure 22. Funnel plot of 79 studies including ten imputed studies right of the mean... | 159 |

List of Tables

| | |
|---|-----|
| Table 1. Engagement Search Terms by Database..... | 65 |
| Table 2. Engagement Assessment Instruments Included in the Literature Search..... | 66 |
| Table 3. Engagement Type Indicators | 69 |
| Table 4. Correlation Assumptions for Combining Variances..... | 84 |
| Table 5. Descriptive Statistics for Predictor Classification | 97 |
| Table 6. Descriptive Statistics for Engagement Outcomes..... | 98 |
| Table 7. Effect Sizes and Null Tests for Publication Status | 101 |
| Table 8. Meta-regression Model for Publication Status | 101 |
| Table 9. Effect Sizes and Null Tests for Peer Review Status | 103 |
| Table 10. Meta-regression Model for Peer Review Status | 103 |
| Table 11. Effect Sizes and Null Tests for School Structure | 105 |
| Table 12. Effect Sizes and Null Tests for School Type..... | 106 |
| Table 13. Effect Sizes and Null Tests for School Setting..... | 106 |
| Table 14. Meta-regression Model for School Setting (Without Rural) | 107 |
| Table 15. Meta-regression Model for School Setting (With Rural) | 109 |
| Table 16. Effect Sizes and Null Tests for Socioeconomic Status..... | 110 |
| Table 17. Effect Sizes and Null Tests for Geographic Location | 111 |
| Table 18. Meta-regression Model for Geographic Location | 111 |
| Table 19. Effect Sizes and Null Tests for Study Methodology | 112 |
| Table 20. Meta-regression Model for Study Methodology | 113 |
| Table 21. Effect Sizes and Null Tests for Instrument Validity..... | 115 |
| Table 22. Effect Sizes and Null Tests for Instrument Reliability..... | 115 |

| | |
|---|-----|
| Table 23. Meta-regression Model for Instrument Reliability | 116 |
| Table 24. Effect Sizes and Null Tests for Authors | 118 |
| Table 25. Meta-regression Model for Authors | 118 |
| Table 26. Summary of Effect Sizes and Regression Models for Moderators..... | 120 |
| Table 27. Distribution of Point Estimates..... | 121 |
| Table 28. Distribution of Point Estimates by Predictor Classification | 125 |
| Table 29. Effect Sizes and Null Tests for Predictor Classification: Type | 126 |
| Table 30. Meta-regression Model for Predictor Classification: Type | 126 |
| Table 31. Effect Sizes and Null Tests for Predictor Classification: SDT..... | 128 |
| Table 32. Meta-regression Model for Predictor Classification: SDT | 129 |
| Table 33. Meta-regression Models for Combined Moderators and Predictors: Type | 131 |
| Table 34. Meta-regression Models for Combined Moderators and Predictors: SDT | 132 |
| Table 35. Effect Sizes and Null Tests for Engagement Type | 134 |
| Table 36. Distribution of Affective Point Estimates by Predictor Classification..... | 135 |
| Table 37. Effect Sizes & Null Tests for Affective Engagement by Predictor Classification: Type | 138 |
| Table 38. Meta-regression Models for Affective Predictor Classification: Type..... | 138 |
| Table 39. Meta-regression Models for Combined Moderators and Affective Predictors: Type | 141 |
| Table 40. Effect Sizes and Null Tests for Affective Engagement by Predictor Classification: SDT..... | 142 |
| Table 41. Meta-regression Model for Affective Predictor Classification: SDT..... | 142 |

| | |
|--|-----|
| Table 42. Meta-regression Models for Combined Moderators and Affective Predictors: SDT | 143 |
| Table 43. Distribution of Cognitive Point Estimates by Predictor Classification | 145 |
| Table 44. Effect Sizes and Null Tests for Cognitive Engagement by Predictor Classification: Type | 147 |
| Table 45. Meta-regression Model for Cognitive Predictor Classification: Type..... | 148 |
| Table 46. Effect Sizes and Null Tests for Cognitive Engagement by Predictor Classification: SDT..... | 149 |
| Table 47. Meta-regression Model for Cognitive Predictor Classification: SDT | 150 |
| Table 48. Distribution of Behavioral Point Estimates by Predictor Classification..... | 151 |
| Table 49. Effect Sizes and Null Tests for Behavioral Engagement by Predictor Classification: Type | 153 |
| Table 50. Effect Sizes and Null Tests for Behavioral Engagement by Predictor Classification: SDT..... | 153 |
| Table 51. Distribution of Point Estimates by Engagement Type and Predictor | 156 |
| Table A1. Coding Scheme for Included Studies..... | 218 |
| Table B1. Statistics and Moderators by Point Estimate..... | 220 |
| Table C1. Overview of Included Studies..... | 227 |
| Table D1. Descriptive Statistics for Included Studies | 230 |
| Table E1. Selection and Use of Predictor and Criterion Variables by Study | 233 |
| Table F1. Studies Included in the Meta-Analysis: References..... | 241 |

List of Appendices

| | |
|--|-----|
| Appendix A. Coding Scheme | 218 |
| Appendix B. Statistics and Moderators by Point Estimate | 220 |
| Appendix C. Overview of Included Studies | 227 |
| Appendix D. Descriptive Statistics for Included Studies..... | 230 |
| Appendix E. Selection and Use of Predictor and Criterion Variables by Study..... | 233 |
| Appendix F. Studies Included in the Meta-Analysis: References | 241 |

Acknowledgements

The phrase “it takes a village” rings true as I think of who I want to acknowledge as I complete this dissertation. I am indebted to the physics department family at SPU, who lead inspiring teacher staff development that turned my attention toward the university and its doctoral programs. Thank you to Stamatis Vokos, Lane Seeley, Rachel Scherr, Amy Robertson, Lezlie DeWater, Abigail Daane, Kara Gray, Katey Houmiel, and Laurie Mendes.

My doctoral committee was second to none. My chair, Dr. Arthur Ellis, gave thoughtful, profound, and critical feedback that pushed me to think more deeply about my topic and its implications for educational practice. Though some might say a dissertation is a laborious process, I truly anticipated and enjoyed Dr. Ellis’ feedback. I will miss the intellectual banter with him as I journey out of the doctoral program. Dr. Stamatis Vokos supported me both intellectually and emotionally through career transitions I experienced during the program. His support was critical to my perseverance and ability to continue the program. I also owe thanks to Dr. Rick Eigenbrood, who stepped onto my committee late in the process to offer valuable feedback concerning statistical methods in my dissertation. Last, I want to thank Dr. Andrew Lumpe for guiding me masterfully through the initial stages of the process, and for providing calm, sure, guidance through the proposal process and subsequent work.

Though the support of peers is emphasized early in the program, I could not have imagined the friendships I would make, nor the importance of the the peer support I would receive by way of both cheerleading and critical review. Thank you to Nalline Baliram, David Hartman, Amy Wright, Kathleen Cifu, and countless others in my cohort

for answering incessant questions, providing moral support, and otherwise ensuring I stayed on track. Thanks to Kimberly Jensen for sharing her expertise in writing a meta-analysis.

Most importantly, I'd like to thank my husband, Ross Aker, for providing endless encouragement and perspective throughout this process. He tolerated me being glued to my computer for a long time—I am looking forward to spending more quality time with him now that this process is winding down. I'd also like to thank my father-in-law, Stephen Aker, for his support he has given in numerous ways from the beginning through the end of my doctoral program.

Lastly, I'd like to thank the researchers I contacted who took the time to respond to my questions about their studies: Dr. Ching-Huei Chen, Meghan Bathgate, Dr. Brian Matthews, Dr. Barry Fraser, Dr. Dana Vedder-Weiss, Dr. Hana Swirski, Dr. Linnenbrink-Garcia, Sarah Blanchard, Dr. Jonathon Osborne, Dr. Maria Adamuti-Trache, and Dr. Andrew Martin. These very busy people contributed knowledge and valuable information that ensured that my dissertation was as rich and complete as it could possibly be.

Abstract

A Meta-Analysis of Middle School Science Engagement

by

Leanna Aker

Seattle Pacific University

Dissertation Chair: Dr. Arthur K. Ellis

Researchers and educational practitioners have long been concerned with declines in science engagement reported by students as they transition into the middle school setting. Though the operationalization of engagement is still nascent, an emerging consensus on a three-faceted model of student engagement has recently emerged in the research literature (Fredricks, Blumenfeld, & Paris, 2004). Thus, a synthesis of existing primary research of early adolescents' science engagement under this emerging conceptualization was warranted. The results of this meta-analysis indicate that instructional methods, class characteristics and competence predictors had the strongest relationship with self-reported science engagement in early adolescence. These predictors also show the strongest relationship with affective and cognitive engagement sub-types. Though affective and cognitive engagement were well represented in primary studies, behavioral engagement was underrepresented in student self-reports.

Keywords: meta-analysis, engagement, behavioral engagement, cognitive engagement, affective engagement, science, middle school, junior high school, early adolescence, self-determination theory, stage-environment fit theory

Chapter 1: Introduction

Background

Engagement is an important area for education research, both as a desired outcome of schooling, and as a predictor of a number of positive educational outcomes. Learning to remain engaged and persist is an important education goal (Finn & Zimmer, 2012). However, engagement is also linked to a number of positive outcomes such as achievement (Bresó, Schaufeli, & Salanova, 2011; Chang, Singh, & Mo, 2007; Finn & Zimmer, 2012; Fredricks, Blumenfeld, & Paris, 2004; Nolen, 2003) and lowered educational risk (Finn & Rock, 1997). Though many educational constructs link to positive educational outcomes, engagement is intuitively understood by practitioners as malleable, and engagement is responsive to school and teacher practices (Finn & Zimmer, 2012; Singh, Granville, & Dika, 2002; Skinner & Pitzer, 2012). Thus, engagement can be positively influenced by school and teacher practices to improve student achievement.

Despite the link between engagement and achievement, the operationalization of the construct is still in its infancy, showing overlap with existing terms, theories and constructs. Engagement research has also evolved over the years from a broad focus on dropout prevention to a finer focus on task and personal-level variables (Finn & Zimmer, 2012; Newmann, 1981; Reschly & Christenson, 2012; Sinatra, Heddy, & Lombardi, 2015). Despite these overlaps and changes, an emerging consensus on a three-faceted engagement model including behavioral, affective, and cognitive components is apparent in the research literature (Fredricks et al., 2004). Additionally, psychometric engagement instruments appropriate for use in K-12 settings have been developed which verify this three-faceted model (Fredricks et al., 2011; Veiga, Reeve, Wentzel, & Robu, 2014). This

emerging operationalization will provide a framework with which assimilate existing engagement research.

Lack of engagement with school science has been a concern of science researchers and practitioners for many decades as student interest in, and attitudes toward, school science have waned (Jenkins & Pell, 2006; Lee & Anderson, 1993; Osborne, Simon, & Collins, 2003). These declines often coincide with the transition into middle school (Braund & Driver, 2005; Eccles et al., 1993; Eccles & Roeser, 2010; Mahatmya, Lohman, Matjasko, & Farb, 2012). However, researchers demonstrated that declining engagement is not an inevitable outcome of the transition to middle school (Anderman & Maehr, 1994; Eccles et al., 1993; Vedder-Weiss & Fortus, 2011). Early adolescent students possess rich developmental potential to cognitively engage by reasoning abstractly, considering multiple perspectives, and weighing several strategies simultaneously (Mahatmya et al., 2012; Piaget, 1972).

Self-determination theory (SDT) and stage-environment fit (SEF) theory can guide an evaluation of research about early adolescents' engagement with middle school science. SDT posits that students are most likely to be motivated when they feel a sense of competence, autonomy, and relatedness (Roeser & Eccles, 1998; Ryan & Deci, 2000). SEF theory suggests that a good fit between the educational environment and students' developmental needs will lead to increased engagement (Eccles & Midgley, 1989, Eccles et al., 1993). As early adolescents are unique in their increasing developmental need for autonomy and relatedness, these two theories provide a lens with which to evaluate engagement research at this age level.

Purpose of the Study

The purpose of this study is to conduct a quantitative synthesis of existing engagement studies using the three-faceted theoretical framework. Predictors of middle school science engagement will be identified and ranked in terms of their practical effect. Possible commonalities about effective predictors of engagement will be identified and analyzed using stage-environment fit (SEF) theory and self-determination theory (SDT). Additionally, analyses will be conducted with the purpose of identifying predictors specific to each sub-type of engagement (behavioral, affective, cognitive).

This meta-analysis of middle school science engagement will focus upon classroom and task level engagement, rather than a broad focus on student engagement with science as a discipline. The reason for this is two-fold. There is a great deal of existing literature about students' engagement with science as a discipline, vis-à-vis attitudes, self-efficacy, motivation, and sense of science task value (Britner & Pajares, 2006; DeBacker & Nelson, 2000; Osborne et al., 2003). Another reason is that classroom and task level engagement is something over which educational practitioners have influence and perceive that they have influence. Thus, a synthesis of classroom and task level engagement variables will yield new and useful information for educational practitioners.

Significance of the Study

As conceptual and operational clarity is beginning to emerge about engagement, a meta-analysis of existing engagement studies is a crucial next step for bringing coherence to engagement research. Studies exist in the research literature that purport to measure engagement but use operationalizations that are incongruent with the emerging consensus

about the construct. In 1991, a meta-analysis of engagement was conducted that focused almost exclusively on behavioral indicators of engagement, a limited subset of what is now considered engagement (Kumar, 1991). There are studies that are not identified as engagement-related, yet assess indicators of behavioral, affective, or cognitive engagement. A purposeful, updated synthesis of engagement and engagement-related research will help to solidify an operational definition for the construct.

The identification of practically significant predictors of engagement will also benefit educational practitioners. In the current era, when accountability for student achievement is embedded in many teacher and school evaluation models, teachers prioritize student achievement. Though engagement is linked to achievement, many factors can contribute to achievement. Engagement is intuitively understood by educators and viewed as malleable and responsive to teacher practices (Finn & Zimmer, 2012; Singh et al., 2002; Skinner & Pitzer, 2012). Synthesis of existing research can inform possible interventions to positively impact student engagement with specific science tasks. Identifying effective predictors of each type of engagement can inform targeted interventions to address more specific engagement issues. For example, the results from a meta-analysis could provide direction for a teacher who wishes to positively influence a student's negative affective engagement.

As conceptual clarity about the three facets of engagement emerges, it is expected that some sub-types of engagement will be more heavily represented in the research literature with some predictors of engagement having stronger representation. The results of this study may identify gaps in the literature and identify areas with inconsistent results that suggest further research. Without a comprehensive meta-analysis, such judgments

about omissions and inconsistencies in engagement research are determined based on a limited focus on statistical significance in published studies. As statistical significance is influenced by sample size, and as there is a bias to publish statistically significant results, a meta-analysis is an effective method to inform further research on this construct.

The documentation of declines in student engagement and achievement during or after the transition to middle school supports the focus of this meta-analysis on the early adolescent age group. Researchers have indicated that declining engagement is not an inevitable outcome of the middle school transition (Anderman & Maehr, 1994; Eccles et al., 1993). Early adolescent students possess rich developmental potential to cognitively engage by reasoning abstractly, considering multiple perspectives, and weighing several strategies simultaneously (Mahatmya et al., 2012; Piaget, 1972). Thus, it is imperative to determine which aspects of science class and science tasks more effectively engage middle school students.

Definitions

Engagement. Though definitions for engagement abound, they all coalesce around the notion that engagement involves multiple aspects such as participation, investment, and effort. Newmann (1992) wrote that engagement is “the student’s psychological investment in and effort directed toward learning, understanding, or mastering the knowledge, skills, or crafts that academic work is intended to promote” (p. 12). Marks (2000) referred to engagement as “the attention...investment, and effort students expend in the work of school” (p. 155). In 2004, Fredricks et al. published a seminal review of engagement literature, and proposed that engagement consists of behavioral, affective, and cognitive components.

Behavioral engagement. Students who are behaviorally engaged show on-task actions such as attention, participation and school attendance (Caraway & Tucker, 2003; Fredricks et al., 2004). When engagement first began to appear in the research literature in the late 1970’s in relation to school dropout studies, engagement was conceptualized as solely behavioral in nature, with empirical evidence such as “time on task” and “engaged time” (Anderson, 1975; Stallings, 1980). Behavioral engagement is typically discernible by an external observer. For example, an observer can quantify “time on task” by watching students’ actions.

Affective engagement. Affectively engaged students are interested, see value in the tasks they are given, and have positive emotions about what they are experiencing (Fredricks et al., 2004). Researchers recently suggested that affective engagement in science could be differentiated into feelings about science, feelings in science class, and feelings within science (Jaber, 2014; Jaber & Hammer, 2016). Feelings about science can

be seen to have overlap with well-developed bodies of knowledge about attitudes and interests toward science as a discipline (Osborne et al., 2003). Feelings within science include aspects such as being driven to resolve inconsistencies and delighting in the discovery of new information. As this meta-analysis is focused upon classroom and task-level variables, feelings in science class and feelings within science will be of greater focus than feelings about science as a discipline. Affective engagement is difficult for an external observer to assess and analyze. A student who is frowning could be enjoying a particular task; a student who is frustrated could be deeply cognitively engaged in resolving inconsistencies in thought. Student self report is an important aspect of assessing affective engagement.

Cognitive engagement. Cognitively engaged students think critically and creatively, reflect on their learning, and use multiple strategies for learning. Typical indicators of cognitive engagement include effort, goal orientation, and help-seeking behavior. Two alternative models in the research literature identify an additional, fourth facet of engagement indicating the importance of help-seeking behavior to cognitive engagement. Some refer to this fourth facet as self-initiated cognitive engagement (Lee & Anderson, 1993) while others refer to it as agentic engagement (Sinatra et al., 2015; Veiga et al., 2014). For this meta-analysis, help-seeking behavior will be considered a component of cognitive engagement. As with affective engagement, cognitive engagement is difficult for an external observer to assess and analyze. For example, a student may show no help-seeking behaviors, but be profoundly engaged in mental processing. Student self-report is important for assessing cognitive engagement.

Research Questions and Null Hypotheses

Research questions. Classroom and task-level predictors of engagement with middle school science as assessed by student self-report will be examined using meta-analysis. The research questions examined in this study include the following:

1. What moderators have statistically significant practical effects on early adolescents' science engagement as assessed by student self-report?
2. What predictors have the largest practical effect on early adolescents' science engagement as assessed by student self-report?
3. What commonalities exist among predictors that have the largest practical effect on early adolescents' science engagement as assessed by student self-report?
4. What predictors have the largest practical effect on early adolescents' behavioral, affective, and cognitive engagement in science as assessed by student self-report?
5. What commonalities exist among predictors that have the largest practical effect on early adolescents' behavioral, affective, and cognitive engagement in science as assessed by student self-report?
6. What predictors are underrepresented in the research literature on middle school science engagement as measured by student self-report?
7. What types of engagement are underrepresented in the research literature on middle school science engagement as measured by student self-report?

Null Hypotheses. Two null hypotheses derive from the research questions above:

H₀: There are no statistically significant differences in the practical effects of predictors on engagement with middle school science as measured by student self-report.

H₀: There are no statistically significant differences in the practical effects of different predictors on behavior, affective, or cognitive engagement with middle school science as measured by student self-report.

For each null hypothesis, the independent variables are the predictors of engagement, and the dependent variable is engagement, as measured by student self-report.

Content of the Following Chapters

The remainder of this dissertation is divided into four chapters titled Literature Review, Research Methods, Results, and Summary of Findings. The Literature Review includes the historical evolution of the construct, alternative models, overlaps, and empirical evidence for engagement. The Research Methods chapter describes the research design, criteria for inclusion/exclusion of studies, methodology, and data analysis. The Results chapter presents descriptive statistics and meta-analytic results related to the research questions and hypotheses. In the Summary of Findings chapter, the results are discussed in relation to the research questions and hypotheses. Possible patterns in highly effective predictors will be identified, as well as gaps and inconsistencies in middle school science engagement research.

Chapter 2: Literature Review

Engagement

The term “engagement” is ubiquitous in the educational field, appearing in teacher evaluation criteria, educator vernacular, and educational research. Part of the reason that the term is so pervasive is that it has such an intuitive meaning in education. This intuitive meaning is reflected in different definitions of engagement found in the research literature: “the student’s psychological investment in and effort directed toward learning, understanding, or mastering the knowledge, skills, or crafts that academic work is intended to promote” (Newmann, 1992, p. 12), “the attention...investment, and effort students expend in the work of school” (Marks, 2000, p. 155), and “constructive, enthusiastic, willing, emotionally positive, and cognitively focused participation with learning activities in school” (Skinner & Pitzer, 2012, p. 22). Thus, engagement refers to a student’s quality of participation in school and its academic tasks.

Despite this intuitive meaning, or perhaps because of it, engagement has only recently begun to become operationalized as a construct. While differing engagement models can be found in the research literature, they each fundamentally attempt to describe and differentiate between high and low quality engagement. Some researchers criticize engagement as subsuming, duplicating, or overlapping existing educational constructs, such as motivation or attitudes toward a discipline (Azevedo, 2015; Fredricks et al., 2004). Lastly, due to historical changes in both the construct itself, and its grain size of interest, differentiating between facilitators, indicators, and outcomes of engagement has been challenging. Nevertheless, the seminal literature review by

Fredricks et al. (2004) created a synthesis of the engagement construct that has guided engagement research since.

Origins and historical evolution of the construct. Engagement began to appear in research literature in the late 1970's in relation to school dropout studies (Finn, 1989; Finn & Zimmer, 2012; Reschly & Christenson, 2012). Students who dropped out were believed to be disengaged with school. Early empirical evidence for the construct included "time on task," "engaged time," and school attendance (Anderson, 1975; Stallings, 1980). Engagement was operationalized as observable student behaviors indicating students were participating in school and academic tasks. The construct quickly broadened in scope as behavioral indicators such as time on task failed to completely describe or explain engagement. Even the definition given in the earliest literature review of engagement, "the attitude leading to, and the behavior of, participation in the school's programs," (Mosher & McGowan, 1985, p. 14) suggested engagement included behavioral and affective aspects.

Three early models reflected this broader conceptualization of engagement, and suggested predictors or mediators of school engagement. The school reform model emphasized the role of context in student engagement, and originated from a literature review about student alienation in schools (Newmann, 1981). This model proposed that features of the school environment and culture determined student engagement; one could foster student engagement by fixing the school (Newmann, 1981; Wehlage, Rutter, Smith, Lesko, & Fernandez, 1989). Proponents of the school reform model advocated smaller class sizes, better relationships between teachers and students, and increased involvement of students in school policy decisions (Newman, 1981). With these changes,

students would have more opportunities to participate and would have better affective perceptions of their school experience.

Alternatively, the self-system process model emphasized the role of personal needs in student engagement (Connell & Wellborn, 1991; Skinner & Belmont, 1993). This model proposed that students engaged or disengaged from school and academic work based on their perceptions of having their needs met. Informed by self-determination theory, the self-system process model identified those needs as competence, autonomy, and relatedness (Deci & Ryan, 1985). To the extent that a student feels able to complete school tasks, to have control over his or her experiences, and to be socially connected to others, he or she will be engaged. Thus, the self-systems process model suggests that students' affective perceptions are an important predictor of school engagement.

Lastly, the participant-identification model emphasized the interaction of contextual and intrapersonal features. In his seminal work on school dropout prevention, Finn (1989) suggested that engagement was determined by how behavior (participation) and affect (identification with school) interact to impact the likelihood of school success. Identification consisted of not only a sense of belonging, but also of valuing one's school experience. Participation was differentiated into four qualitative levels: appropriate conduct, student initiation of questions, extracurricular opportunities, and opportunities for student governance. The relationship between participation and identification is iterative in Finn's model; as students participate and experience success in school, they can identify with the school, which further impacts engagement. What distinguishes Finn's participant-identification model from both the school reform and self-system

process models is this focus on the interaction between behavioral and affective aspects of engagement.

Though behavioral and affective engagement components appear consistently in these early and many subsequent models of engagement, the notion of engagement quality or degree appears in varied ways in a number of engagement models as well. Finn's four levels of behavioral engagement, vis-à-vis participation, were one of the first attempts to suggest that engagement had different qualitative levels (Finn, 1989). Soon thereafter, Nystrand and Gamoran (1991) distinguished between types of engagement in terms of commitment and purpose—substantive engagement is a sustained commitment to the content of schooling, and procedural engagement is a commitment to completing task requirements, which lasts only as long as the task itself. Ainley's (2012) engagement model is similar, including “high gear” and “low gear” categories; engagement occurs when students have connected with the content of a task “rather than simply performing the activity mechanically or pretending to perform the activity” (p. 286). Greene and Miller (1996) differentiated between shallow and meaningful cognitive engagement, while Meece, Blumenfeld, and Hoyle (1988) differentiated between superficial and active engagement. Productive disciplinary engagement is another model that implies a level of engagement (Engle & Conant, 2002). Productive disciplinary engagement distinguishes between low-level engagement such as time on task, and engagement that results in student progress in understanding the discipline of study. Another model divides behavioral engagement into academic (time on task) and behavioral (participation) components (Appleton, Christenson, Kim, & Reschly, 2006).

In addition to engagement quality, one feature that appeared in several engagement models was the distinction between teacher-initiated and student-initiated engagement. Two of Finn's four categories of behavioral engagement reflect this distinction (1989). Level one participation reflects students merely attending to teachers' requests, while level two participation involves students proactively initiating the process of asking questions. Lee and Anderson (1993) proposed the idea of self-initiated cognitive engagement, which reflected students initiating learning activities and going beyond the requirements of a particular task. Agentic engagement is a similar idea, reflecting students exerting their agency in the learning process by personalizing, modifying or seeking instruction (Reeve & Tseng, 2011; Sinatra et al., 2015).

Engagement has also been conceptualized in terms motives or purposes. Nystrand and Gamoran's (1991) substantive and procedural engagement reflect this distinction. Students who are substantively engaged have a sustained commitment to school and academic tasks, while students who are procedurally engaged are interested in simply completing the tasks in front of them. Schlechthy's engagement model (2011) reflects five qualitative levels: engagement, strategic compliance, ritual compliance, retreatism, and rebellion. Engaged students are authentically interested in the task at hand. Strategically compliant students do what is asked because of an ulterior motive (e.g., to obtain a good grade). Ritually compliant students do what is asked to avoid getting into trouble. Retreatist students do not participate in the activity at hand, and rebellious students actively do something other than what was asked.

One model of engagement differs substantially from those mentioned previously. Bresó et al. (2011) proposed a three-faceted model consisting of vigor, dedication, and

absorption. Vigor is characterized by effort and resilience; dedication by enthusiasm and inspiration; and absorption by full concentration on a task. Though this model is qualitatively quite different than those considering behavioral and affective components of engagement, one can identify aspects of those components. For example, dedication can be seen to have overlap with affective engagement. Nevertheless, Bresó et al.'s model provides a different perspective on engagement that may yield useful information to education practitioners and researchers.

Still other models give special consideration to disengagement. Schlechty's model, mentioned previously, has five levels, one of which—rebellion—differs in that it reflects a student actively doing something other than the task in front of him or her (Schlechty, 2011). The level above this in Schlechty's model—retreatism—reflects the simple absence of engagement. Some models propose that engagement and disengagement are fundamentally distinct, rather than opposite ends on a continuum. The rationale for this is elucidated in a simple analogy from medicine: disease is not simply the absence of health (Reschly & Christenson, 2012). Similarly, disengagement is not simply the absence of engagement; anxiety is fundamentally distinct from the absence of emotion. Two models reflect this distinction between engagement and disengagement. Skinner, Kindermann, & Furrer (2008) distinguished between four categories of engagement: behavioral engagement, behavioral disaffection, emotional engagement, and emotional disaffection. Martin (2007) described four higher order factors of engagement: adaptive cognition, adaptive behavior, maladaptive behavior, and maladaptive/impeding cognition.

Accepted Model of Engagement. Despite these varied models, a seminal synthesis of engagement research suggested a model of engagement that has been generally adopted by educational researchers since (Fredricks et al., 2004). This review proposed that engagement is a meta-construct with three facets—behavioral, cognitive, and affective (Fredricks et al., 2004; Linnenbrink & Pintrich, 2003). Behaviorally engaged students show on-task actions such as attention, participation, and school attendance (Caraway & Tucker, 2003). Affectively engaged students are interested, see value in the tasks they are given, and have positive emotions about what they are experiencing (Fredricks et al., 2004). Cognitively engaged students are self-regulated learners, use multiple strategies for learning, and show effort above and beyond what is required (Azevedo, 2015; Fredricks et al., 2004; Pintrich & DeGroot, 1990; Wang, Willet, & Eccles 2011).

In this three-faceted model, one can assimilate prior models and identify the foundation for subsequent models. For example, the attempt to distinguish between levels or degrees of engagement (Ainley, 2012; Appleton et al., 2006; Engle & Conant, 2002; Finn, 1989; Greene & Miller, 1996; Meece et al., 1988; Nystrand & Gamoran, 1991; Schlechty, 2011) can be reflected in the addition of cognitive engagement as an aspect distinct from behavioral engagement. Cognitive engagement reflects a deeper, more authentic engagement with the content of education, while behavioral engagement reflects a more superficial participation. The notion of students' pro-active role in engagement can be seen as a subcategory of cognitive engagement (Finn, 1989; Lee & Anderson, 1993; Reeve & Tseng, 2011; Sinatra et al., 2015). For example, a student can use a variety of learning strategies (cognitive engagement) in response to a teacher

request, or because he or she decided to do so. Even Bresó et al.'s (2011) qualitatively distinct model can be seen to reflect a deep level of cognitive engagement.

While some models informed the seminal literature review by Fredricks et al. (2004), other models were developed subsequent to the review (Appleton et al., 2006; Bresó et al., 2011; Martin, 2007; Reeve & Tseng, 2011; Schlechty, 2011; Sinatra et al., 2015; Skinner et al., 2008). However, with the exception of agentic engagement, these newer models have not been validated psychometrically and have not generally gained acceptance in the research community (Reeve & Tseng, 2011; Sinatra et al., 2015; Veiga & Robu, 2014). Furthermore, many subsequent models can be seen to add to, rather than fundamentally alter, the three-faceted model. For example, differentiating behavioral engagement into participatory and academic components does not fundamentally differ from Fredrick et al.'s (2004) model, but rather suggests an addition or alteration (Appleton et al., 2006). Additionally, Jaber and Hammer's research (2016) can be interpreted to expand affective engagement to include engagement with the attitudes and interests necessary to participate in a discipline, such as an interest in rectifying conflicting results as a measure of affective science engagement.

Another consideration about the three-faceted model of engagement concerns a potential sequence of the facets relative to each other. For example, does one type of engagement precede, mediate, or predict the other? Educational practitioners might intuitively suppose if they can obtain student participation (behavioral engagement), affective and cognitive engagement will follow. Some hypothesize that a student's affect is either a precursor to or a consequence of engagement (Eccles & Wang, 2012; Pekrun & Linnenbrink-Garcia, 2013). A model in the research literature suggests a different

sequence for the three facets; Reschly & Christenson (2006) suggested that cognitive and affective engagement predict changes in a student's behavior. Regardless, many researchers agree that engagement effects are iterative (Reschly & Christenson, 2012). For example, cognitive engagement in a task could predict or mediate future affective engagement with similar tasks.

The three-faceted model of engagement is dominant in the research literature—it has been validated psychometrically, used to examine and categorize psychometric instruments, taken up and cited by researchers in subsequent studies, and used to interpret existing research about engagement (Doğan, 2014; Fredricks et al., 2004; Fredricks et al., 2011; Sinatra et al., 2015; Veiga et al., 2014; Wang & Holcombe, 2010; Wang et al., 2011). Furthermore, behavioral, affective, and cognitive engagement can be intuitively understood as distinct. One can imagine a situation in which a student is behaviorally but not cognitively engaged, or affectively but not cognitively engaged. The three-faceted model will be used to guide this meta-analysis of middle school students' engagement in science.

Related constructs. One criticism of the engagement construct is that there is a great deal of overlap between it and other theoretical constructs. For example, engagement research overlaps with research on student attitudes, motivation, and self-regulated learning (Ford, 1992; Osborne et al., 2003; Zimmerman, 1990). Fredricks et al. (2004) acknowledged this problem:

Because there has been considerable research on how students behave, feel, and think, the attempt to conceptualize and examine portions of the literature under the label “engagement” is potentially problematic; it can

result in a proliferation of constructions, definitions, and measures of concepts that differ slightly, thereby doing little to improve conceptual clarity. (p. 60)

However, Fredricks et al. (2004) suggested that combining behavior, emotion, and cognition under the label “engagement” is valuable because it may provide a “richer characterization of children than is possible in research on single components” (p. 61). Nevertheless, it is necessary to differentiate engagement from related constructs for psychometric and theoretical reasons.

Flow. The relationship between flow theory and engagement is strong and deserves further elucidation. Csikszentmihalyi (1990) defined flow “a state of deep absorption in an activity that is intrinsically enjoyable, as when artists or athletes are focused on their play or performance.” Flow is an amalgamation of concentration, interest, and enjoyment; all three aspects must be present for the flow experience to occur (Shernoff, Csikszentmihalyi, Schneider, & Shernoff, 2003). The experience of flow occurs when a task is uniquely matched to a person’s skillset, with those skills neither being inadequate nor underutilized for the task. This suggests that flow experiences can be created by considering the zone of proximal development (Vygotsky, 1978). From the perspective of flow theory, the most ideal way to engage students is to provide appropriate challenges and scaffolded opportunities to enhance skills (Shernoff et al., 2003). Because the flow experience is itself intrinsically rewarding, individuals who experience this phenomenon seek to have more flow experiences, and thus become increasingly intrinsically motivated.

Intuitively, flow describes a deep level of cognitive engagement. However, flow overlaps with both cognitive and affective aspects of engagement. Concentration relates to cognitive engagement, while interest and enjoyment overlap with affective engagement. While behavioral engagement is not explicitly present in flow theory, one can infer that it is taken for granted; behavioral engagement should be present when someone is deeply absorbed in an activity. Flow theory thus seems to support the idea of high and low levels of engagement reflected in some alternative models (Ainley, 2012; Engle & Conant, 2002; Finn, 1989; Greene & Miller, 1996; Meece et al., 1998; Nystrand & Gamoran, 1991; Schlechty, 2011), and it suggests possible predictors of engagement, such as the use of high-interest tasks uniquely suited to the ability level of students.

Motivation. The relationship between flow theory and intrinsic motivation suggests a second construct with which engagement has a great deal of overlap—motivational theory (Ford, 1992). Fredericks et al. (2004) suggested that engagement, as a meta-construct, subsumes motivation; and in fact, some researchers have used the terms engagement and motivation interchangeably (Martin, 2007; Reschly & Christenson, 2012). One current researcher conducted studies suggesting that engagement fully mediates the relationship between motivation and achievement (Reeve, 2012; Reeve & Tseng, 2011). Still others have suggested that motivation is the theoretical framework that undergirds engagement (Connell & Wellborn, 1991; Yazzie-Mintz & McCormick, 2012). Briefly, Ford's motivational systems theory posits that achievement is the result of motivation, skill, and a responsive environment (Ford, 1992). Further, motivation is a combination of goals, emotions, and personal agency beliefs. Both historical and current engagement models reflect aspects of motivational theory. For example, the self-systems

process model suggests that students' personal agency beliefs (self-efficacy and context beliefs) relate to students' goals (Bandura, 1977; Connell & Wellborn, 1991; Skinner & Belmont, 1993). Affective engagement parallels the idea that students' emotions impact their motivation. Cognitive engagement can be influenced both by skill and personal agency beliefs. Engagement is clearly grounded in, if not heavily overlapping with, Ford's motivational theory.

Some researchers have differentiated motivation and engagement by conceiving of motivation as intent, and engagement as the action that results from that intent (Connell & Wellborn, 1991; Reschly & Christenson, 2012; Skinner & Pitzer, 2012). In this view, motivation is internal and precedes engagement. However, this view of motivation as internal prerequisite, and engagement as external, manifested action, conflicts with existing three-faceted models of engagement. While behavioral engagement is clearly external, manifested action, affective and cognitive engagement are internal and do not necessarily present as observable "action," except by inference. Nevertheless, affective and cognitive engagement can be seen as outcomes of motivation, even if those outcomes are not necessarily visible action. (Finn & Zimmer, 2012; Skinner & Pitzer, 2012). For example, indicators of a student's affective engagement include enthusiasm, enjoyment, and satisfaction (Skinner & Pitzer, 2012). Those indicators can intuitively be understood as possible outcomes of a student's motivation. Motivation is necessary, but not sufficient, for engagement (Appleton et al., 2006). Ford's motivational theory suggests that motivation interacts with contextual variables to determine whether internal inclinations develop into those actions or outcomes that define engagement and lead to achievement.

Behavioral Intent. Another construct that reflects the idea of attitudes and beliefs as potential precursors to observable action is behavioral intent (Ajzen, 1991; Ajzen & Fishbein, 1977). Ajzen and Fishbein's work began with an attempt to identify situations in which a person's attitude toward something predicted their manifested behavior. Ajzen later expanded on this work and developed the theory of planned behavior, which proposes that attitudes, subjective norms, and perceived behavioral control determined behavioral intent (Ajzen, 1991). Attitudes have several components, including target, action, environment, and time. For example, a student can have attitudes about hands-on activities (action), in science (target), in a specific teacher's classroom (environment) in eighth grade (time). Subjective norms refer to the likelihood that people would agree or disagree with the behavior. Perceived behavioral control (PBC) reflects a perception about how easy or difficult it will be to perform a particular behavior; Ajzen likens PBC to self-efficacy (Bandura, 1977). Ajzen further suggested that the relative weights of each component would differ in different contexts. To extrapolate this to a specific educational example, in a classroom with a very strict teacher, the role of perceived behavioral control might carry more weight than either subjective norms or attitudes toward the task at hand.

The work of Ajzen and Fishbein (1977) can be interpreted as an attempt to determine predictors of behavioral engagement. Their work suggested that engagement researchers look toward context-specific attitudes, peer perceptions, and student perceptions of ability, as possible predictors of behavioral engagement. Affective engagement is also reflected within the construct of behavioral intent and the theory of planned behavior, though interestingly, affective engagement can be interpreted as a

precursor to behavioral engagement in this model. This notion of affective engagement preceding or predicting behavioral engagement is reflected in some conceptualizations of engagement (Reschly & Christenson, 2006). Regardless, Ajzen's research has suggested an examination of the link between behavioral and affective engagement. Work on behavioral intent and the theory of planned behavior suggests that a focus on a smaller grain size is warranted in engagement research.

Situational interest. Like behavioral intent, situational interest suggests a focus on a smaller grain size, such as with specific tasks or situations. Though situational interest is a construct that was developed to explain differential student engagement with reading, its categories and assumptions can be applied more generally to educational tasks. Situational interest refers to a temporary and context-dependent desire to engage with a task (Schraw & Lehman, 2001). The construct is thus differentiated from personal interest, which is enduring and irrespective of context. Situational interest is speculated to influence a variety of outcomes which relate to engagement: use of specific learning strategies and the extent to which one engages in deeper processing (cognitive engagement), feelings toward a task (affective engagement), and how one allocates attention (behavioral engagement) (Hidi, 1990; Schiefele, 1999; Schraw, 1998).

There are three aspects of situational interest: text, task, and knowledge. Text-based situational interest refers to aspects of a text (or any content, by extension) that affect interest. Such text-based aspects can include coherence, vividness, or ease of comprehension. Task-based situational interest refers to features of the task itself, such as guiding a student's goals or altering a task to make it more approachable. Knowledge-based situational interest refers to the relationship prior knowledge has on interest. The

effects of prior knowledge on situational interest are not linear; extremely low or high prior knowledge would seem to predict low interest (Kintsch, 1980).

The intersection of situational interest with engagement research is complex. On the one hand, situational interest seems to relate closely to affective engagement, as both constructs reflect emotions, attitudes and values about a particular task or topic.

Situational interest has also been speculated to be a precursor to affective engagement (Schiefele, 1999; Schraw, 1998). Situational interest could also be considered a mediator of cognitive and/or behavioral engagement. Schank (1979) coined the phrase interest-based parsing to describe one's allocation of his or her cognitive resources based on interest. Thus, Schank's work supports the notion that one type of engagement may precede the other; affect may predict cognitive or behavioral engagement (Reschly & Christenson, 2006). Situational interest can contribute to engagement research by not only suggesting characteristics of individual tasks that might predict engagement, but also by highlighting a role for affective engagement as the gatekeeper of cognitive resources, and thus, cognitive engagement.

Theoretical Frameworks

While the focus of the previous section was to link as well as differentiate engagement from related constructs in the research literature, those related constructs could also be seen as frameworks with which to synthesize engagement research. Nevertheless, this meta-analysis will utilize two other theoretical frameworks, self-determination theory (SDT) and stage-environment fit (SEF) theory. These frameworks allow one to more explicitly consider engagement in relationship to the unique developmental needs of early adolescent students in a middle school setting.

Self-Determination Theory. Self-determination theory (SDT) explains conditions that sustain and encourage motivation. While the theory focuses on the idea that motivation arises from the fulfillment of intrapersonal needs, it stresses the role of social contexts in promoting or hindering motivation. SDT posits that people are most motivated to learn when they feel a sense of competence, autonomy, and relatedness (Roeser & Eccles, 1998; Ryan & Deci, 2000). Competence refers to a sense that one can accomplish a task, autonomy refers to the sense that one has control over those tasks, and relatedness refers to the need to connect with others. Social contexts that promote these three needs serve to foster intrinsic motivation, while social contexts that do not promote these needs, or promote one at the expense of the other, serve to diminish intrinsic motivation. For example, the use of rewards may encourage competence, but will likely decrease one's sense of autonomy.

In addition to describing the characteristics of social contexts that promote intrinsic motivation, SDT establishes a typology of motivation, which is organized by the degree to which values and behaviors are internalized and integrated by a person. This typology distinguishes amotivation, extrinsic motivation, and intrinsic motivation; but more interestingly, subdivides extrinsic motivation into four types—external regulation, introjected regulation, identified regulation, and integrated regulation. Externally regulated extrinsic motivation is characterized by compliance and external rewards and punishments. Introjected regulation also involves rewards and punishments, but they are more internally regulated, such as through ego involvement. Identified regulation refers to a sense of conscious valuing of a behavior, while integrated regulation goes further in that this conscious valuing is integrated with a sense of self and one's own goals. Deci

and Ryan (1985) referred to external regulation and introjected regulation as controlled; while identified regulation, integrated regulation, and intrinsic motivation are considered autonomous. Intrinsic motivation differs from extrinsic motivation in that intrinsic motivation is characterized by interest, enjoyment, and inherent satisfaction.

Self-determination theory can inform engagement research in a number of theoretical and practical ways. SDT provides a framework with which to view engagement interventions in social context; fostering autonomy, competence, and relatedness can enhance a student's level of motivation, and thus by extension, engagement. Furthermore, these characteristics can be promoted in several different realms of a student's experiences—within student-teacher relationships, peer relationships, and school environments. The motivation typologies in SDT provide a way to predict and explain different levels of engagement. For example, the use of rewards and punishments is likely to lead to low levels of behavioral engagement, as such practices diminish autonomy, and lead to more externally regulated extrinsic motivation. Conversely, helping a student to see the relevance of a particular task or topic is likely to help that student internalize the value of the task, leading to identified regulation. The typologies of extrinsic motivation in SDT further suggest that deep levels of engagement are possible, even with material that does not hold inherent interest or enjoyment for a student. Both identified regulation and integrated regulation both are characterized by self-regulation, often cited as an indicator of cognitive engagement (Appleton et al., 2006; Greene & Miller, 1996; Linnenbrink & Pintrich, 2003).

Stage-Environment Fit Theory. Stage environment fit (SEF) theory posits that decreases in motivation and affective engagement are caused by of a mismatch between developmental student needs and existing school environments (Eccles & Midgley, 1989; Eccles et al., 1993). In other words, to the extent that a school or classroom uses developmentally appropriate practice, the better the “fit” between the students and the environment, and the more students will engage and achieve. As perceptions of school decline for many students during or after the transition to middle school, a developmental perspective on engagement, such as that afforded by SEF theory, is warranted (Braund & Driver, 2005; Eccles et al., 1993; Eccles & Roeser, 2010; Mahatmya et al., 2012).

There is extensive evidence, both from external observation and student self-report, that the developmental match between early adolescents and their middle school classroom environments is poor (Anderman & Maehr, 1994; Anderman & Mueller, 2010; Eccles et al., 1993 Eccles & Roeser, 2010; Midgley, Feldlaufer, & Eccles, 1989). Early adolescents have increasing needs for autonomy, yet experience less control (Anderman & Mueller, 2010). Research has suggested that middle school teachers focus more heavily on behavior management and control than their elementary school teacher colleagues (Hoy, 2001; Midgley et al., 1989; Roeser & Eccles, 1998, Ryan & Patrick, 2001; Wentzel, 2010). Other research documents that students are given less opportunities for choice, self-management, and decision-making at both the school and classroom level (Feldlaufer, Midgley, & Eccles, 1988). Additionally, the transition to middle school itself disrupts peer relationships at a time when students have a growing peer orientation. Because of this increasing value placed on peer interaction, early adolescents also experience increased self-consciousness. Here again, there is a mismatch

between needs and environments: students report their middle school classrooms are characterized by competition, performance goal orientations, public evaluation of work; and a decreased level of nurturing in the teacher-student relationship (Eccles & Midgley, 1989; Eccles et al., 1993; Lepper, Corpus, & Iyengar, 2005; Midgley et al., 1989; Roeser & Eccles, 1998). Furthermore, research has suggested that middle school students receive lower grades than at any other time, which has been linked to teachers having higher academic standards. These characteristics of middle school classrooms are likely to increase social comparison and decrease self-efficacy at a time when students are increasingly self-conscious.

From a strengths-based perspective, early adolescent students have rich developmental potential to cognitively engage by reasoning abstractly, considering multiple perspectives, and weighing several strategies at the same time (Anderman & Mueller, 2010; Mahatmya et al., 2012; Piaget, 1972; Ryan & Patrick, 2001). Students entering middle school have a developmental need for more abstract, cognitively demanding academic tasks (Anderman & Mueller, 2010; Piaget, 1972). However, student self-reports indicate that the cognitive demand of tasks decreases after the middle school transition (Uekawa, Borman, & Lee, 2007; Walberg, House, & Steele, 1973). One study documented that in 11 seventh grade science classrooms, the most frequent activities were copying information from the board and filling in worksheets (Mergendoller, Marchman, Mitman, & Packer, 1988). The content covered in middle school science classrooms is more academic as well, which may cause students to question the relevance of what they are learning. Thus, from the perspective of SEF theory, not only do the

affective qualities of a classroom conflict with early adolescents' developmental needs, but the cognitive characteristics of academic tasks conflict as well.

There is clearly a link between SEF theory and SDT. While personal needs for competence, autonomy and relatedness exist throughout life, the early adolescent years demand special attention to these needs. Early adolescents not only need competency, autonomy, and relatedness, they are learning *how* to be competent, autonomous, and related effectively to others. Deci and Ryan (2002) suggested that while all three needs are relevant, developmental characteristics can change the importance of one need relative to the other. For example, many of the developmental changes and needs of early adolescents are social in nature. For this reason, perhaps relatedness is more concern than competence in middle school. Regardless, the combination of SDT and SEF theory will provide a rich and flexible developmental perspective on the match between science classrooms and early adolescent students that may explain declining engagement and predict interventions that are likely to have a large practical effect in the classroom.

Measurement of Engagement

Psychometric instruments. Typical measures of engagement include self-report questionnaires, classroom observations, interviews/focus groups, teacher reports, discourse analysis, and physiological measures such as eye movements (Azevedo, 2015; Fredricks et al., 2011; Fredricks & McColskey, 2012; Greene, 2015). Fredricks et al. (2011) identified 21 engagement measures suitable for K-12 use in the research literature between 1979 and 2009. Of those, 14 were student report instruments, three were teacher reports, and four were classroom observation measures. Veiga et al. (2014) conducted a

similar, but more limited review, which focused on multidimensional student self-report measures.

Classroom observation protocols and teacher reports are effective methods for measuring behavioral engagement, which can be easily operationalized as observable actions. Indicators of behavioral engagement include time on task, eye contact, and participation. However, classroom observation protocols and teacher reports also have the potential to measure cognitive engagement, vis-à-vis relating tasks to prior knowledge, requesting clarification, and using analogies (Lee & Anderson, 1993). Goal orientation has also been used as a measure of cognitive engagement, and could be assessed through teacher observation (Ames & Archer, 1988; Pintrich & DeGroot, 1990). The idea of differentiating between performance (task) goals and mastery (learning/understanding) goals parallels other models of engagement as well, including authentic vs. strategic engagement (Schlechty, 2011) and substantive vs. procedural engagement (Nystrand & Gamoran, 1991). Nevertheless, only one of the three teacher reports included a measure of cognitive engagement (Fredricks et al., 2011).

Self-report measures are the predominant method of assessing student engagement. Of 14 available self-report measures, 11 included behavioral engagement, ten included affective engagement, and eight included cognitive engagement (Fredricks et al., 2011). Items for each sub-type of engagement were Likert-type items, including statements such as “I work several examples of the same type of problem when studying mathematics so I can understand the problems better” (cognitive engagement, on Attitudes Toward Mathematics-ATM), “I feel excited by the work in school” (affective engagement, on Student School Engagement Survey-SSES), and “I outline the chapters

in my book to help me study” (behavioral, on Motivated Strategies for Learning Questionnaire-MSLQ) (Appleton et al., 2006; Fredricks et al., 2011; Pintrich & DeGroot, 1991).

Though the use of self-report measures is often criticized in psychological research due to potential desirability biases (Chan, 2009; Field, 2013), self-report is the preferred method of assessing student engagement. While behavioral engagement can be observed, cognitive and affective engagement are problematic to discern from an external perspective. Classroom observations and teacher reports can only infer cognitive and affective engagement (Appleton et al., 2006; Finn & Zimmer, 2012; Fredricks et al., 2004; Linnenbrink & Pintrich, 2003). For example, “effort” is troublesome for external observers to assess, since it reflects both observable phenomena (quality and quantity of work) and internal processes (level of understanding, connections with prior knowledge, etc.) (Fredricks et al., 2004). Furthermore, research has suggested that student and teacher perceptions often differ, so this calls into question whose perspective is more valid (Fraser, 1982; Lee & Reeves, 2012; Skinner et al., 2008). One study found that students reported being more behaviorally engaged, and more emotionally disengaged, than their teachers observed them to be. (Skinner et al., 2008).

Despite concerns about the validity of external observations of engagement, there are other compelling reasons to use student self-report to assess this construct. Intuitively, it makes sense to ask students about their perceptions of their learning experiences: “there is something fundamentally amiss about building an entire [education] system without consulting at any point those it is ostensibly designed to serve” (Cook-Sather, 2002, p. 3). Above and beyond other variables, student perceptions have consistently shown strong

predictive validity for engagement and achievement (Fraser, 1982; Greene, Miller, Crowson, Duke, & Akey, 2004; McCombs, 2010; Uekawa et al., 2007). If a student reports that he or she does not feel safe expressing opinions in a class, this perception predicts decreased engagement and achievement. This suggests that student perceptions are important, regardless of their accuracy (Mahatmya et al., 2012). Other research has confirmed that students are able to give valid assessments of classroom characteristics, and that they are able to distinguish between what they like and what they need (Uekawa et al., 2007; Walberg & Hartel, 1980). Early adolescent students are developmentally able to use metacognitive skills to reflect on their cognitive engagement, vis-à-vis the use of multiple strategies. (Fredricks et al., 2004).

Validity and reliability. As engagement is a nascent construct, its operationalization in psychometric instruments is still in its infancy as well. Facilitators, indicators, and outcomes of engagement are often confused (Eccles & Wang, 2012; Fredricks et al., 2004). Facilitators can be thought of as all of the things that affect or predict engagement, such as teacher warmth, student autonomy within the lesson, and student interest in the topic of the lesson (Lam, Wong, Yang, & Liu, 2012). Indicators are those things that suggest engagement is occurring, such as the use of multiple learning strategies (cognitive engagement), enjoyment (affective engagement), and on-task behaviors (behavioral engagement) (Lam et al., 2012). Outcomes are those things that result from student engagement, such as graduation and increased attendance. Thus, facilitators (motivation, other constructs and contextual factors) predict indicators of engagement, and lead to several possible outcomes (graduation, increased attendance, achievement) (Eccles & Wang, 2012). Nevertheless, outcomes such as attendance and

drop out rates have been used as indicators of engagement (Finn & Zimmer, 2012).

Available instruments for engagement reflect this confounding of the iterative aspects of engagement. Some engagement measures assess predictors of engagement, such as student perceptions of their abilities and teacher expectations; while others assessed consequences of engagement, such as attendance, boredom, and graduation (Finn & Zimmer, 2012). Eccles and Wang (2012) warned about the lack of clarity in the operationalization of engagement:

If “engagement” encompasses everything from feeling like one belongs in the school to doing one’s homework, or to participating in the school band, then almost anything we do to improve schools can be seen as an intervention to increase engagement. (p. 138)

In addition to the issue of finding the most proximal engagement indicators, engagement instruments show other construct operationalization issues, reflecting clarity about behavioral, affective and cognitive engagement. Some measures use the same scale items to represent different types of engagement. The Attitudes Toward Mathematics (ATM) and the Motivated Strategies for Learning Questionnaire (MSLQ) include items assessing self-regulation—a blend of cognitive and behavioral engagement (Fredricks et al., 2011). In other measures, the same term describes different types of engagement (Reschly & Christenson, 2012). For example, while some instruments considered extent of participation to represent behavioral engagement, others considered it indicative of cognitive engagement (Fredricks et al., 2011).

Furthermore, there is disagreement about how best to consider behavioral, affective, and cognitive engagement. Some researchers have argued that engagement

should be measured as a holistic construct, and that it is ill advised to consider behavioral, affective, and cognitive components separately (Csikszentmihalyi, 1990; Fredricks et al., 2004). Others have argued that measuring sub-types of engagement might allow for targeted interventions to improve engagement for individual students (Appleton et al., 2006). Of 21 instruments, five assessed all three components, and one yielded a single score reflecting the combination of the three components (Fredricks et al., 2011). Some researchers posited that disengagement should have its own scale; that disengagement is not simply the absence of engagement (Martin, 2007; Reschly & Christenson, 2012). One teacher report measure (Engagement vs. Disaffection with Learning-EvsD) reflects both positive and negative aspects of behavioral and emotional engagement (Fredricks et al., 2011).

Despite operationalization concerns, the 21 engagement measures reviewed by Fredricks et al. (2011) showed promising levels of reliability and validity. The internal consistency of the majority of the 21 measures showed Cronbach's alphas from .70-.80. Inter-rater reliabilities were reported for all four classroom observation measures, with three showing acceptable ranges of .80-1. Additionally, five instruments showed construct validity by way of correlations with other measures. For the EvsD, the correlations were stronger for behavioral engagement than emotional engagement, confirming that outward behaviors are easier for external observers to assess. Criterion validity was suggested by positive correlations between 13 of the 21 tests with measures of student achievement. Two of the 21 tests reported positive correlations between student engagement measures and student attendance.

The factor structure of several of the engagement instruments has been analyzed. Some factor analyses reveal a number of factors, but many suggest three factors, congruent with the emerging consensus on a three-faceted model. Betts and his colleagues (2010) conducted a confirmatory factor analysis of the Student Engagement Instrument (SEI), which revealed the most parsimonious fit with a five-factor model of engagement (Betts, Appleton, Reschly, Christenson, & Huebner, 2010). Three factors were categorized as affective engagement: teacher-student relationships, peer support for learning, and family support for learning. Two factors were categorized as cognitive engagement: control and relevance of school work, and future aspirations and goals. This study also confirmed the factorial invariance of the instrument across age groups, which is an important step to establishing construct validity. However, many of these factors, and their related items, do not assess items that are most proximal to engagement (Veiga et al., 2014).

Factor analyses of ex post facto data from the Maryland Adolescent Development in Context Study (MADICS) support the factorial validity of the three-faceted model (Wang & Holcombe, 2010; Wang et al., 2011). In one study, six factors were identified, coalescing under a second-order factor structure with behavioral, affective, and cognitive factors (Wang et al., 2011). Behavioral engagement was comprised of factors representing attention and compliance; affective engagement comprised of belonging and valuing; and cognitive engagement comprised of self-regulation and strategy use. Another study confirmed a three-factor structure of the Student Engagement Scale for middle school and high school students (Doğan, 2014). Until clarity about the construct improves, researchers are advised to analyze particular instruments, and the items within,

in terms of the fit with their accepted models and research purpose (Azevedo, 2015; Fredricks et al., 2004).

Grain size. While the engagement construct has broadened over time from a focus on behavioral engagement in dropout prevention research to a broad meta-construct, it has also shown a narrowing in the grain size of interest, from school and subject-specific engagement to classroom and personal-level engagement (Greene, 2015; Lau & Roeser, 2008; Sinatra et al., 2015). Intuitively, an instrument measuring a student's general attitudes toward science would have less utility than longitudinal measures of that student's engagement with specific science tasks, as the latter would suggest how to design or change activities to promote engagement.

For engagement at the classroom and personal level, Sinatra et al. (2015) proposed that measurement approaches could be visualized along a continuum, comprised of three general theoretical orientations. On one end is a person-oriented approach, grounded in cognitive and motivational frameworks; such an approach favors student self-report. On the opposite end of the continuum is a context-oriented approach, which is informed by situated and sociocultural theoretical frameworks. In this approach, external observation is favored via discourse analysis, observations, and teacher ratings. A blend of the two can be found in the person-in-context approach. Methods used in this approach include observations of interactions, triangulated self-report, and the experience sampling method (ESM). ESM is a unique blend of student self-report at numerous timed intervals during class, matched up with external observation of contextual features occurring within the task or class at the times of student self-report (Larson & Csikszentmihalyi, 1983). As such, ESM provides a longitudinal measure of a student's

changing engagement with contextual features. A consideration of using this approach within the science classroom is the extent to which it would interrupt the flow of thought and activity with complex science activity. Another interesting approach to measuring engagement at a smaller grain size is to use a person-centered, rather than a variable-centered approach. Rather than identifying variables that predict engagement, the researcher would use inverse factor analysis to identify personality-types that predict engagement (Lau & Roeser, 2008).

Review of Engagement Research

Engagement is a burgeoning area of research in education. It has proven to be a robust predictor of many student outcomes: learning, grades, test scores, retention, and graduation (Bresó et al., 2011; Chang et al. 2007; Finn & Rock, 1997; Finn & Zimmer, 2012; Fredricks et al., 2004; Furrer & Skinner, 2003; Nolen, 2003). Engagement is intuitively understood by educators as malleable and essential for learning (Finn & Zimmer, 2012; Singh et al., 2002; Skinner & Pitzer, 2012). Furthermore, research has suggested that it is responsive to school and teacher practices, allowing for the possibility to improve achievement for students who are not highly engaged (Finn & Zimmer, 2012; Raphael, Pressley, & Mohan, 2008). Additionally, researchers have observed a reciprocal relationship between teacher and student involvement; disengaged students received fewer positive teacher behaviors that encourage and promote engagement, furthering the problem (Skinner & Belmont, 1993; Skinner & Pitzer, 2012). This last finding seems particularly cogent to the issue of student disengagement with middle school science, considering student self-reports of declining nurturant qualities in their teachers after the transition to middle school (Midgley et al., 1989).

The studies in this literature review were selected to represent the range of research about science engagement. They are grounded in a variety of theoretical frameworks, use different methodologies, and include students from middle childhood through late adolescence. The three-faceted engagement model guided the interpretation of each article (Fredricks et al., 2004). In some studies, the conceptualization of engagement was explicitly stated, while in others, the three-faceted model was used to interpret the author's conceptualization. Selected studies emphasize malleable instructional practices rather than static characteristics such as school size, or community characteristics. The studies are grouped into four categories: meta-analyses, studies linking engagement with achievement, studies examining predictors of engagement, and qualitative research.

Meta-Analyses.

Kumar Study. Kumar (1991) conducted a meta-analysis of 16 studies, between 1979 and 1990, investigating the relationship between science instructional methods and engagement. Studies were excluded from the meta-analysis if statistics were not reported or inferentially inadequate (means with no standard deviations, etc.). The studies represented 4518 students and 376 teachers in the United States and Australia. In this meta-analysis, instructional methods included both teaching and management strategies; assessed through observation, coding, and narratives. Engagement was conceptualized as behavioral engagement, and referred to as “on-task” engagement, which is consistent with early research on the engagement construct (Anderson, 1975; Stallings, 1980).

The mean correlation coefficient for the 16 studies was 0.48 (minimum = -.01, and maximum = .83), and 89% of the coefficients were positive. Instructional methods (n

= 70) found in three or more of the 16 studies were grouped together and coded into 17 variables, while instructional methods found in two or fewer articles were coded into 22 variables and considered in a separate group. The highest correlation coefficients for the 17 instructional variables were for giving directions and explanations ($r = .73$), reinforcing and encouraging student effort ($r = .66$), and demonstrating ability to work with individuals and groups ($r = .66$).

A one-way ANOVA was conducted on the 17 instructional variables. The effect was nonsignificant ($F(16,15) = 1.02$; $p = .45$), suggesting that the instructional variables did not differ significantly in their relationship with student engagement in science. A one-way ANOVA investigating the relationship between grade level and engagement was also nonsignificant ($F(1,8) = 2.31$, $p = .17$), though the comparison groups—grades 5-8 and grades 6-8—were overlapping. Studies, pre- and post-1985 (reflecting division based on “A Nation at Risk”) were compared via one-way ANOVA, since publication year had been shown to have effects in previous meta-analyses (Wilson & Rachman, 1983). No statistically significant effect was found here, either ($F(1,81) = 1.81$, $p = 0.99$). Engagement instrument used ($N = 7$) did not show a significant result ($F(1,14) = 9.98$, $p = 0.32$) when considering the most commonly used instrument, the Transaction in Science (TIS) instrument against five others (Kumar, 1991).

While an abundance of nonsignificant results might seem cause for concern, instructional method and publication year might arguably be the only variables for which one might have expected to see a significant result. However, it was possible that the coding of instructional methods into 17 categories reflected a grouping that was not warranted or valid. Also, though engagement is a relatively new construct, one might

have expected publication year to correlate with engagement, considering the one-dimensional focus in the early years on behavioral engagement. However, the years of the meta-analysis fell within the early “on-task” focus for engagement research. Given that 89% of the correlation coefficients were positive, and that the mean correlation coefficient was a moderate .48, a logical recommendation would be to perform an updated meta-analysis utilizing the more current models for engagement.

Links Between Engagement and Achievement.

Chang et al. Study. Chang et al. (2007) conducted a longitudinal study, using ex post facto data from the National Education Longitudinal Study (NELS) for 12,144 middle and high school science students (Curtin, Ingels, Wu, & Heuer, 2002). The authors investigated the relationship between science engagement and science achievement. Ethnicity, gender, socioeconomic status (SES), self-concept, and locus of control were considered as covariates. Three composite engagement variables were created from NELS questionnaire items showing face validity with engagement: Engagement 1 (E1: student choice in curricular activities), Engagement 2 (E2: intellectual involvement in science learning), and Engagement 3 (E3: involvement in routine whole-class seat work). E2 parallels cognitive engagement, while E3 parallels behavioral engagement. E1, however, might be conceptualized in the Eccles and Wang model as a facilitator, rather than an indicator of engagement (2012). The selected science achievement measure used item response theory (IRT) to correct for examinee’s ability in relation to item difficulty, ensuring that scores at different points in time were comparable. A baseline measure was established (8th grade, 1988) and compared to follow-up data in 10th and 12th grade.

Data were inferentially analyzed using correlation and hierarchical multiple regression. Correlations between engagement composite variables and IRT scores were statistically significant, but minimal. In 10th grade, E2 (cognitive) had a small correlation with IRT scores ($r = .09, p < .01$), and E3 (behavioral) had a similar, small correlation. However, E1 (student choice), had a small negative correlation with IRT ($r = -.12, p < .01$). In 12th grade, E1 had a small positive correlation, and E2 and E3 had small negative correlations. The beta-weights in the hierarchical multiple regression model for 10th grade, while significant, confirm that engagement composite variables were not strong predictors of achievement (E1: $\beta = -.15$, E2: $\beta = .04$, E3: $\beta = .05$). Two level longitudinal analyses revealed that while engagement did not predict test scores, engagement did have a significant association with scores at the growth rate ($\beta = 0.02, p < .01$). In other words, students showing high science engagement increased their scores faster than those with low engagement scores. However, this result only held true for Caucasians in the study. Also interesting was that self-concept, locus of control, and engagement declined over time, while IRT scores continued to grow.

This study had a number of positive aspects, including the use of longitudinal data and a large, varied, national dataset. While there were many significant findings concerning engagement and achievement, the sample size was large ($N = 12,144$), making significance easier to attain. Furthermore, the items for the composite engagement variables were selected based only on face validity. For example, one item in E2, the cognitive engagement composite, was “watch the teachers demonstrate an experiment.” Arguably, this observation should be considered a behavioral engagement measure. The use of a measure with stronger construct validity with the sub-types of

engagement would be advised. However, such a study would fail to have the large and diverse sample size afforded by the ex post facto NELS data. Additionally, self-reports of engagement were subject to the same shortcomings that all self-reports are; triangulation with a second data source would have been beneficial.

Singh et al. Study. Singh et al. (2002) examined the effects of motivation, attitude, and academic engagement on achievement for eighth grade math and science students. The authors' conception of academic engagement represented an amalgam of behavioral and cognitive engagement. The inclusion of attitudes represents an aspect of affective engagement, and motivation includes several aspects of engagement, as previously discussed. It is also the authors' belief that motivation and academic engagement are reciprocally related.

The authors drew a 25% random sample from the original 8th grade NELS data from 1988 (Curtin et al., 2002). After eliminating cases for which science attitudes, engagement, and motivation items were missing, 3,227 cases were included in the study. Two composite motivational variables were created from items on the NELS questionnaire data, and analyzed via confirmatory factor analysis. Motivation 1 (M1) represented attendance related items, such as how often the student missed school, skipped school, or was tardy. Motivation 2 (M2) represented preparedness items, such as coming to class with materials and homework. One attitudinal variable (A1) was created from three items about looking forward to math/science class, utility of math/science to the student's future, and boredom. Lastly, the engagement variable represented academic time, including time spent on homework and watching TV. Academic performance represented class grades and standardized test scores for math/science.

Structural equation modeling was used to determine direct and indirect effects of the selected factors on each other and on achievement. The overall model of two motivational variables, one attitudinal variable, and one engagement variable explained 38% of the variance in science performance. Engagement had the strongest direct effect on student learning ($\beta = .61$). Attitude toward science had the next largest effect, though indirect ($\beta = .32$). Both motivational variables had strong indirect effects on science performance (M1: $\beta = .31$ and M2: $\beta = .24$). The results of this structural equation analysis support the notion that motivation and attitudes were strong facilitators of engagement; the effects of M1 ($\beta = .50$), M2 ($\beta = .40$) and science attitudes ($\beta = .53$) reflected the strongest total effects (direct plus indirect) on engagement in the model.

The benefits of this study included the large, nationally representative NELS data and the use of structural equation modeling to examine non-experimental data. However, the author's conception of engagement was limited and indirect, in that "academic time" was represented by time spent on homework and time spent watching TV. Neither of these items reflects direct engagement with school, and both are subject to social desirability bias. While the inclusion of science attitudes in the model might appear to represent the inclusion of affective engagement, Eccles and Wang might argue that attitudes should be considered an outcome of engagement (2012).

Predictors of Engagement

There are a number of factors that could potentially predict a student's engagement in school, including both personal beliefs and attitudes; as well as external factors such as family, teacher, peer, classroom, and school characteristics. One broad study assimilated self-report data from 42,754 students across the United States who

completed the High School Survey of Student Engagement (HSSSE) to determine classroom and teacher characteristics that best engaged students (Yazzie-Mintz & McCormick, 2012). The results suggested that teacher lecture was least engaging to students; while discussion and debate, group projects, and projects involving technology were most engaging. The reasons students cited for being bored included uninteresting material, irrelevant material, not enough interaction with the teacher, and the work not being challenging enough. Personal beliefs and attitudes, such as self-efficacy, goal orientation, and fear of failure can predict engagement as well (Ames & Archer, 1988; Bandura, 1977; Caraway & Tucker, 2003).

Additionally, there are potential predictors that are specific to science engagement, such as epistemic cognition, scientific attitudes, topic emotions, gender and minority issues, and misconceptions (Greene, 2015; Greene & Miller, 1996; Jaber, 2014; Sinatra et al., 2015). Epistemic cognition refers to a student's methods and beliefs about knowing and knowledge. For example, a student with a relativistic viewpoint may find it more difficult or unpleasant to engage in science, which is a more positivist domain. Scientific attitudes are general, evaluative feelings about science or scientists. A student who believes that scientists are morally corrupt are less likely to engage. Topic emotions are feelings about specific ideas in science, such as climate change, genetically modified foods, and the like. Gender and minority issues can influence science engagement in that female students may not see themselves as able or worthy to participate in the science endeavor.

Uekawa et al. Study. Uekawa et al. (2007) explored student engagement in urban high school math and science classes. This study investigated how a variety of factors

predicted student engagement during class activities. A convenience sample was comprised of 320 students, representing eight schools in four geographic areas—Chicago, El Paso, Memphis, and Miami. Three methods were used to collect data: student focus groups, classroom observations, and the experience sampling method (ESM). Students wore beepers for five contiguous days. When the beepers activated, students completed a Likert-type survey comprised of affective and cognitive items; classroom observers coded the type of activity occurring (i.e., seatwork). The authors achieved an average of 6.8 observations per student over five days, which resulted in 2,360 engagement observations. The analyzed sample was reduced to 1,936 cases, due to missing data.

Engagement data, converted to *Z* scores, were analyzed inferentially using several hierarchical multiple regression models. Variance was found to be an effect between individuals (51%) as well as within individuals (39%), with little explained by teachers and classes (10%). As basic covariates were considered in the model, the between class variance was reduced to zero, suggesting support for ESM as a method to examine temporal and contextual perceptions of engagement within individuals. When teacher-controlled variables were considered in the model, group work showed a higher level of engagement (.20*SD*) compared to lecturing, seatwork, and testing. Some temporal effects were noted; lecture was typically used at the start of class, with group work occurring toward the end of class.

When student perceptions and class conversations were considered in models, some main effects were suggested. Students were more engaged with classwork that had relevance to their present concerns (.17 for everyday lives; .16 for tests) than to their future (negligible differences for college and career). Students were highly engaged when

they felt cooperative (.36), competitive (.25), not confused (.19), and not sleepy (-.41). When students had academic conversations with teachers, their engagement was higher than during silent moments (.23) or than moments of social talk with classmates (.45). As group work was the teacher-controlled variable with the largest main effect, the authors then examined the reduction in the advantage of group work to lecture when these student-level variables were added to the model. Sleepiness and conversations each accounted for 20% of the group work advantage over lecture, and were thus considered the primary intervening factors.

This study examined many possible predictors and interactions predicting engagement. The ESM method used is a promising one for future engagement studies, as it allows for longitudinal, repeated measures data lacking in many self-report engagement studies. Results of this study suggested that while the majority of class time was spent in lecture and seatwork (75%), student engagement benefitted from group work. However, broad-sweeping generalizations from this study should be conducted with caution. The urban environment was the only one considered in this study. Furthermore, student participation in this study was voluntary and required parental consent. Thus, it is possible that the data reflected a positive engagement skew.

Lau and Roeser Study. Lau and Roeser (2008) examined the engagement and achievement of high school science students with a person-centered, rather than a variable-centered approach. They established subgroups of individuals with similar configurations of cognitive, motivation, and affective characteristics; and then explored the relationship of those groups with engagement and achievement. The creation of the groups was guided largely by Ford's motivational theory and Snow's aptitude theory, and

included many variables, such as personal goal orientation, task value, classroom emotions, test anxiety, competence-related beliefs, context beliefs, regulatory processes, and cognitive abilities (Ford 1992; Snow, 1992). Items to assess these variables were taken from the MSLQ, Patterns of Adaptive Learning Scales (PALS), and NELS (Curtin et al., 2002; Fredricks et al., 2004). Science test scores (NELS, National Assessment of Education Progress [NAEP], and Trends in International Mathematics and Science Study [TIMSS]), grades, classroom engagement, and extracurricular engagement were used to validate the groupings (Curtin et al., 2002).

Inverse factor analysis was used to establish groupings of individuals ($N = 318$) based on 39 attributes, and was conducted separately for boys and girls to determine the generalizability of the solution by gender. A two-factor solution was selected, resulting in four classifications. For girls, type 1 were able and confident ($n = 50$), type 2 were anxious and ego-involved ($n = 51$), type 3 were intrinsically-motivated and task-involved ($n = 19$), and type 4 were able but work avoidant ($n = 24$). For boys, type 1 were able ($n = 30$), type 2 had positive perceptions of the classroom ($n = 33$), type 3 were confident and task-involved ($n = 46$), and type 4 were anxious and ego-involved ($n = 30$). ANOVA was used to validate the groupings, with Newman-Keuls test used for post hoc comparisons. Group differences were significant both on the derivation measures used to distinguish groups, and on the validation measures.

The types that reported the highest engagement were type 1 boys, type 3 boys, and type 3 girls ($M_s = .43, .51, \text{ and } .52$, respectively). Some of the characteristics shared by these higher engagement types were high levels of competence-related beliefs, task orientation, intrinsic motivation, and positive perceptions of the classroom environment.

Type 4 boys and girls were uniformly low on engagement measures ($M_s = -.96$ and $-.89$, respectively), and reported low competence-related beliefs, task goal orientation, and intrinsic motivation. It is important to note that for type 4 boys and girls, the poor outcomes were not necessarily associated with low ability. These results suggested possible targeted interventions. For example, type 4 boys would not likely benefit from cognitive interventions, since their deficits (i.e. low task goal orientation) were largely motivational in nature. Type 4 girls, alternatively, showed below average ability, and would benefit from both cognitive and motivational interventions.

The advantage of studies such as this one is a more holistic conception of intrapersonal variables that might be predictors of engagement. Additionally, grouping students in this manner could allow for interventions targeted to the needs of specific groups. However, one recommendation would be to repeat this investigation with a more diverse sample (2/3 of the students' parents attended four or more years of college), and to obtain longitudinal data to determine how different types interact with contextual variables in the classroom.

Assor et al. Study. Assor, Kaplan, and Roth (2002) conducted an engagement study of 862 Israeli students in grades 3-8; students in grades 3-5 ($n = 498$) were considered separately from students in grades 6-8 ($n = 364$). They used simultaneous multiple regression to determine which teacher behaviors best predicted student affect and engagement. Self-determination theory guided the categorization of teacher behaviors as autonomy-supportive or autonomy-suppressing. Autonomy-supportive behaviors included allowing criticism, fostering relevance, and providing choice; while autonomy-suppressing behaviors including suppressing criticism, forcing meaningless

activities, and intruding. Each of these three categories within autonomy-supportive and autonomy-suppressing behaviors was determined to be distinct via smallest space analysis (SSA) of student self-report data. Engagement was defined as behavioral and cognitive, and was assessed through student self-report on Likert-style items. Because elements of affect (positive and negative feelings) were included in the student self report measure, they could be considered analogous to affective engagement.

The Assor et al. study reported standardized beta weights for their simultaneous multiple regression. For grades 3-5, the best predictors of behavioral and cognitive engagement were fostering relevance ($\beta = .25, p < .001$) and suppressing criticism ($\beta = -.20; p < .001$). Additionally, the best predictors of affective engagement, vis-à-vis positive and negative feelings were fostering relevance ($\beta = .39, p < .001$) and providing choice ($\beta = .19, p < .001$). Thus, fostering relevance is the biggest predictor of all three types of engagement for these grade 3-5 Israeli students.

For grades 6-8, the best predictors of behavioral and cognitive engagement were also fostering relevance ($\beta = .24, p < .05$) and suppressing criticism ($\beta = -.15, p < .05$). The best predictors of affective engagement were the autonomy-suppressing behaviors; intruding behaviors ($\beta = .38, p < .001$) and criticism suppression ($\beta = .24, p < .001$) best predicted negative student feelings. Thus, for grades 6-8, autonomy-suppressing behaviors best predicted decreased affective engagement, while fostering relevance best predicted behavioral and cognitive engagement.

A comparison of grades 3-5 with grades 6-8 revealed other important findings. One is that fostering relevance is the best predictor of all types of engagement for both grade levels. In fact, fostering relevance was more important than providing choice in

increasing positive feelings. Interestingly, providing choice did not have a statistically significant relationship with behavioral or cognitive engagement for either age group. The authors proposed that these findings clarify what is meant by autonomy: “the essence of autonomy enhancement is not minimisation of the educator’s presence, but making the educator’s presence useful for the student who strives to formulate and realise personal goals and interests” (Assor et al., 2002, p. 273). In other words, there is a misconception that enhancing a student’s autonomy means minimizing the role of the teacher in education. This study suggested that increased student freedom, such as through incorporating student choice into assignments, is not as effective as fostering the relevance of those assignments.

Also, while providing choice was a statistically significant predictor of positive feelings in both middle childhood students ($\beta = .19, p < .001$) and early adolescent students ($\beta = .27, p < .001$), it was not a statistically significant predictor of behavioral or cognitive engagement. Taken as a whole, the autonomy enhancing and autonomy suppressing behaviors accounted for 51% of the variance in positive feelings, 41% of the variance in negative feelings, but only 19% of the variance in behavioral and cognitive engagement for students in grades 6-8. Similar percentages resulted from the grades 3-5 data. One possible conclusion is that the autonomy enhancing and suppressing behaviors are better predictors of affective than of either behavioral or cognitive engagement. Perhaps if autonomy supports were considered in conjunction with competence and relatedness—the other aspects of self-determination theory, they would more fully predict all types of engagement. Alternatively, one could question the role of emotions in engagement. Are emotions indicators of engagement or do they *predict* engagement?

One important aspect of early adolescent engagement is revealed by this study—suppression of autonomy is a stronger predictor of affective engagement for early adolescents than for middle childhood students. For example, intruding behaviors better predicted negative feelings for students in grades 6-8 ($\beta = .38, p < .001$) than for students in grades 3-5 ($\beta = -.12, p < .01$). In fact, intruding behaviors predicted slightly *decreased* negative feelings for students in grades 3-5. This suggested that educators of early adolescents should pay special attention to the suppression of autonomy in their classrooms, as these types of behaviors can negatively impact students' affective engagement in class.

Qualitative Research.

Olitsky Study. Olitsky (2007) conducted an ethnographic study, examining classroom conditions and teacher behaviors that encouraged positive interaction rituals (IR) for 33 eighth grade science students in an urban magnet school in Philadelphia. IRs are characterized by high levels of emotional energy, feelings of group membership, and sustained interest in the subject. The theoretical grounding for this work is in the community of practice model (Lave, 1991), which stresses social learning and co-construction of meaning. In this way, learning involves not only the development of knowledge, but also the acquisition of an identity associated with the group.

The author became a participant-observer during the 2001-2002 school year, collecting data via videotape, field notes, student work, interviews, and informal conversations. The author participated in the study by leading a weekly science review session, and occasionally co-teaching the class. The creation of successful interaction rituals and a community of practice was challenging due to the competitive nature of the

school. Not only was school admittance selective (based on test scores, etc.), but also many eighth grade students did not make the cut into the high school into which this magnet school fed.

One interesting IR vignette involved a whole class discussion about balancing chemical equations. A student, Anita, was asked to go to the board to balance a difficult equation. She struggled with the problem, while her classmates murmured quietly at their seats in conversation about the problem. At this point, what the author referred to as “entrainment” occurred, in which students became attuned to each other’s gestures and voices. Most students began paying attention to what Anita was doing, calling out suggestions like “you shouldn’t have erased that.” Anita persisted, despite her difficulties, and everyone began clapping at Anita’s resolution of the problem. This vignette showed all three aspects of successful IRs. Students of all levels contributed animatedly to the conversation, and provided suggestions (rather than answers, which would have ended the ritual). The author posited several reasons that this IR was successful. One reason was the difficulty of the problem, as easier problems had not elicited as much emotional energy. Additionally, peers were interacting with other peers, rather than students responding to a teacher who already knew the answer. What is interesting is that the topic, balancing equations, is not particularly relevant to students’ everyday lives, yet students persisted. This suggested that the focus in current literature on making science relevant to students might not have as much leverage as structuring the social situation in the classroom to allow for authentic, rigorous interactions. Furthermore, the use of IRs to investigate engagement in the science classroom reflects authenticity with the discursive

and social nature of the discipline. More research on science classrooms in schools with different demographics will add to the research on IRs.

Raphael et al. Study. Raphael et al. (2008) conducted a case study of nine sixth grade teachers who taught in a variety of content areas. The authors were interested in identifying and examining teacher practices that produced greater student engagement. Observations, teacher interviews, and class artifacts were used to both identify the different ways in which teachers attempted to engage their students, and to determine if those attempts were successful. Engagement was conceptualized as behavioral, with indicators such as on-task behaviors and conversations, though the level of the task was also included in the assessment of engagement; researchers only classified student behavior as engaged if the task required effort and thoughtfulness on the part of the student. For example, tasks were considered to require thoughtfulness and effort if students needed to think before acting, had to exert multiple attempts to achieve success, and/or needed to ask for help from peers or the teacher. A grounded theory approach was used until saturation was achieved, in which the authors found no new themes emerging in terms of methods to engage students. Forty-four practices were identified that promoted engagement, and they were collapsed into 14 categories; 17 practices were identified that discouraged engagement, and they were collapsed into seven categories.

From these observations, teacher interviews, and classroom artifacts, classrooms were classified as highly engaging, moderately engaging, and low engaging. A highly engaging classroom was defined by at least 90% of students engaged 90% of the time, moderately engaging classrooms by at least 50% of students engaged 50% of the time, and low engaging classrooms by less than 50% of students engaged with 50% or more of

students off task. Percentages were created by averaging multiple observations; teachers were observed from seven to 23 hours, with longer observation times to elucidate the instructional practices of the moderately engaging classrooms. Three classrooms were classified as highly engaging, four as moderately engaging, and two as low engaging.

Cross-case analyses were conducted to compare the instructional practices in each of the three categories in terms of their ability to affect or undermine engagement. Highly engaging classrooms were characterized by the variety of instructional practices used; all three highly engaging classrooms used all 14 categories of engagement-supporting instructional practices. Also, the highly engaging classrooms used none of the instructional practices characterized as engagement-hindering. The two low engaging classrooms differed in which engagement-supporting practices were used, though they both used far fewer of those practices—four for one teacher and seven for the other. The authors conclude that in efforts to identify what promotes student engagement, increasing the variety of practices may have more impact than ranking individual practices as more or less capable of increasing engagement.

Raphael et al. (2008) drew several other salient conclusions from this study. One is that engagement did not result simply from classroom management. In other words, defining appropriate behaviors, enforcing rules, and addressing misbehavior did not produce engagement. In fact, the authors found classroom management policies and procedures easiest to discern in the low-engaging classrooms, and implicit or unable to be identified in the highly-engaging classrooms. A possible explanation is that by promoting engagement, teachers create an environment in which (mis)behavior concerns are minimized. Another interesting finding is while engagement differed from teacher to

teacher, engagement did not differ greatly from one class to another for the same teacher. This runs counter to intuition that different students are easier or more difficult to engage. The researchers observed the same students disengaged in some classrooms and disengaged in others, and a group of students engaged with a curriculum in one teacher's classroom but disengaged with the same curriculum in another teacher's classroom. This suggests that engagement is indeed malleable, and that instructional practices can impact student engagement.

Logan and Skamp Study. Logan and Skamp (2008) conducted a longitudinal, observational case study of the engagement of 21 students as they progressed from year six (primary) to year seven (secondary) at a government school in New Zealand. Engagement was characterized by attitudes toward, and interest in, science; and thus, represented affective engagement. The researchers utilized multiple data sources, including personal interviews, same-sex focus groups, artifact observations, and a science attitude interest survey (Pell & Jarvis, 2001). When triangulating the data, the authors created a narrative for each student (one from year six and one from year seven) in order to understand each student's individual science engagement story. Additionally, the multiple data sources were analyzed by groups and sub-groups (year by year, advanced versus mixed ability groups, boys versus girls, etc.). The authors approached the study from a symbolic interactionism theoretical background which proposes that the way students define their world determines how they behave within it; thus, student perspectives and voice were valued highly.

At the end of primary school, the 21 students were enthusiastic and interested in science. Features of the classroom environment and teaching practices were consistently

cited by students as reasons for liking or disliking science. What was interesting about this study is that for the 21 participants, their interest in science did not decline from year six to year seven, while interest in science for a comparison, non-participant group **did** decline (the longitudinal qualitative study was part of a larger one involving cross-sectional data). The reason this is interesting is that there was no treatment or intervention. The authors proposed that this maintenance of science engagement for the students in the longitudinal study might actually represent a Hawthorne effect—students perceived themselves as special when the researchers showed interest in what they had to say. The Hawthorne effect is generally perceived to be a failure in quantitative studies, insofar as it calls into question the conclusions or generalizability of an investigation. In this case, the retained interest was not due to differential experiences by participants and comparison group. This finding lends support to the idea that autonomy supports, such as allowing for student voice, is a beneficial practice to engage middle school science learners (Cook-Sather, 2006; Ryan & Deci, 2000).

Summary

A three-faceted model of engagement—comprised of behavioral, affective, and cognitive components—has begun to appear more frequently in educational research literature. Not only have researchers adopted such a model, but reliable and valid psychometric instruments have also been developed that support one or many of the three facets. Most importantly, engagement has intuitive appeal and comprehensibility to educators, and has been shown to have predictive validity for achievement and a number of other achievement outcomes. As student attitudes toward and engagement in science decrease most drastically at or after the middle school transition, research about methods

to positively impact student engagement is warranted. A synthesis of research about engagement predictors, guided by this three-faceted model is a logical next step in advancing the coherence of the engagement construct.

Chapter 3: Research Methods

A meta-analysis of engagement research is warranted at this time in the evolution of the engagement construct. An emerging consensus about the operationalization of the construct is present in the research literature, yet there is variety in how researchers measure, conceptualize, and discuss engagement in primary studies (Fredricks et al., 2004). A synthesis of engagement research, conducted through the lens of the three-faceted model of engagement, offers coherence to the existing body of research. A meta-analysis affords an examination of broad engagement patterns that cannot be accomplished by any single study. The present study represents an attempt to identify inconsistencies and omissions in engagement research. Meta-analytic methods are also effective at guiding theory development; despite an emerging consensus about the construct, an engagement theory with predictive power does not yet exist.

As a means of synthesizing disparate research literature, meta-analysis provides a number of benefits over narrative literature reviews. Rosenthal and DiMatteo (2001) suggested that it may be “too tempting for authors of narrative reviews consciously or unconsciously to select and describe studies to support their own understanding of the literature and/or their own established theoretical positions” (p. 62). The more studies considered in a literature review, and the more disparate the results in those studies, the more problematic it becomes to synthesize the research in a meaningful way. Different narrative reviews of the same body of research can yield markedly different results. Though meta-analysis is not immune to subjectivity concerns, it affords transparency and consistency about how studies are weighted and considered in the synthesis. Meta-analysis is a systematic, quantitative technique that allows the researcher to effectively

summarize results and quantify dispersion across multiple studies.

An additional benefit to meta-analysis for synthesizing research is its focus on practical, rather than statistical significance. Though statistical significance is the accepted metric by which primary studies are compared in the research community, statistical significance is somewhat arbitrary and is often misinterpreted. For example, a non-significant p -value could reflect that there is no effect or a small effect, but it could indicate that there was a large effect in a study with a small sample size (Borenstein, Hedges, Higgins, & Rothstein, 2009). Because p -values confound effect size with sample size, it is difficult to use techniques such as vote counting of statistically significant versus statistically nonsignificant studies to evaluate a body of research. While an alpha value of .05 is standard and used almost universally in educational research, more liberal alpha values may be appropriate in smaller studies. Rosnow and Rosenthal (1989) acknowledged the arbitrary nature of alpha values: "Surely God loves the .06 nearly as much as the .05" (p. 1277). A consideration of practical significance via effect sizes rather than statistical significance via p -values provides a way to compare the practical meaning of results from one study to another.

Criticisms and Limitations of Meta-Analysis

Though meta-analysis is a powerful quantitative method to synthesize research, it is not without criticism. In fact, some researchers' criticism of the method is scathing—Feinstein (1995) referred to meta-analysis as "statistical alchemy for the 21st century," while Shapiro (1994) published an article titled "Meta-Analysis/Shmeta Analysis." A major criticism lies in skepticism that disparate studies can be validly summed and compared. It seems dubious that numerous studies could be accurately represented as a

single summary effect size. Even if the body of research were large enough to afford researchers the ability to combine only studies with many similar methodological and theoretical characteristics, summing effects minimizes potential meaningful differences between studies. One potential response to this criticism is that the purpose of meta-analysis is not simply to sum results, but also to evaluate the dispersion of effects and to suggest further clarifying studies in a particular area of research (Borenstein et al., 2009). In fact, depending on the meta-analytic question, dispersion may be of more interest than a summary effect.

Some criticize meta-analysis for comparing studies which should not be compared. Because studies can differ in a myriad of methodological, theoretical, and qualitative characteristics, it is challenging to compare the effect sizes from such diverse studies. Meta-analysis often asks larger questions than those addressed in primary studies (Borenstein et al., 2009; Schmidt & Hunter, 2015). Rosenthal expressed the same idea metaphorically by saying that combining apples and oranges makes sense if your goal is to produce a fruit salad (Borenstein et al., 2009). For example, this meta-analysis has a broad focus on comparing predictors of middle school science engagement. Thus, it is desirable to include a variety of middle school courses, teaching techniques, geographic locations, socioeconomic levels, etc. in the analysis. Glass, McGaw, and Smith (1981) responded to this criticism by suggesting that meta-analysis is not unique in pooling data; educational research frequently considers effects for populations with dissimilar individuals. By combining a breadth of studies, the meta-analytic researcher can assess how comparable and generalizable disparate studies are. Where anomalies are found, future research questions can be generated.

While criticisms about summarizing and comparing disparate studies question the method of conducting meta-analyses, other criticisms are more methodological in nature. One such criticism concerns the criteria for inclusion of studies. Criteria should both exclude low-quality studies and include important studies that relate to the research questions. However, the criteria for what renders a study of low quality are subjective. For one researcher, correlational research might be low quality; while for another, studies not published in peer-reviewed journals might be classified as low quality. Meta-analysis has been criticized for excluding important studies. In this meta-analysis, middle school student engagement is a focus; important studies about high school student engagement, for example, are not included. Researchers can address these criteria inclusion criticisms by clarifying the relationship of studies to the research question being asked and examining data for patterns related to quality.

Related to the issue of inclusion criteria is publication or availability bias. Studies that are statistically significant and/or show larger effects are more likely to be submitted and accepted for publication (Borenstein et al., 2009; Schmidt & Hunter, 2015). The bias for statistically significant results is apparent at all phases of the research process. Self-report research shows that researchers are more likely to submit articles showing significant results, reviewers rate statistically-significant studies more favorably, and editors are more likely to publish those statistically-significant studies (Coursol & Wagner, 1986; John, Loewenstein, & Prelec, 2012). A meta-analysis that integrates only published studies would be likely to overestimate possible effects. However, published studies are easy to procure through online databases and journals, while unpublished studies are logistically more difficult to obtain. The meta-analytic researcher must make

special efforts to find unpublished research such as conference proceedings and dissertations.

Despite the intuitive logic of publication bias leading to inflated summary effects, some researchers find no such inflation, or that preferentially including published results is not a source of bias. Rosenthal (1984) examined several hundred effect sizes from 12 meta-analyses, and found the mean effect size of the unpublished studies to be larger, while the median effect size was larger for the published studies. Similarly, other researchers found no difference in effect sizes between published and unpublished studies, or between small and large sample sizes (Hedges, 1992; Schmidt & Hunter, 2015; Schmidt, Oh, & Hayes, 2009). Alternatively, other researchers have suggested that publication bias could have a preferred effect—studies that are accepted for publication are likely to reflect stronger methodologies than those that are not (Schmidt & Hunter, 2015). Others have argued that while publication bias might result in an inflation of effect sizes, it will not produce type I error. One study of 302 psychological interventions showed less than 1% of those interventions resulted in no effect, while another study of 322 meta-analyses found that only 8 showed no or nearly no effect size (Lipsey & Wilson, 1993; Richard, Bond, & Stokes-Zoota, 2003).

Another perspective of publication bias is to consider the prominence of a particular hypothesis and the number of hypotheses within a study. This perspective is particularly germane to research on the construct of engagement. As a nascent construct, engagement is often ancillary to other outcome variables such as achievement. Thus, studies which reveal a statistically significant, positive effect on achievement could get published, regardless of the statistical significance of the engagement outcomes. In other

words, questions of interest in a meta-analysis may be irrelevant to the central hypotheses of the primary studies (Cooper, 1998). Similarly, primary research studies often test more than one hypothesis; the likelihood that all hypotheses would be nonsignificant is low (Schmidt & Hunter, 2015). This suggests that concerns about non-publication of statistically nonsignificant results may be merely theoretical. Nevertheless, a number of techniques were used to assess publication bias in this meta-analysis (see Publication Bias Analysis).

Literature Search Methods

The investigator conducted a comprehensive literature review to obtain both published and grey literature concerning student engagement in middle school science classrooms. Five databases were used to find published studies—Academic Search Premier, Education Full Text, Education Resource Information Center (ERIC), PsychInfo, and JSTOR. ProQuest Dissertations and Theses and Google Scholar were also searched as potential sources of grey literature. The investigator connected with scholars active in engagement research via social media sources such as LinkedIn and Google Plus to find further unpublished research.

Search terms for this meta-analysis effectively located studies about the correct research topic (engagement), grade level (middle school/junior high), and school subject (science). Subject terms that returned database results relevant to a student's engagement with school included *student engagement* and *learner engagement*. The use of these phrases, rather than simply *engagement*, was necessary to eliminate studies that concerned civic or political engagement. To locate studies for the proper grade level, checkboxes for “grades 5-9” or “early adolescence” were selected if available within

each database, and if not available, the Boolean search phrase “*middle school*” or “*junior high*” was used. The search term “*science*” was included as well to limit results to those including mention of the proper disciplinary topic (see Table 1 for a complete list of search terms by database).

Table 1

Engagement Search Terms by Database

| Database | Exact Terms and Phrases | Selections within Database |
|---------------------|---|--|
| ERIC | “learner engagement” “science” | Middle school Junior high Grades 5,6,7,8,9 |
| PsychInfo | “student engagement” “science” | Adolescence (13-17) |
| Education Full Text | “student engagement” “science” “middle school” or “junior high” | |
| JSTOR | “student engagement” “science” “middle school” or “junior high” | |
| Google Scholar | “student engagement” “science” “middle school” or “junior high” | |

Note. All database searches limited to 2006-2015.

The investigator examined the located studies to discern their congruence with the inclusion criteria. The titles or abstracts that clearly addressed a topic other than early adolescent engagement with school science were eliminated. If the characteristics of studies were unclear from the title or abstracts, the studies were retained for further examination. The reference lists of articles that did not meet the inclusion criteria were further mined for potential relevant studies.

Table 2

Engagement Assessment Instruments Included in the Literature Search

| <u>Name of Instrument</u> |
|---|
| Academic Motivations Scale (AMS) |
| Achievement Goal Questionnaire (AGQ) |
| Academic Self-Regulation Questionnaire (SRQ-A) |
| Approaches to Learning Questionnaire (ATL) |
| Approaches to Learning Science (ATLS) |
| Attitude Scale Toward Science (ASTS) |
| Attitudes Toward Mathematics Survey (ATM) |
| Dimensions of Continuing Motivation to Learn Science (DCMLS) |
| Effort and Persistence in Learning (EPL) |
| Experience Sampling Form (ESF) |
| Five Component Scale for Self-Regulation (FCSSR) |
| Intrinsic Motivation Inventory (IMI) |
| Motivated Strategies for Learning Questionnaire (MSLQ) |
| Patterns of Adaptive Learning (PALS) |
| Relative Autonomy Index (RAI) |
| Science Motivation Questionnaire (SMQ) |
| Secondary School Student Questionnaire (SSSQ) |
| Situational Interest (SI) |
| Science Achievement Influences Survey (SAIS) |
| Student Attitude to Science Survey (SASS) |
| Student Motivation Questionnaire (SMQ) |
| Students' Motivation Toward Science Learning (SMTSL) |
| Student Perceptions of Class Questionnaire (SPOCQ) |
| Test of Science-Related Attitudes (TOSRA) |
| Waering Attitudes toward Science Protocol (WASP) |
| <u>What is Happening in this Class? Questionnaire (WIHIC)</u> |

Because the engagement construct overlaps with other existing bodies of research, such as motivation, it is possible that research studies not explicitly measuring engagement could yield relevant data for this meta-analysis. For example, goal orientation and self-regulated learning are considered indicators of cognitive engagement. However, broadening the literature search to include the array of currently accepted

indicators of each type of engagement would be logistically burdensome. Informed by primary engagement research and reviews of self-report instruments used to assess indicators of engagement at the classroom level (Fredricks et al., 2011; Fredricks & McColskey, 2012; Veiga et al., 2014), the investigator included additional searches by assessment instrument in Google Scholar. Only assessment instruments that did not explicitly name “engagement” were included in additional searches. Boolean search phrases were generated by adding the name of the assessment instrument to the search phrase “*science*” and “*middle school*” or “*junior high*” (see Table 2 for a list of engagement instruments included in the literature search).

Inclusion/exclusion criteria. The investigator screened studies for various source, study, and methodological characteristics. As much as possible, studies that yielded pertinent information about early adolescent engagement with school science were retained. Studies were coded to reflect differences in source, study, and methodological characteristics for further analysis via descriptive statistics and meta-regression.

Source characteristics. Characteristics of the source of the study included language, date and publication status. Studies that were not in English or able to be translated into English were excluded. Google Translate was used to translate one study from Spanish to English (Liu, 2014). To reflect recent research that could be responsive to the seminal literature review of engagement by Fredricks et al. (2004), only studies published in 2006 or after were included. Both published and unpublished studies were included to address potential publication bias, though they were coded appropriately for further analysis to determine differences (see “Criticisms and Limitations of Meta-

Analysis”). Similarly, both peer-reviewed and non peer-reviewed studies were included and coded. These decisions were made to allow for the most complete consideration of the nascent engagement construct.

Study characteristics. A number of study characteristics were used as inclusion criteria. Included studies assessed indicators of engagement either explicitly or implicitly. The decision about whether a study implicitly measured engagement was informed by guidelines from the research literature (Fredricks et al., 2004; Skinner & Pitzer, 2012) (See Table 3). For example, one study implicitly assessed cognitive engagement vis-à-vis students’ mastery approach goals (Kahraman & Sungur, 2013). The investigator made the final determination if measures within each study showed face validity with accepted engagement indicators. Studies which included one, several, or a combination of engagement types were included. For example, the Spearman and Watt study (2013) assessed only indicators of affective engagement, the Wolf and Fraser study (2008) assessed behavioral and affective engagement separately, and the Zheng and Spires study (2014) assessed an amalgam of all three types of engagement.

As this meta-analysis examined the most practically significant predictors of engagement, studies were excluded if they did not include predictors of engagement. However, studies that did not assess engagement as the criterion variable were included if engagement was assessed as a mediator variable. This afforded the inclusion of many additional studies and study methodologies, such as structural equation models that focus primarily on achievement as a criterion variable (e.g., Mo, 2008).

Studies including predictors of engagement were included in the analysis if the predictors were malleable at the classroom or task level. For example, studies that

primarily examined socioeconomic status, grade level, or science content as predictors of engagement were excluded. When possible, less malleable characteristics were coded and considered as possible engagement moderators. Similarly, studies that examined students' attitudes toward science were excluded as such studies reflect a grain size larger than the classroom or task-level. However, if students' attitudes toward science were assessed in response to a specific classroom or task-level intervention, they were retained in the analysis. Studies assessing the impact of extra-curricular science interventions, such as science clubs, field trips, or summer programs, were only retained in the analysis only if the intervention identified a specific methodology that could be implemented within the classroom.

Table 3

Engagement Type Indicators

| Type | Indicators |
|------------|--|
| Affective | Attitudes Interest Situational interest Enjoyment Valuation |
| Behavioral | Time on task Participation Completion Compliance with teacher requests Persistence Effort |
| Cognitive | Goal orientation Reflective strategies Use of cognitive strategies (rehearsal, elaboration, critical thinking) Initiation of questions (agentic engagement) Self-regulation (monitoring, regulating) Flexibility Metacognition |
| All Three | Flow |

Content area and age range. Included studies assessed engagement of early adolescent students with science. The investigator included studies of student engagement from grades five through nine (ages 10-15) in order to be inclusive of alternative school configurations. Studies that assessed secondary science were evaluated; if they contained early adolescent data that was separable from high school data, they were retained in the analysis. High school science engagement studies were retained if the participants were restricted to grade nine. Similarly, studies that examined K-8 science engagement were evaluated; the investigator retained the study in the analysis if the early adolescent data was separate.

Instrumentation. Though there are a number of methods to assess student engagement, this meta-analysis excluded studies that did not assess engagement through student self-report surveys. In some cases, the self-report measure was designed specifically to assess engagement. In other cases, the self-report measure was a sub-scale or smaller portion of a larger instrument. Survey questions may or may not have been originally designed to assess engagement, but used for that purpose in the included studies. Such ex post facto aggregate measures of engagement were included only if the investigator or researchers determined that the aggregate showed face validity with engagement. Studies assessing student engagement through qualitative self-report methods such as journaling or focus group interviews were excluded.

Methodology and experimental design. Though experimental research with randomized assignment is the gold standard in many fields of research, it can be considered a threat to ecological validity in educational research (Bronfenbrenner, 1976). Experimental designs are rare in education, as random assignment is logistically difficult

in school settings. However, establishing causation is not required for an educational study to be useful. The purpose of many educational studies, including this meta-analysis, is not to claim cause and effect, but to identify relationships and the strength of relationships between variables, or to identify predictors of a desired outcome variable.

For the aforementioned reasons, a variety of methodological designs were included in this meta-analysis, despite their inability to establish causation. Experimental, quasi-experimental, repeated measures, correlational (e.g., correlational, structural equation modeling, and regression), and ex post facto study methodologies were included. The inclusion of single group repeated measures designs deserves further discussion. Though a seminal review of experimental and quasi-experimental research is often cited as a rationale for excluding single group, repeated measures designs (Campbell & Stanley, 1963), the single-group repeated measure design affords higher precision and power than independent group designs (Schmidt & Hunter, 2015). Despite a number of potential threats to validity, some researchers suggest that such designs do not often suffer from those threats (Schmidt & Hunter, 2015). Lipsey and Wilson (1993) found that this study design overestimates effect sizes by up to 61%. However, repeated measures designs result in mean gain *difference* effect sizes, while independent group designs produce mean gain effect sizes. Thus, the effect sizes reported from repeated measures designs will naturally be inflated, as the effects are calculated by dividing the mean by the standard deviation of the difference. Fortunately, formulas exist to correct and standardize effect size measures from repeated measures designs in order to compare them to other research designs (Schmidt & Hunter, 2015). The type of study was coded to allow for separate analysis and/or moderator analysis of study methodology.

Statistical considerations. Meta-analysis allows the researcher to quantify the size of an effect and its precision. Thus, studies in a meta-analysis must report statistics that afford the opportunity to determine both of these characteristics. To quantify the size of an effect, included studies should report effect sizes directly, or statistics necessary to calculate effect sizes. Effect sizes fall into two categories: measures of group differences and measures of association. Common measures of group differences are Cohen's d , Hedges' g , or Glass' Δ (Ellis, 2010). To calculate group difference measures, the minimum required statistics include means and standard deviations. Studies that did not report a standard deviation, but did report standard error and sample size, were also included, as it is possible to calculate the standard deviation from the standard error and sample size. Alternatively, F -statistics and t -values can be used in conjunction with sample sizes to calculate group difference effect sizes. In order to appropriately weight studies and determine precision via confidence intervals, studies should provide a sample size. In cases where required statistics were not reported in a study, the investigator attempted to obtain the missing statistics through personal communication with the primary author of the article.

As this meta-analysis included a variety of study designs, studies that reported measures of association as effect sizes were also included. Measures of association include correlation indices, such as Pearson's r , or proportion of variance indices, such as r^2 , R^2 , or η^2 (Ellis, 2010). Multiple regression studies and structural equation modeling studies reporting beta weights were also included, as β can be considered a substitute for Pearson's r (Becker & Wu, 2007; Borenstein et al., 2009). Thus, studies reporting measures of association and sample sizes were included in this meta-analysis.

Study characteristics and coding. After selecting studies based on the inclusion criteria, the investigator coded a number of possible predictors, moderators, or covariates. In addition to reporting descriptive statistics for these covariates, the investigator conducted a meta-regression on covariates with at least ten studies, following suggested minimum variable requirements for multiple regression (Borenstein et al., 2009; Field, 2013). The investigator collapsed or categorized variables that were either continuous or had fewer than 10 studies. For example, the reliability of the student self-report instrument was collapsed into five groups: studies referencing external instrument reliabilities, those with instrument reliabilities greater than .7, less than .7, or not reported. This categorization reflects general recommendations for minimum criteria for reliability of attitudinal instruments (Nunnally, 1978). See Table A1 for a coding schematic (Appendix A).

Source characteristics. Studies were coded for publication as unpublished or published and for publication type as non-peer reviewed or peer-reviewed.

Study characteristics. Studies were coded to reflect predictor type and engagement conceptualization as well as a number of possible covariates. The investigator selected variables for coding based both on evidence from prior research suggesting a relationship with the variables and engagement, and also the likelihood that such variables would yield at least 10 studies per variable. The latter is a minimum requirement to conduct multiple regression analyses of the covariates (Borenstein et al., 2009).

Predictor classification: Type. The investigator coded the type of predictor or intervention into four categories: instructional method, technology, class characteristics

and social characteristics. For example, autonomy support was coded as a class characteristic predictor, while teacher relatedness was coded as a social characteristic predictor. The investigator selected only those predictors that were malleable at the classroom or task level. For some studies, only a portion of selected predictors were included (see Table B1 for a detailed list of predictors by study).

Predictor classification: Self-determination theory. The investigator coded the type of predictor or intervention as primarily one facet of self-determination theory: autonomy, competence, or relatedness. For example, project-based learning was coded as primarily an autonomy intervention, as project-based learning can differ on competence scaffolding and degree of peer interaction. See Table B1 (Appendix B) for a list of coded characteristics by point estimate.

Engagement conceptualization. Studies were coded for which facet(s) of engagement were measured. Seven categories were created which reflected different permutations of the engagement facets: 1=behavioral, 2= affective, 3=cognitive, 4=behavioral and affective, 5=behavioral and cognitive, 6=affective and cognitive, and 7=all three facets (see Table 3 for indicators of engagement used by the investigator). If studies assessed two facets of engagement, but reported them separately, they were entered and coded separately. For example, if a study assessed both behavioral and affective engagement, and reported separate scores for each, the behavioral measure was coded as a 1 and the affective measure was coded as a 2, rather than the study being coded as a 4. Alternatively, if the study produced an aggregate measure of behavioral and affective engagement that could not be separated, it was coded as a 4. Each aspect of engagement was thus considered separately in the analysis when possible. The

investigator selected scales, sub-scales, or combinations of scales within each study to reflect measures of affective, behavioral, and/or cognitive engagement (see Table B1 for a detailed list of decisions by study). For meta-regression, the categories were collapsed into four categories to achieve the minimum number of studies per variable:

1=behavioral, 2=affective, 3=cognitive, and 4=two or more facets.

Participant and school characteristics. Studies were coded by school type, structure, setting, socioeconomic status, and geographic location. School type was recorded as unspecified, public, private, charter, independent, or alternative/other. School structure was coded as unspecified, elementary school, middle school, junior high, K-8, high school, or other/mixed. Descriptive statistics noting the specific grade/age level in the study were also recorded. School setting was coded as unspecified, rural, suburban, urban, or mix. Geographic location was coded as United States and not United States. The specific country was recorded and reported in the descriptive statistics. School socioeconomic status was coded as not specified, low (greater than 60% free and reduced lunch (FRL)), average (35-59% FRL), high (less than 35% FRL), or mixed. Last, the age of sample participants was recorded as 5th grade (10-11 years old), 6th-8th grade (11-14 years old), 9th grade (14-15 years old), or a mix of those categories.

Instrumentation reliability and validity. The investigator recorded both the Cronbach's alpha for the instrument, and whether or not the reliability measure was reported from external studies or as an internal measure within the study. For meta-regression, the continuous reliability data was collapsed into categories: 0=not reported, 1=references external instrument, 2=references external instrument reliability, 3=internal reliability < .70, 4=internal reliability > .70.

The investigator also recorded validity measures when given, and differentiated them as internal or external, low level (e.g., face/content) or high level (e.g., exploratory factor analyses, concurrent validity), and whether validity was determined by the investigator, study, or external study. The data was collapsed into categories: 0=not reported or some content face validity assessed by investigator, 1=face/content validity assessed by investigator, 2=face/content validity assessed by study, 3=reference to external measure, 4=reference to external measure validity, 5=internal reliability measure (EFA, CFA, etc.).

Methodology and experimental design. The investigator coded the type of study (1=correlation or regression, 2=single group pre-post, 3=quasi-experimental, 4=experimental). Though a common effect size, Hedges' g was used to compare all studies in the meta-analysis, this coding allowed for comparison of effect-sizes for different methodological designs—measures of group difference from the d -family, or a measures of association from the r -family.

Research Synthesis Methods

The investigator used Comprehensive Meta-Analysis (CMA), Version 3 (Biostat, 2015) to conduct the meta-analysis, an online effect size calculator (Wilson, 2015) for effect size calculations not offered within the program, and Microsoft Excel to perform sub-calculations and examine descriptive statistics. Statistics that were reported for individual studies included effect sizes, variances, confidence intervals, Z -scores, p -values and sample sizes. Studies were synthesized both to produce summary effects by type of engagement and to differentiate variance as observed, expected, or true. Sub-

analyses were conducted using meta-regression to determine the relationship between various predictors or moderators and engagement outcomes.

Calculations for individual studies.

Effect size calculation. Effect sizes are standardized measures that allow meta-analytic researchers to compare studies to one another using a common metric. There are two main categories of effect sizes: measures of group difference and measures of association. Measures of group differences, often called the *d*-family of effect sizes, include Cohen's *d*, Hedges' *g*, or Glass' Δ (Ellis, 2010). For each measure, a mean difference is divided by a standard deviation in order to produce a number that reflects the size of the effect, largely regardless of effect size or sample size (Ferguson, 2009). The following formula reflects the general formula for calculating a group difference effect size:

$$ES = \frac{\overline{x_1} - \overline{x_2}}{SD}$$

The decision about which mean is x_1 and which is x_2 is arbitrary. However, desired effects should be represented by positive numbers, and undesired effects by negative numbers.

The three measures of group measures—Cohen's *d*, Hedges' *g*, and Glass' Δ , differ in the standard deviation used to standardize the mean differences between groups. Cohen's *d* pools the standard deviations of the control and experimental groups; problems can arise when the two groups are of different sizes as both groups' standard deviations are given equal weight in the formula for Cohen's *d*. Additionally, Cohen's *d* can overestimate the effect size in small samples, which tend to be common in educational research (Borenstein et al., 2009). Glass' Δ uses the standard deviation of the control

group, with the rationale that the control group's standard deviation is likely to be closer to the population mean's standard deviation (Ferguson, 2009). However, in many educational studies, there is no specific control group, only groups to be compared to one another. Hedges' g weights each group's standard deviation by its sample size. Due to the biases of the other two measures, and due to the likelihood of unequal group sizes in educational research, Hedges' g was used as a measure of group differences in this meta-analysis.

Hedges' g was automatically calculated by CMA for the majority of methodological study designs included in this meta-analysis. The variety of entry formats within the program take into account considerations in producing comparable effect sizes, such as possible inflation of group difference effect sizes when compared with independent group designs (Schmidt & Hunter, 2015). The Campbell Collaboration online effect size calculator was used for calculations that were not offered by CMA, such as calculating an effect size when given t -values and sample sizes (e.g., for Vedder-Weiss & Fortus, 2011). For studies with incomplete statistics to calculate an effect size through CMA or the online calculator, the investigator made efforts to obtain the statistics through personal communication with the author, or to calculate the missing statistics. For example, if the standard deviations were missing, but the standard error and sample size were given, the investigator calculated the standard deviation: $SD = SE \times \sqrt{n}$.

As a number of study designs were included in this meta-analysis, measures of association were also used as effect sizes. Measures of association include correlation indices, such as Pearson's r , or proportion of variance indices, such as r^2 , R^2 , or η^2 (Ellis, 2010). Correlation indices are preferred to proportion of variance indices, as proportion

of variance indices are small and can be misinterpreted in education research (Schmidt & Hunter, 2015). Variables which account for a small variance in an outcome often have important effects on that outcome (Schmidt & Hunter, 2015). Proportion of variance indices were thus converted to measures of association directly, such as in the case of r^2 to r . Some meta-analysts propose that r be converted to Fisher's z scale before synthesis, as the correlation and variance are confounded (Borenstein et al., 2009). However, others suggest that this bias is trivial and there is no need for the transformation (Schmidt & Hunter, 2015). The investigator used Pearson's r in CMA to obtain Hedges g .

In multiple regression studies, β can be used as a measure of association. The use of β as an effect size is considered controversial by some researchers (Schmidt & Hunter, 2015), though it is gaining acceptance and popularity in social science research (Becker & Wu, 2007; Bowman, 2012). The statistic is highly correlated to Pearson's r , regardless of the number of variables in the regression equation (Peterson & Brown, 2005). The high correlation between β and r is even higher for smaller β values, which are common in education research (Peterson & Brown, 2005). For these reasons, a number of meta-analytic researchers have suggested that β values can be substituted directly for r (Becker & Wu, 2007; Borenstein et al., 2009; Bowman, 2012; Ferguson, 2009; Rosenthal & DiMatteo, 2001). The investigator included β values from regression, multiple regression, and structural equation models in this analysis.

Converting between effect sizes. Hedges' g was selected as a common metric for comparing studies included in this meta-analysis. However, a diverse array of research designs were included in the meta-analysis, including measures of association that differ in interpretation from measures of group differences, such as Hedges' g . For example, an

r of 0.5 is about twice as large as a d of 0.5 (Ellis, 2010). For this reason, measures of association from the r -family of effect sizes were converted to the d -family of effect sizes within CMA. Additionally, the Cohen's d values produced by the online effect size calculator from these calculations were converted to Hedges' g within CMA.

Variance. Variances are used both to calculate confidence intervals for effect sizes and to properly weight studies within meta-analyses. Variance is a measure of the amount of spread in a set of data. Variances can be calculated for effect sizes with high variances indicating lower precision and low variances indicating higher precision. A meta-analysis assigns more weight to precise studies and less weight to less precise studies (see "Synthesis of Studies" for weighting methods).

Confidence intervals. Confidence intervals identify the amount of uncertainty (or precision) in an effect size estimate. A confidence interval gives a range of values within which 95% of similar samples should fall. The upper and lower values for that range are given by the following formulae:

$$UL_Y = \bar{Y} + 1.96 \times SE_Y$$

$$LL_Y = \bar{Y} - 1.96 \times SE_Y$$

Where SE is the standard error of the effect size estimate. Confidence intervals for individual studies were represented graphically by forest plots.

Complex data structures. Educational studies are often complex, involving multiple subgroups, time points, and outcome measures. Thus, the meta-analyst must make a number of decisions about how best to synthesize complex data. There are two categories of decisions that must be addressed in meta-analyses—how to deal with complex data structures involving independent subgroups, and how to deal with complex

data structures involving non-independent groups, such as repeated measures or multiple outcomes for the same participants. These decisions can be influenced by how discordant the results are for the groups.

Independent sub-groups. For studies that include multiple independent sub-groups, the meta-analyst must decide how to synthesize data from the sub-groups. Two main approaches exist—consider subgroups as separate studies within the meta-analysis, or combine the studies into a single effect size (Borenstein et al., 2009). For the former approach, the meta-analyst would consider a study with three independent groups as three separate studies within the meta-analysis, each weighted by its own subgroup sample size. This approach would be warranted in situations where the researcher was interested in the differences between the subgroups, or when the subgroups yielded different outcomes. However, some criticize that this approach weights complex studies more heavily in the overall analysis—a large study with six subgroups has six data points in the final analysis, whereas a small study with no subgroups only has one data point (Borenstein et al., 2009).

The other approach is to combine the subgroups into one effect size for the study. If the subgroups do not yield differing results on the outcome measure, or if the subgroups are not of interest in the research questions, then considering the study as the unit of analysis is appropriate. Furthermore, the subgroups can be coded as potential moderators if the researcher is interested in determining a possible moderating effect. If the study is considered the unit of analysis, a sort of mini meta-analysis is conducted to yield a single effect size for the study. Effect sizes are calculated for each subgroup and then multiplied by the subgroup's weight. The product of the effect size and the

subgroup's weight is then divided by the summed weights of the groups to produce an overall effect size for the study. Calculating each subgroup's effect size is preferred to first pooling the means or other raw data and then calculating a mean effect size. The latter can produce a situation called Simpson's paradox, in which an observed positive trend reverses or disappears when the data are combined (Borenstein et al., 2009). It is important to compare each group's outcome with its own control group to avoid Simpson's paradox (Borenstein et al., 2009).

For this meta-analysis, the study was considered the unit of analysis for independent groups, and data were pooled together for subgroups to yield a single effect size for the study. One such example is the McConney, Oliver, Woods-McConney, Schibeci, and Maor (2014) study. This study examined affective engagement differences between low inquiry and high inquiry classrooms. The researchers reported separate affective engagement measures for three countries: Australia, Canada, and New Zealand. A separate effect size was calculated for each country and multiplied by the inverse of the variance for each subgroup. The resulting products were averaged to yield a single effect size for the study.

Non-independent subgroups. Some complex data structures involve multiple comparisons, outcome measures or time points. For these structures, the multiple data points do not yield independent information, because they came from the same participants. For non-independent subgroups, conducting a mini meta-analysis is not appropriate because the precision of the summary effect is improperly estimated (Borenstein et al., 2009). To combine data from non-independent groups, a mean effect size is first calculated for the different outcomes. Then the variance is calculated in order

to assign a weight to the mean effect size. The formula for the variance corrects for the correlation among the outcomes for non-independent groups:

$$V_{\bar{Y}} = \left(\frac{1}{m}\right)^2 \left(\sum_{i=1}^m V_i + \sum_{i \neq j} r_{ij} \sqrt{V_i} \sqrt{V_j} \right)$$

where V is the variance, \bar{Y} is the mean of the effect sizes, m is the number of outcomes, and r is the correlation amongst the studies. Thus for a study with two outcomes, and a correlation of .5 between the two outcomes, the formula for the variance would be:

$$V_{\bar{Y}} = \left(\frac{1}{2}\right)^2 (V_1 + V_2 + 0.5\sqrt{V_1}\sqrt{V_2})$$

In many studies in engagement research specifically, and education research generally, the correlations among the outcome variables are not given. When the correlations are not known, the researcher can use a reasonable estimate of r from similar research, use a default value of zero that assumes no correlation, or use a default value of one that assumes perfect correlation (Borenstein et al., 2009). As the outcome measures for engagement vary from study to study, despite being classified as behavioral, affective, or cognitive engagement; it is difficult to find a reasonable estimate of r in the literature. Further, assuming no correlation will overestimate variance and underestimate precision. Conversely, assuming perfect correlation will underestimate variance and overestimate precision (Schmidt & Hunter, 2015). Thus, the investigator analyzed the multiple outcomes to be combined to determine if they were likely to be highly correlated, moderately correlated, or weakly correlated. Following guidelines proposed by Ferguson (2009), the values assigned to those levels of correlations were .8, .5, and .2 respectively (see Table 4 for further correlation assumptions used by the investigator). For example, in the McConney et al. (2014) study, five variables reflecting affective engagement were

assessed. As these measures were from the same facet of engagement, they were assumed to be highly correlated. Thus, .8 was used as the correlation in the variance formula used to combine the five affective engagement measures into a single, aggregate measure of affective engagement. Microsoft Excel was used to calculate pooled variances for non-independent groups.

Table 4

| <i>Correlation Assumptions for Combining Variances of Non-independent Group Measures</i> | |
|--|------------------|
| Outcome measures | Correlation used |
| Behavioral and Behavioral Affective and Affective Cognitive and Cognitive | .8 |
| Behavioral and Affective Behavioral and Cognitive Affective and Cognitive | .5 |

Synthesis of multiple studies.

Statistical model. There are two statistical models for conducting meta-analyses—fixed effects and random effects. The fixed effect model assumes that there is one true effect size for all studies in the analysis. This model attributes differences in observed effects solely to sampling error. Studies that are synthesized using a fixed effect model tend to have similar characteristics; differences in any feature of the study would be likely to introduce real variation in the effect size. Conversely, the random effects model assumes that the true effect varies from study to study depending on participant characteristics, implementation of interventions, and differing methodologies. Thus, in a random effects model, differences in observed effects are attributed not only to sampling error but also to true variation between studies. As engagement effect sizes were expected

to vary based on study characteristics, the investigator used a random effects model for this meta-analysis.

The random effects and fixed effect models differ in how weights are assigned to each study's effect size to compute a summary effect. For both models, the weight for any study's effect size is the inverse of the variance. Thus, studies with more precision (less variance) are weighted more heavily in the meta-analysis; while studies with less precision (higher variance) are weighted less heavily. In a random effects model, the variance is the sum of the within-study variance (sampling error) and the between-studies variance. The between-studies variance within the sample of studies included in the meta-analysis is designated by T^2 :

$$T^2 = \frac{Q - df}{C}$$

where Q is the observed or total variance for the meta-analysis, df is the degrees of freedom for the meta-analysis ($k-1$), and C is a scaling factor that corrects for the fact that Q is a weighted sum of squares. The total observed variance (Q) is calculated using the following formula:

$$Q = \sum_{i=1}^k W_i Y_i^2 - \frac{(\sum_{i=1}^k W_i Y_i)^2}{\sum_{i=1}^k W_i}$$

where the weights are the inverse of the within-studies variance for each study. The formula for the scaling factor (C) is given by the following formula:

$$C = \sum W_i - \frac{\sum W_i^2}{\sum W_i}$$

Summary statistics. A summary or pooled mean effect size and confidence interval was calculated for all of the studies in the meta-analysis. The mean effect size

was calculated by dividing the sum of the product of the weight and effect sizes for the study by the sum of the weights for the studies. A variance was calculated by taking the inverse of the sum of the study weights. Subsequently, a confidence interval was calculated using the formulae described in the section on confidence intervals. Similar summary statistics were calculated for behavioral, affective and cognitive engagement.

A number of statistics were reported to quantify the heterogeneity of values for the included studies. Cochran's Q is a commonly reported statistic that represents the total observed variation for the studies. Q is the weighted sum of squared differences between individual study effects and the pooled study effect (see "Statistical Model" for the formula). Q is a test of statistical significance, and thus, its precision is given using p -values. The degrees of freedom (df) represents the amount of expected within studies variation, and is given by $k-1$, where k is the number of studies. Thus, subtracting the within studies variation (df) from the observed variation (Q) yields the excess variation, which can be attributed to true differences from study to study.

Because Q is a sum, dependent on the number of studies, other summary statistics that quantify dispersion were also reported. Tau-squared (T^2) and tau (T) are statistics which correct Q for the number of studies. T^2 represents the variance of the true effects between studies, and is given in the section on statistical methods. In examining the formula for T^2 , it is clear that if the observed variation (Q) is greater than the expected variation (df), T^2 will be positive. If the observed variation is less than the expected variation, T^2 will be negative. However, since the actual variance of true effects will never be less than zero, T^2 is considered to be zero when the estimate is negative (Borenstein et al., 2009). T^2 is also used to establish weights in a random effects meta-

analysis model (see “Statistical Model”). Taking the square root of T^2 gives T , which is the standard deviation of the pooled effect size. Thus, T can be used to calculate the confidence interval for the summary effect size.

Another statistic that reflects the true variation between studies is I^2 , which is a ratio of excess to total dispersion, given by the following formula:

$$I^2 = \frac{(Q - df)}{Q} \times 100\%$$

Thus, I^2 can be viewed as a measure of signal to noise. Suggested benchmarks for I^2 are that 25% is low, 50% is moderate, and 75% is high (Higgins, Thompson, Deeks, & Altman, 2003). A low value suggests that there is little true difference between studies, while a large value indicates that there is a large difference between studies. The benefit to using I^2 to quantify dispersion is that it is not dependent on sample size (as Q and p are), and it is not dependent on the scale in the study (as T^2 and T are).

Issues of precision and variance. Just as there are considerations about precision and variance in combining multiple groups within individual studies, there are similar considerations when combining studies to produce summary effect sizes for the meta-analysis and outcomes within the meta-analysis. One such issue is how to include a single study with two separate reported outcomes in calculations of the overall effect size. One option is to use the mean of the two outcomes. For example, the Moote, Williams, and Sproule study (2013) has affective and cognitive engagement outcomes. One could combine average the effect sizes for the two outcomes and include this mean in the overall effect size calculation.

However, as one research question in the study concerns predictors of each facet of engagement, combining affective and cognitive outcomes is not advisable. The second

option is to consider each outcome separately in the meta-analysis. While this option gives accurate effect size statistics for each outcome, it assigns more weight in the overall analysis to studies with more outcomes. More concerning, considering each outcome separately assumes that the two outcomes are independent of each other, which underestimates error and overestimates precision, when the two outcomes are positively correlated, which they would be in the case of engagement measures (Borenstein et al., 2009). In considering outcomes separately, the chance of Type I error is inflated: the summary effect size's standard error will be smaller than it should and its confidence interval will be more narrow than it should (Schmidt & Hunter, 2015).

In order to obtain accurate precision and variance estimates for the summary effect size, the investigator averaged effect sizes for multiple outcomes, yielding a single effect size per study. However, in analyzing outcomes for affective, behavioral, and cognitive engagement, each outcome was considered separately. By using both approaches, averaging multiple outcomes for a study to obtain overall effect size estimates, and considering each outcome separately to analyze predictors of each type of engagement, statistically sound conclusions can be drawn for each of the research questions within the study.

Similar precision and variance concerns arise with multiple comparisons in a single study. For example, in the Linnebrink-Garcia, Patall, and Messersmith study (2013), there are five predictors of affective engagement. These can be averaged to yield accurate precision and variance estimates for the summary affective engagement effect size, or they can be considered separately to determine the differences. In this analysis, comparisons were averaged to yield accurate estimates of precision and variance for the

summary effect size and for the effect sizes for each type of engagement. Considering multiple comparisons separately was problematic, as the comparisons differed from study to study. One of the comparisons within the Linnenbrink-Garcia (2013) study was about the relationship of involvement supports with affective engagement; this specific comparison did not appear in other studies. Analysis by comparison was accomplished through hierarchical meta-regression considering author and predictor type as moderators (see “Meta-regression”). Such an approach differentiated effects of comparisons, while also considering those effects nested by author to avoid violations of statistical independence.

Another issue of precision and variance concerning statistical non-independence is when a number of studies from the same author occur within the same meta-analysis. Though the problem is similar in many ways to the multiple outcome and multiple comparison concerns, there are some unique characteristics to the multiple author problem. Papers by the same author may or may not vary in the sample or data chosen for analysis. Thus, it is difficult to discern the degree of statistical dependence from study to study. Though similar solutions to the multiple comparison/outcome problem exist for the multiple author problem, unique solutions include selecting one study from the author with sample size as a criterion for selection, weighting studies by sample size, using the author rather than the study as the unit in meta-analysis, and sensitivity analyses with and without the author in question (Shin, 2009). This meta-analysis considers the author as a covariate at the meta-regression stage, following recommendations from the research literature (Cheung, 2015; Shin, 2009).

Meta-regression. Though summary effects are often desired in meta-analyses, the research questions in this meta-analysis focus on finding the most practically significant predictors of different types of engagement. Additionally, coded study characteristics could be potential moderators of the effects of predictors on engagement. Thus, it is useful to determine what portion, if any, of the between-studies variance is due to these covariates.

The use of multiple regression in meta-analysis to investigate the relationships of these multiple predictors and moderators with an outcome is termed meta-regression. Procedures that are used in multiple regression, such as hierarchical modeling, can be used in meta-regression as well to examine groups of predictors. As Borenstein et al. (2009) suggested, there should be a minimum of ten point estimates per covariate. For categorical covariates, there should be 10 point estimates per category of the covariate. For example, in considering the effect of instrument reliability, there should be 50 point estimates: 10 for each of the collapsed categories for that potential moderator.

For meta-regression, as for meta-analysis, the investigator can select a fixed or random effects model. In meta-regression, a random effects model assumes that only some of the between studies variation can be explained by the covariate. Thus, a random effects model was selected for meta-regression models conducted within this meta-analysis. Reported statistics for included statistical significance tests of the model: Z -value ($p < .05$) for one covariate or Q -value ($p < .05$) as an omnibus test of the model fit for two or more covariates. Additionally, I^2 was reported to describe the between studies variance with the effect of the covariate(s) removed. Thus, I^2 ($p < .05$) was used in a goodness of fit test of the null hypothesis that the variance not explained by the model is

zero. R^2 was reported to reflect the proportion of variance explained by the covariates. Lastly, β -values and their corresponding significance and confidence intervals were also reported for each covariate.

Publication bias analysis. Because of conflicting results about whether or not publication bias exists and the extent of its impact, it was assessed in this meta-analysis (see “Criticism and Limitations of Meta-Analysis”). There are a number of methods available to assess publication bias. Mimicking Rosenthal’s (1984) study, the investigator compared the mean effects of published and unpublished studies to determine possible differences. A “file drawer” analysis was conducted to estimate the number of unpublished studies, the “fail-safe N ,” that would bring the effect size to zero (Orwin, 1983; Schmidt, Pearlman, Hunter, & Shane, 1979). The rationale of such a technique is that if the fail-safe N is an exceedingly large number, there is little reason for concern about publication bias (Borenstein et al., 2009). In practice, this technique produces paradoxical results; a meta-analysis affected by publication bias will inflate the effect size—yet the larger the effect size, the larger the number of unpublished studies needed to change the cumulative effect (Schmidt & Hunter, 2015).

Funnel plots were used as a visual method of assessing publication bias. In a funnel plot, studies are organized by effect size on the x-axis and standard error on the y-axis (Light & Pillemer, 1984). Consequently, studies with larger sample sizes cluster at the top of the graph with a smaller dispersion left to right, while smaller sample sizes spread out across the bottom of the graph with a wider dispersion left to right. Publication bias is indicated by a skew of studies toward the right-hand side (higher effect size) of the funnel plot (see Figure 21 for an example of a funnel plot). In particular, meta-analysts

look for possible skew in smaller studies because effect size dispersion in smaller studies is greater and easier to discern on a funnel plot, and because larger studies tend to get published, positive effects or not, due to money and time investments in those studies (Borenstein et al., 2009).

Because the interpretation of funnel plots is subjective and because there are other explanations for positive skew in effect sizes, the investigator also used a trim and fill procedure to determine an unbiased estimate of the true effect size (Duval & Tweedie, 2000). Using this procedure, studies that make the funnel plot asymmetrical are removed, an adjusted effect size is calculated, and then both the asymmetrical studies and their missing counterparts are added back to the funnel plot. This adjusted effect size can be used as the summary effect and compared with the raw estimate to determine potential differences. It is important to note that the trim and fill procedure corrects for skew, regardless of the reason for the skew. Thus, care should be taken in using this procedure when there is considerable heterogeneity and/or significant covariates in included studies (Terrin, Schmid, Lau, & Olkin, 2003). What appears to be publication bias might be true heterogeneity among studies.

Limitations and delimitations of the study. The exclusive focus on the early adolescent age group means that the results of this meta-analysis cannot be generalized to high school age students. Furthermore, there are valid engagement studies that focus on the high school age group, which could provide a valuable, holistic picture of secondary science engagement. However, the decision to focus on the early adolescent age group is deliberate, as this is the age range which first shows marked declines in engagement

(Braund & Driver, 2005; Eccles et al., 1993; Eccles & Roeser, 2010; Mahatmya et al., 2012).

Characteristics of the student engagement measure imposed further limitations on the study. This meta-analysis excluded student engagement that was measured through external observation. There are valid teacher report and classroom observation protocols available for assessing engagement that could provide useful information about practices that best engage early adolescent students in science (Fredricks et al., 2011). Because teacher and student perceptions of engagement differ, the predictors identified as strong in this study may not appear to strongly predict student engagement from the perspective of teachers.

The inclusion of studies with a wide array of research designs means that causality between potential predictors and engagement outcomes cannot be established. Thus, while it might seem logical to use the results of this meta-analysis to inform educational practices to enhance student engagement, the results would better be used to identify predictors for further quasi-experimental or experimental research.

The interpretability of the research as a whole is complicated not only by the broad inclusion of research designs, but also by construct validity issues. These issues manifest in instrumentation, adoption of varying engagement operationalizations by researchers, and utilization of terms which mean different things to different researchers. Though the investigator examined the instrumentation in studies in regards to their congruence with established conceptualizations, this examination and possible recategorization of studies is itself, based on the interpretation of the investigator. Thus, transparency was considered important in terms of the decision the investigator made in

terms of categorizing study conceptualizations of engagement (see Table 3 for a list of indicators for classifying engagement conceptualization).

Summary

The investigator conducted a random effects meta-analysis of classroom and task-level predictors of middle school students' self reports of engagement in science. Studies with a variety of methodological designs were included in the meta-analysis; including correlational, ex post facto, and quasi-experimental studies. Potential moderators were coded, including publication status, predictor type, engagement conceptualization, school age range, school location, instrument reliability, geographic location, and methodological design. The impact of moderators was investigated through meta-regression when there were at least 10 studies per moderator. Though the investigator performed extensive searches to obtain published and unpublished studies, publication bias was assessed using visual analysis of funnel plots, comparison of means between published and unpublished studies, calculation of Orwin's fail-safe N , and use of a trim and fill procedure. In the remaining chapters, the investigator presents the results of the main meta-analysis and meta-regression, and analyzes the results in terms of theory, existing research, and the research hypotheses and questions.

Chapter 4: Results

Descriptive Statistics

The investigator found 75 studies that met the inclusion criteria. Due to personal communications with study authors, the investigator determined that an additional four studies met inclusion criteria, resulting in 79 total studies for the analysis (M. E. Bathgate, personal communication, 2016; S. Blanchard, personal communication, 2016; B. J. Fraser, personal communication, 2016; & J. Osborne, personal communication, 2016). See Table C1, *Overview of Included Studies* (Appendix C).

Source characteristics. Descriptive statistics for source characteristics include publication and peer-reviewed status, as well as sample size. The majority of the included studies were published ($k = 58$, 73.4%) and peer-reviewed ($k = 52$, 67.6%). Sample sizes for the studies ranged from 20 to 10,437, with an overall sample size of 53,971 for the meta-analysis. The majority of studies ($k = 48$) had sample sizes larger than 100. Detailed information about source characteristics can be found in Figure 1, *Stem and leaf plot of sample sizes*, and Table D1, *Descriptive Statistics for Included Studies* (Appendix D).

Study characteristics. Included studies reflected a range of methodologies and instrument characteristics. The most common study design was quasi-experimental ($k = 31$), followed by correlational ($k = 23$), single-group, repeated measures ($k = 18$), and experimental ($k = 7$). The majority of studies utilized psychometric instruments with high internal reliability, with Cronbach alphas of greater than .70 ($k = 46$). While some studies ($k = 16$) reported internal validity measures by way of exploratory or confirmatory factor analyses or concurrent validity, many studies ($k = 45$) referenced an external instrument

as a way of establishing validity. Detailed information about reliability and validity can be found in Table D1, *Descriptive Statistics for Included Studies* (Appendix D).

| Ste | Leaf | | | | | | | | | | | | | | | | | | | | | | | | |
|-----|------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2 | 3 | 3 | 3 | 4 | 4 | 4 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 6 | 6 | 6 | 7 | 7 | 7 | 8 | 8 | 8 | 9 |
| 100 | 0 | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 6 | 7 | 7 | 7 | 7 | | | | | | | | | | | | |
| 200 | 0 | 2 | 2 | 3 | 4 | 5 | 5 | 6 | 7 | 9 | | | | | | | | | | | | | | | |
| 300 | 1 | 6 | 7 | | | | | | | | | | | | | | | | | | | | | | |
| 400 | 4 | | | | | | | | | | | | | | | | | | | | | | | | |
| 500 | 3 | | | | | | | | | | | | | | | | | | | | | | | | |
| 600 | 1 | 6 | | | | | | | | | | | | | | | | | | | | | | | |
| 700 | 3 | 5 | | | | | | | | | | | | | | | | | | | | | | | |
| 800 | 0 | 6 | | | | | | | | | | | | | | | | | | | | | | | |
| 900 | 0 | 0 | 8 | | | | | | | | | | | | | | | | | | | | | | |
| 100 | 2 | | | | | | | | | | | | | | | | | | | | | | | | |
| 110 | 5 | 8 | | | | | | | | | | | | | | | | | | | | | | | |
| 120 | 3 | | | | | | | | | | | | | | | | | | | | | | | | |
| 130 | 8 | | | | | | | | | | | | | | | | | | | | | | | | |
| 140 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 150 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 160 | 1 | 4 | | | | | | | | | | | | | | | | | | | | | | | |
| 170 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 180 | 1 | 9 | | | | | | | | | | | | | | | | | | | | | | | |
| 190 | 3 | 6 | | | | | | | | | | | | | | | | | | | | | | | |

Figure 1. Stem and leaf plot of sample sizes from 76 studies. Outliers omitted ($n = 3,281$, $n = 8,544$, $n = 10,437$)

A range of school characteristics were reflected by the included studies. Studies investigating engagement in public schools were most common ($k = 46$) and private schools were least common ($k = 4$). Twenty-one studies did not report school type. Common school structures included middle schools ($k = 21$) and K-8 schools ($k = 16$), though 26 studies did not report school structure. School settings included urban ($k = 13$), followed by suburban ($k = 8$) rural settings ($k = 5$). Eight studies reflected a mix of school settings, while 45 studies did not report school setting. For studies reporting socio-economic status, high status (<35% FRL) was most common ($k = 13$), followed by low status ($k = 6$), and average status (35-59% FRL, $k = 3$). Geographically, the majority of studies took place in schools outside of the United States ($k = 44$), with high representation in Turkey ($k = 18$), Taiwan ($k = 4$), and Israel ($k = 3$). Detailed information

about school characteristics can be found in Table B1, *Statistics and Moderators by Point Estimate* (Appendix B), Table C1, *Overview of Included Studies* (Appendix C), and Table D1, *Descriptive Statistics for Included Studies* (Appendix D).

Sixteen of the 79 studies yielded multiple predictors of engagement. Predictors were coded both by type and by self-determination theory component (see Table 5). Instructional method ($n = 57, k = 40$) and class characteristics ($n = 60, k = 20$) were the most common predictor types. Predictors representing aspects of autonomy were most common ($n = 94, k = 22$), followed by relatedness ($n = 35, k = 49$) and competence ($n = 29, k = 21$). The number of studies sums to more than 79 for predictor type ($k = 83$) and self-determination theory classification ($k = 82$) as some studies included more than one engagement predictor. Detailed information about predictor classifications can be found in Table B1, *Statistics and Moderators by Point Estimate* (Appendix B), Table D1, *Descriptive Statistics for Included Studies* (Appendix D), and Table E1, *Selection and Use of Predictor and Criterion Variables by Study* (Appendix E).

Table 5

Descriptive Statistics for Predictor Classification

| Predictor classification | Point estimates | | Studies | |
|---------------------------|-----------------|---------|---------|---------|
| | n | Percent | k | Percent |
| Type | | | | |
| Instructional Method | 57 | 36.1% | 40 | 48.2% |
| Technology | 15 | 9.5% | 13 | 15.7% |
| Class Characteristics | 60 | 37.9% | 20 | 24.1% |
| Social Characteristics | 26 | 16.5% | 10 | 12% |
| Self-determination theory | | | | |
| Autonomy | 94 | 59.4% | 22 | 23.9% |
| Competence | 29 | 18.4% | 21 | 22.8% |
| Relatedness | 35 | 22.2% | 49 | 53.3% |

Twenty-three of the 79 studies yielded multiple engagement outcomes. The most common outcome provided by the studies was affective engagement ($n = 84, k = 56$), followed by cognitive engagement ($n = 49, k = 31$), combinations of two engagement outcomes ($n = 13, k = 9$), behavioral engagement ($n = 10, k = 7$), and combinations of all three engagement outcomes ($n = 2, k = 2$). The number of studies summed to more than 79 ($k = 105$) because some studies provided data about more than one engagement outcome. Detailed information about engagement outcomes can be found in Table 6, *Descriptive Statistics for Engagement Outcomes*, Table C1, *Overview of Included Studies* (Appendix C), Table D1, *Descriptive Statistics for Included Studies* (Appendix D), and Table E1, *Selection and Use of Predictor and Criterion Variables by Study* (Appendix E).

Table 6

Descriptive Statistics for Engagement Outcomes

| Engagement type | Point estimates | | Studies | |
|-------------------------|-----------------|---------|----------|---------|
| | <i>n</i> | Percent | <i>k</i> | Percent |
| Behavioral | 10 | 6.3% | 7 | 6.7% |
| Affective | 84 | 53.2% | 56 | 53.3% |
| Cognitive | 49 | 31% | 31 | 29.5% |
| Two outcomes combined | 13 | 8.2% | 9 | 8.7% |
| Three outcomes combined | 2 | 1.3% | 2 | 1.9% |

Summary Effect Size

Though this meta-analysis focused primarily on dispersion of effect sizes, a summary effect size was calculated. When combining across predictors and outcomes to ensure precision of the point estimate, the summary mean effect size was $g = .37$, 95% CI [30, .43]. Statistical significance tests confirm that the mean effect of predictors in these studies was likely not zero ($Z = 10.98, p < .0001$). Tests of heterogeneity confirmed that the random effects model was appropriate, as there was variation in the effect size among studies that was likely not due to sampling error ($Q = 1884, p < .0001$). The standard

deviation of the expected true predictor effects ($T = .26$) suggested that 95% of engagement predictors should be distributed $\pm .51$ around the mean effect size ($g = .37$, 95% CI $[-.14, .87]$). Finally, 95.86% of the observed variance is attributable to authentic differences in engagement predictors ($I^2 = 95.86$). The high I^2 value suggests a closer examination of engagement predictors through meta-regression.

| Stem | Leaf |
|------|---|
| -.7 | 5 |
| -.6 | |
| -.5 | 7 8 |
| -.4 | 1 2 2 8 |
| -.3 | 0 3 |
| -.2 | 0 2 2 6 8 |
| -.1 | 0 0 1 2 2 4 4 6 |
| -0.0 | 0 0 1 2 2 3 3 4 4 6 7 8 |
| 0.0 | 0 2 3 3 3 4 4 4 5 7 8 8 8 9 9 9 |
| .1 | 0 0 0 2 2 3 3 3 3 3 4 4 4 5 6 6 8 8 8 9 9 9 |
| .2 | 0 1 3 3 4 6 7 7 7 8 8 |
| .3 | 0 0 0 1 2 2 2 3 3 3 4 5 5 6 6 6 6 7 7 9 9 |
| .4 | 1 1 1 1 3 3 5 6 7 7 |
| .5 | 1 2 3 5 5 6 6 7 8 |
| .6 | 0 1 3 7 8 |
| .7 | 1 2 3 4 4 4 6 7 7 7 |
| .8 | 2 5 8 |
| .9 | 0 2 |
| 1.0 | 0 1 2 |
| 1.1 | 7 7 |
| 1.2 | 5 |
| 1.3 | 5 7 |
| 1.4 | |
| 1.5 | 0 4 |
| 1.6 | 7 |
| 1.7 | |
| 1.8 | 0 9 |
| 1.9 | |
| 2.0 | |
| 2.1 | |
| 2.2 | |
| 2.3 | |
| 2.4 | 5 |
| 2.5 | 1 |

Figure 2. Stem and leaf plot of 158 point estimates from 79 studies.

Effect Sizes for Individual Studies

The investigator calculated 158 effect sizes representing each engagement predictor and outcome for the 79 included studies. The effect sizes ranged from $-.75$ to

2.51, with the majority falling between $-.75$ and 1.8 (see Figure 2). Positive effect sizes were most numerous ($n = 124$), though there were 33 negative effect sizes, and one effect size of zero.

Research Questions

Research Question 1: Moderators of Engagement. In order to answer the first research question: *what moderators have statistically significant practical effects on early adolescents' science engagement as assessed by student self-report?* the investigator conducted a meta-regression for coded moderators with a minimum of 10 point estimates per category. See Table C1 for a list of coded moderators and number of point estimates per category and Table E1 for a detailed list of coded moderators by study. For predictors with fewer than 10 point estimates per category, the investigator either conducted a meta-regression with collapsed categories (e.g., combining all of engagement categories reflecting two or more combined engagement types) to achieve the suggested minimum ten point estimates, or simply provided descriptive statistics (e.g., for socioeconomic status). The rationale for each decision is included within the sections for each meta-regression.

Publication status. Point estimates from published studies showed the higher effect size ($g = .40$, 95% CI [.33, .46]), while those from unpublished studies showed the lower effect size ($g = .15$, 95% CI [.04, .25]). The Z-values indicate it was unlikely that the effect size for either publication status was zero (see Table 7).

The descriptive statistics for publication status included a minimum of 10 point estimates per category, indicating that meta-regression was appropriate. A test of the publication status regression model reveals that it was likely effect size differed by

publication status ($Q = 15.70, p = .00007$). However, the model was incomplete, as there was unexplained variance between point estimates with the same publication status ($Q = 4772, p < .0001$). The incremental changes in unexplained variance (T^2) for publication status are presented in Table 8. The proportion of the unexplained variance that represented true variance, rather than error variance, was 96.73% ($I^2 = 96.73, I = .98$). This suggests that the observed variance around subgroup means would shrink by approximately 2% if the error variance were removed. The predictor type model explained a negligible amount of the total between-studies variance in effect sizes ($R^2 < .0001$).

Table 7

Effect Sizes and Null Tests for Publication Status

| Publication status | <i>n</i> | <i>g</i> | 95% CI | <i>Z</i> | <i>p</i> |
|--------------------|----------|----------|------------|----------|----------|
| No | 39 | .15 | [.04, .25] | 19.75 | .0000 |
| Yes | 119 | .40 | [.33, .46] | 39.44 | .0000 |

Table 8

Meta-regression Model for Publication Status

| Publication Status | Variance | | Test of model | | | Regression | | |
|--------------------|----------|-------|---------------|-----------|----------|------------|----------|----------|
| | T^2 | R^2 | Q | <i>df</i> | <i>p</i> | Coeff. | <i>Z</i> | <i>p</i> |
| No (intercept) | .09 | 0 | | | | .15 | 2.70 | .007 |
| Yes | .09 | 0 | 15.70 | 1 | .001 | .25 | 3.96 | .00007 |

An examination of the regression coefficients for the publication status model showed that published studies predicted increases in engagement point estimates ($\beta = .25, p = .00007$) when compared to unpublished studies ($\beta = .15, p = .007$). These statistically significant regression coefficients paralleled the statistically significant null tests for the

effect sizes for published and unpublished studies. Figure 3 shows a scatterplot of the regression model for publication status.

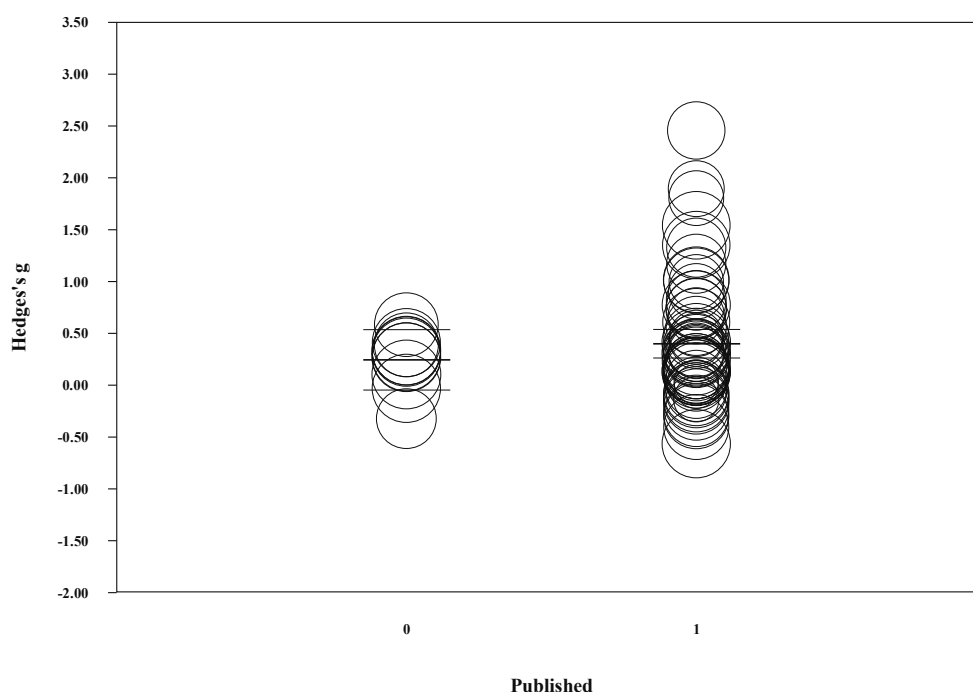


Figure 3. Regression of point estimates on publication status. The values on the x-axis show not unpublished (0) and published (1) codes.

Peer review status. Point estimates from peer reviewed studies showed the higher effect size ($g = .36$, 95% CI [.30, .42]), while those from unpublished studies showed the lower effect size ($g = .27$, 95% CI [.17, .37]). The Z -values indicate it was unlikely that the effect size for either publication status was zero (see Table 9).

The descriptive statistics for peer review status included a minimum of 10 point estimates per category, indicating that meta-regression was appropriate. A test of the peer review status regression model revealed that it was unlikely effect size differed by peer review status ($Q = 2.38$, $p = .123$). The model was incomplete, as there was unexplained variance between point estimates with the same peer review status ($Q = 4819$, $p < .0001$). The incremental changes in unexplained variance (T^2) for publication status are presented

in Table 10. The proportion of the unexplained variance that represented true variance, rather than error variance, was 96.76% ($I^2 = 96.76, I = .98$). This suggests that the observed variance around subgroup means would shrink by approximately 2% if the error variance were removed. The predictor type model explained a negligible amount of the total between-studies variance in effect sizes ($R^2 < .0001$).

Table 9

Effect Sizes and Null Tests for Peer Review Status

| Peer review status | <i>n</i> | <i>g</i> | 95% CI | <i>Z</i> | <i>p</i> |
|--------------------|----------|----------|------------|----------|----------|
| No | 47 | .27 | [.17, .37] | 5.38 | .0000 |
| Yes | 111 | .36 | [.30, .42] | 11.04 | .0000 |

Table 10

Meta-regression Model for Peer Review Status

| Peer review status | Variance | | Test of model | | | Regression | | |
|--------------------|----------|-------|---------------|-----------|----------|------------|----------|----------|
| | T^2 | R^2 | Q | <i>df</i> | <i>p</i> | Coeff. | <i>Z</i> | <i>p</i> |
| No (intercept) | .09 | 0 | | | | .27 | 5.38 | .0000 |
| Yes | .10 | 0 | 2.38 | 1 | .123 | .09 | 1.54 | .1230 |

An examination of the regression coefficients for the peer review status model suggests that peer reviewed studies predicted increases in engagement point estimates ($\beta = .09, p = .123$) when compared to studies that were not peer reviewed ($\beta = .27, p < .0001$). Though the null test for the point estimate for peer reviewed studies suggested that the effect of peer on engagement was not zero, the contribution of peer-review to predicting engagement is not significant. Figure 4 shows a scatterplot of the regression model for peer review status.

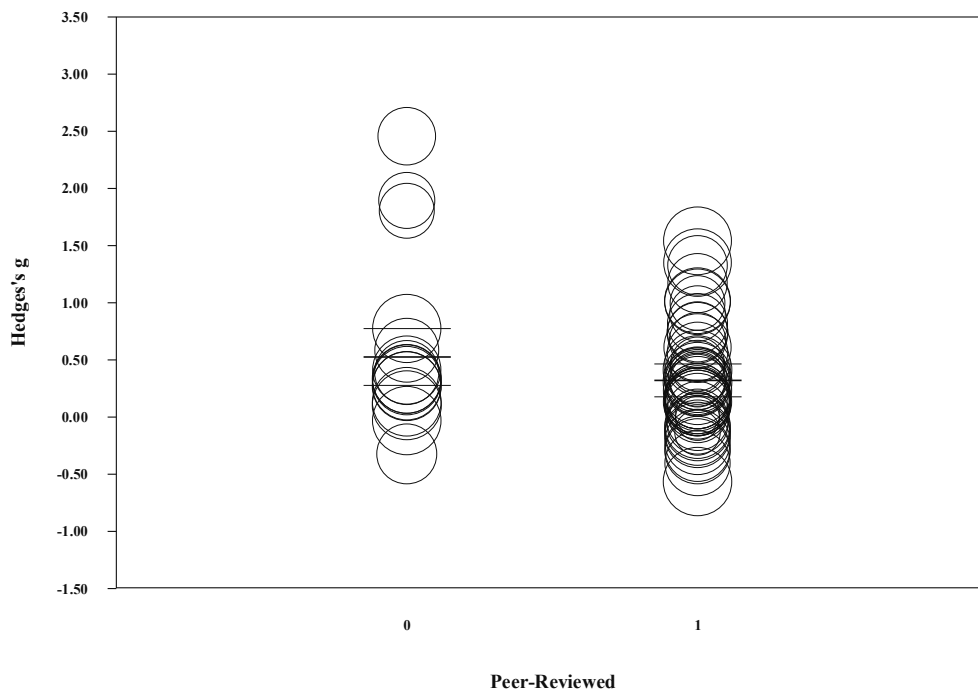


Figure 4. Regression of point estimates on peer review status. The values on the x-axis show not peer reviewed (0) and peer reviewed (1) codes.

School structure. Point estimates from studies sampling high school structures showed the highest effect size ($g = .97$, 95% CI [.59, 1.36]), followed by junior high school structures ($g = .65$, 95% CI [.39, .90]), and K-8 school structures ($g = .42$, 95% CI [.31, .52]). The Z -values suggest that the mean point estimates for each of these categories was statistically significant (see Table 11). The lowest effect sizes came from studies sampling middle schools ($g = .16$, 95% CI [.06, .25]).

The descriptive statistics for school structure showed that there were fewer than 10 point estimates per category, indicating that meta-regression was not appropriate. Despite differences in effect sizes, the investigator decided not to eliminate or condense categories for meta-regression. The rationale for this decision was based upon three main data features. First, there were a limited number of point estimates elementary elementary ($n = 2$), junior high ($n = 7$), and high school categories ($n = 3$). Second, all

K-8 point estimates also represented studies that sampled from Turkey, thus, it would not be possible to separate confounding variables to determine any possible relationship.

Third, logical combinations of school structures (e.g., middle school with junior high school) would combine seemingly disparate mean effect size point estimates.

Table 11

Effect Sizes and Null Tests for School Structure

| School structure | <i>n</i> | <i>g</i> | 95% CI | <i>Z</i> | <i>p</i> |
|------------------|----------|----------|-------------|----------|----------|
| Not specified | 55 | .36 | [.27, .46] | 7.44 | .0000 |
| Elementary | 2 | .17 | [-.28, .62] | .73 | .467 |
| Middle school | 44 | .16 | [.06, .25] | 3.05 | .002 |
| Junior high | 7 | .65 | [.39, .90] | 4.96 | .000 |
| K-8 | 37 | .42 | [.31, .52] | 7.62 | .000 |
| High school | 3 | .97 | [.59, 1.36] | 4.95 | .000 |
| Mix | 10 | .29 | [.09, .50] | 2.79 | .005 |

School type. Studies sampled from unspecified school types showed the highest effect size ($g = .41$, 95% CI [.30, .52]), followed by a mix of school types ($g = .36$, 95% CI [.16, .55]). Point estimates from studies sampling private schools showed the smallest effect size ($g = -.02$, 95% CI [-.31, .27]), though the 95% confidence interval spanned zero and the mean effect was not significant ($Z = -.14$, $p = .888$).

The descriptive statistics for school type showed that there were fewer than ten point estimates per category, indicating that meta-regression was not appropriate. The investigator decided not to eliminate or condense categories for meta-regression, as the category that could provide a useful comparison (private schools) was the smallest one and yielded a nonsignificant effect size point estimate (see Table 12).

Table 12

Effect Sizes and Null Tests for School Type

| School type | <i>n</i> | <i>g</i> | 95% CI | <i>Z</i> | <i>p</i> |
|---------------|----------|----------|-------------|----------|----------|
| Not specified | 43 | .41 | [.30, .52] | 7.36 | .0000 |
| Public | 97 | .32 | [.25, .39] | 9.11 | .0000 |
| Private | 6 | -.02 | [-.31, .27] | -.14 | .888 |
| Mix | 12 | .36 | [.16, .55] | 3.63 | .0000 |

School setting. Point estimates from studies sampling from urban schools ($g = .40$, 95% CI [.25, .54]), and unspecified schools ($g = .40$, 95% CI [.32, .47]), showed the highest effect size while sampling from rural schools showed the lowest effect size ($g = -.11$, 95% CI [-.42, .21]). The *Z*-values indicate it was unlikely that the mean effect sizes for suburban, urban, mixed and unspecified schools was zero. However, it was possible that the mean effect size point estimate for rural schools was zero ($Z = -.66$, $p = .511$) (see Table 13).

Table 13

Effect Sizes and Null Tests for School Setting

| School setting | <i>n</i> | <i>g</i> | 95% CI | <i>Z</i> | <i>p</i> |
|----------------|----------|----------|-------------|----------|----------|
| Not specified | 83 | .40 | [.32, .47] | 10.01 | .000 |
| Rural | 5 | -.11 | [-.42, .21] | -.66 | .511 |
| Suburban | 28 | .20 | [.07, .33] | 2.98 | .003 |
| Urban | 24 | .40 | [.25, .54] | 5.43 | .000 |
| Mix | 18 | .29 | [.11, .46] | 3.17 | .002 |

The descriptive statistics for school setting included a minimum of ten point estimates for all but one category, indicating that meta-regression was not appropriate. Despite having a small number of estimates for rural schools ($n = 5$), the investigator continued with a meta-regression. As this small category yielded a statistically nonsignificant result, the regression was run both with and without the rural subcategory.

The regression model of school setting without the urban category is presented in Table 14. A test of this regression model revealed that it was unlikely effect size differed by school setting ($Q = 7.17, p = .067$). The model was incomplete, as there was unexplained variance between point estimates with the same publication status ($Q = 4905, p < .0001$). The incremental changes in unexplained variance (T^2) for publication status are presented in Table 14. The proportion of the unexplained variance that represented true variance, rather than error variance, was 96.96% ($I^2 = 96.96, I = .98$). This suggests that the observed variance around subgroup means would shrink by approximately 2% if the error variance were removed. This school setting model explained a negligible amount of the total between-studies variance in effect sizes ($R^2 < .0001$).

Table 14

Meta-regression Model for School Setting (Without Rural)

| School Setting | Variance | | Test of model | | | Regression | | |
|-------------------------|----------|-------|---------------|------|------|------------|-------|-------|
| | T^2 | R^2 | Q | df | p | Coeff. | Z | p |
| Unspecified (Intercept) | .0897 | 0 | | | | .3968 | 10.00 | .0000 |
| Suburban | .0902 | 0 | 6.49 | 1 | .011 | -.1955 | -2.49 | .0126 |
| Urban | .1006 | 0 | 6.14 | 2 | .046 | .0011 | .013 | .9897 |
| Mix | .1079 | 0 | 7.17 | 3 | .067 | -.1118 | -1.14 | .256 |

An examination of the regression coefficients for this school setting model showed that studies sampling from suburban ($\beta = -.20, p = .013$) and mixed schools ($\beta = -.11, p = .256$) predicted decreases in engagement point estimates when compared to studies sampling from unspecified school settings ($\beta = .40, p = .000$). However, only the effects of suburban school settings were statistically significant. Figure 5 shows a scatterplot of the regression model for school setting without rural schools.

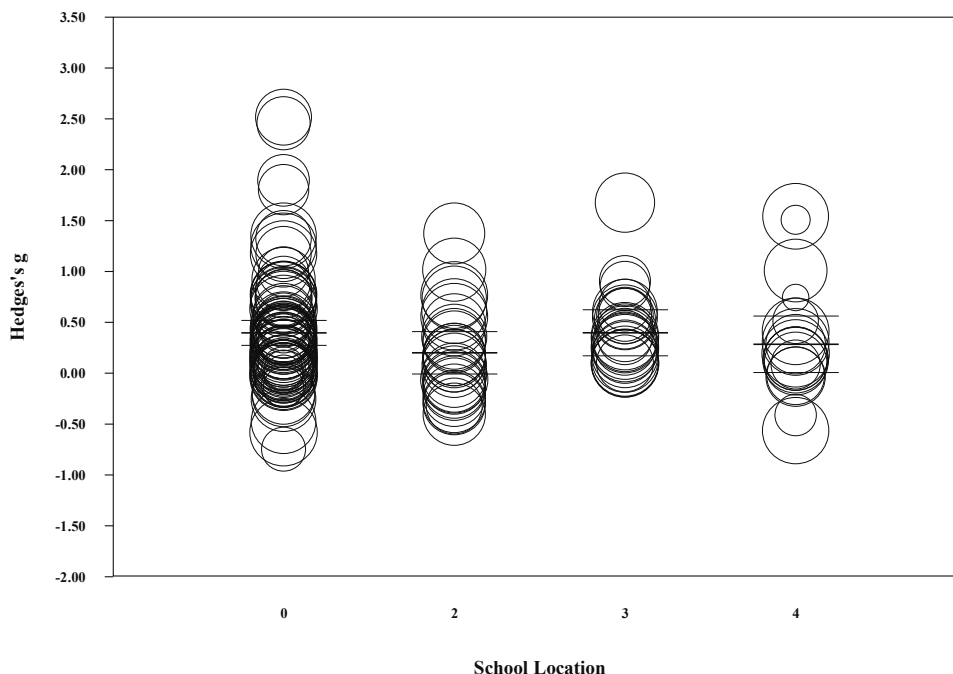


Figure 5. Regression of point estimates on school location (without rural). The values on the x-axis show not unspecified (0), suburban (2), urban (3), and mixed (4) codes.

The regression model of school setting including the rural category is presented in Table 15. A test of this regression model revealed that it was likely effect size differed by school setting ($Q = 14.81, p = .005$). However, this model was incomplete, as there was unexplained variance between point estimates with the same school setting ($Q = 4919, p < .0001$). The proportion of the unexplained variance that represented true variance, rather than error variance, was 96.96% ($I^2 = 96.39, I = .98$). This suggested that the observed variance around subgroup means would shrink by approximately 2% if the error variance were removed. This school setting model explained a negligible amount of the total between-studies variance in effect sizes ($R^2 < .0001$).

An examination of the regression coefficients for this school setting model showed that studies sampling from rural ($\beta = -.50, p = .003$), suburban ($\beta = -.20, p = .013$), and mixed ($\beta = -.11, p = .255$) schools predicted decreases in engagement point

estimates when compared to studies sampling from unspecified school settings ($\beta = .40$, $p = .000$). However, only the effects of rural and suburban settings were statistically significant. Thus, while it was possible that the mean effect size point estimate for rural schools was zero ($Z = -.66$, $p = .511$), the rural school setting was a statistically significant predictor of changes in engagement. Figure 6 shows a scatterplot of the regression model for school setting with rural schools.

Table 15

Meta-regression Model for School Setting (With Rural)

| School setting | Variance | | Test of model | | | Regression | | |
|-------------------------|----------|-------|---------------|------|-------|------------|-------|-------|
| | T^2 | R^2 | Q | df | p | Coeff. | Z | p |
| Unspecified (intercept) | .0898 | 0 | | | | .40 | 10.01 | .0000 |
| Rural | .0894 | .004 | 8.81 | 1 | .0030 | -.50 | -3.02 | .003 |
| Suburban | .0899 | 0 | 15.28 | 2 | .0005 | -.20 | -2.50 | .013 |
| Urban | .1002 | 0 | 14.21 | 3 | .0026 | .001 | .0125 | .990 |
| Mix | .1075 | 0 | 14.81 | 4 | .0051 | -.11 | -1.13 | .255 |

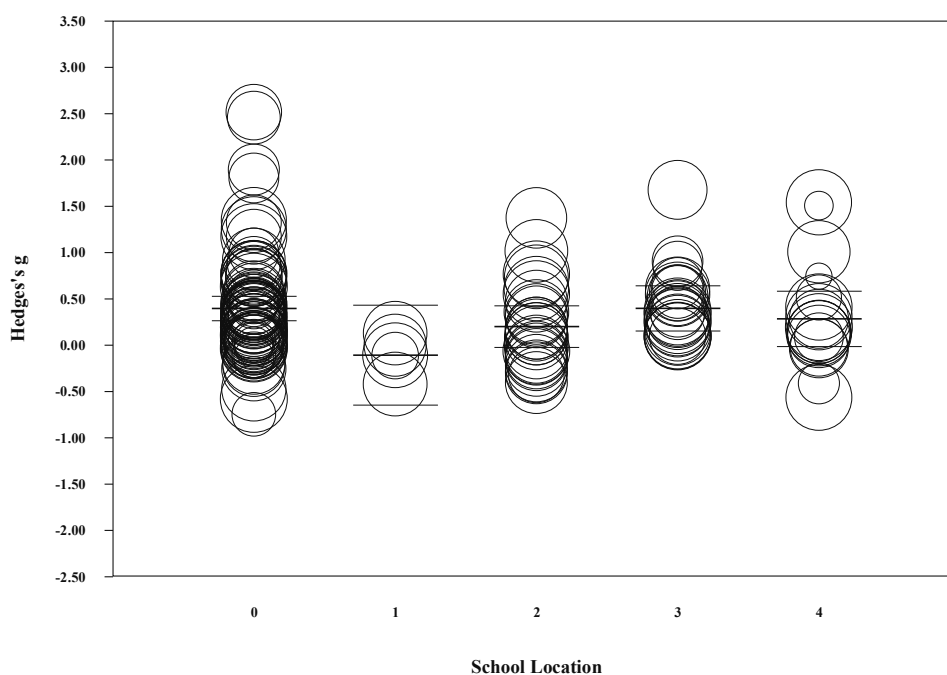


Figure 6. Regression of point estimates on school setting. The values on the x-axis show unspecified (0), rural (1), suburban (2), urban (3), and mixed (4) codes.

Socioeconomic status. Point estimates from studies sampling from average ($g = .49$, 95% CI [.21, .77]), and unspecified ($g = .41$, 95% CI [.34, .49]) SES showed the highest effect size while sampling from high SES showed the lowest effect size ($g = .02$, 95% CI [-.15, .19]). However, the Z -values indicate it was likely that the mean effect size for high SES point estimates was zero (See Table 16).

Table 16

Effect Sizes and Null Tests for Socioeconomic Status

| Socioeconomic status | n | g | 95% CI | Z | p |
|----------------------|-----|-----|--------------|-------|------|
| Not specified | 101 | .41 | [.34, .49] | 10.64 | .000 |
| Low (> 60% FRL) | 9 | .25 | [-.001, .50] | 1.95 | .051 |
| Average (35-59% FRL) | 7 | .49 | [.21, .77] | 3.48 | .001 |
| High (< 35% FRL) | 19 | .02 | [-.15, .19] | .24 | .812 |
| Mix | 22 | .25 | [.08, .42] | 2.85 | .004 |

The descriptive statistics for socioeconomic status included fewer than ten point estimates, indicating that meta-regression was not appropriate. The investigator elected not to conduct a meta-regression as two of the three categories representing specified socioeconomic statuses presented fewer than ten point estimates, and one was not significant.

Geographic location. Point estimates from studies sampling from countries outside the U.S. ($g = .42$, 95% CI [.04, .49]) showed the higher effect size while sampling U.S. schools showed the lower effect size ($g = .24$, 95% CI [.16, .31]). The Z -values indicate it was unlikely that the mean effect sizes for either category was zero. (see Table 17).

The descriptive statistics for geographic location included a minimum of 10 point estimates for both categories, indicating that meta-regression was appropriate. A test of the regression model revealed that it was likely effect size differed by geographic

location ($Q = 11.28, p = .0008$). However, the model was incomplete, as there was unexplained variance between point estimates with the same geographic location ($Q = 4809, p < .0001$). The incremental changes in unexplained variance (T^2) for geographic location are presented in Table 18. The proportion of the unexplained variance that represented true variance, rather than error variance, was 96.76% ($I^2 = 96.76, I = .98$). This suggests that the observed variance around subgroup means would shrink by approximately 2% if error variance were removed. This model explained a negligible amount of the total between-studies variance in effect sizes ($R^2 < .0001$).

Table 17

Effect Sizes and Null Tests for Geographic Location

| | <i>n</i> | <i>g</i> | 95% CI | <i>Z</i> | <i>p</i> |
|------------------|----------|----------|------------|----------|----------|
| United States | 72 | .24 | [.16, .31] | 5.95 | .000 |
| Outside the U.S. | 86 | .42 | [.04, .49] | 11.16 | .000 |

Table 18

Meta-regression Model for Geographic Location

| Geographic location | Variance | | Test of model | | | Regression | | |
|---------------------------|----------|-------|---------------|-----------|----------|------------|----------|----------|
| | T^2 | R^2 | Q | <i>df</i> | <i>p</i> | Coeff. | <i>Z</i> | <i>p</i> |
| United States (intercept) | .0898 | 0 | | | | .24 | 5.95 | .0000 |
| Outside the U.S. | .0944 | 0 | 11.28 | 1 | .0007 | .18 | 3.36 | .0008 |

An examination of regression coefficients for the geographic location model showed that studies sampling schools outside the U.S. predicted increases in engagement point estimates ($\beta = .18, p = .0008$) when compared to studies sampled from schools within the United States ($\beta = .24, p = .000$). Figure 7 shows a scatterplot of the regression model for geographic location.

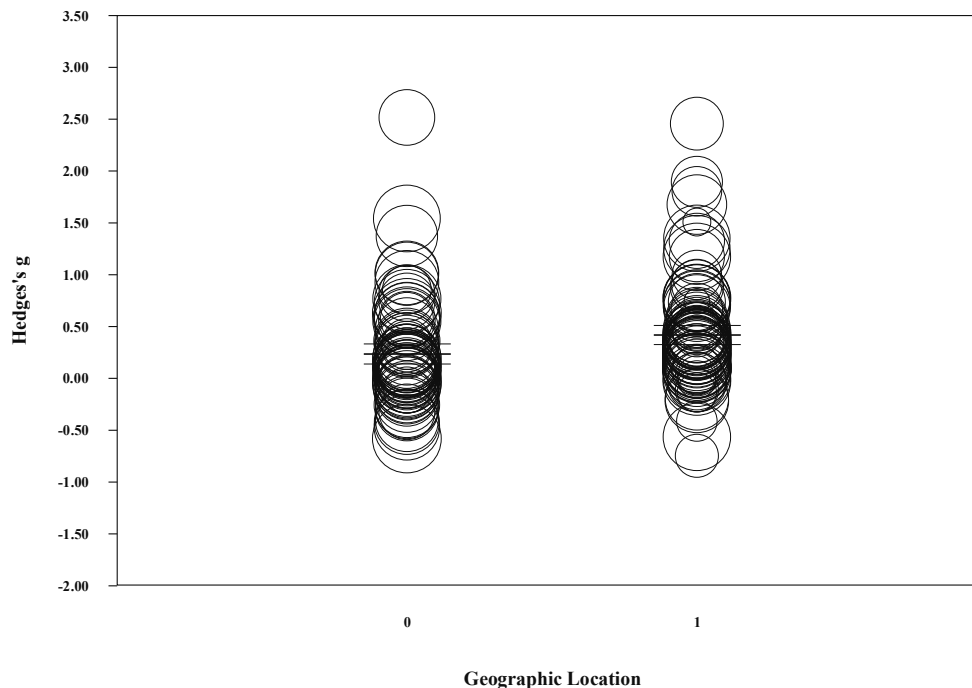


Figure 7. Regression of point estimates on geographic location. The values on the x-axis show U.S. (0) and outside the U.S. (1).

Study methodology. Point estimates from studies with quasi-experimental designs ($g = .42$, 95% CI [.32, .53]) showed the highest effect size while point estimates from studies with experimental designs showed the lowest effect size ($g = .17$, 95% CI [- .05, .39]). While the Z -values for correlational, single group with repeated measures, and quasi-experimental designs indicated it was unlikely that the mean effect was zero, the converse was true for experimental designs ($Z = 1.53$, $p = .127$) (see Table 19).

Table 19

| <i>Effect Sizes and Null Tests for Study Methodology</i> | | | | | |
|--|-----|-----|-------------|------|------|
| Study methodology | n | g | 95% CI | Z | p |
| Correlational | 77 | .32 | [.25, .40] | 8.47 | .000 |
| Single group (repeated measures) | 21 | .25 | [.11, .40] | 3.52 | .000 |
| Quasi-experimental | 45 | .42 | [.32, .53] | 8.14 | .000 |
| Experimental | 15 | .17 | [-.05, .39] | 1.53 | .127 |

The descriptive statistics for study methodology included a minimum of ten point estimates for both categories, indicating that meta-regression was appropriate. A test of the regression model revealed that it was unlikely effect size differed by study methodology ($Q = 6.41, p = .09$). The model was incomplete, as there was unexplained variance between point estimates with the same study methodology ($Q = 4607, p < .0001$). The incremental changes in unexplained variance (I^2) for study methodology are presented in Table 20. The proportion of the unexplained variance that represented true variance, rather than error variance, was 96.66% ($I^2 = 96.66, I = .9832$). This suggests that the observed variance around subgroup means would shrink by approximately 2% if the error variance were removed. The study methodology model explained a negligible amount of the total between-studies variance in effect sizes ($R^2 < .0001$).

Table 20

Meta-regression Model for Study Methodology

| Study methodology | Variance | | Test of model | | | Regression | | |
|----------------------------------|----------|-------|---------------|------|------|------------|-------|-------|
| | I^2 | R^2 | Q | df | p | Coeff. | Z | p |
| Correlational (intercept) | .0898 | 0 | | | | .32 | 8.47 | .0000 |
| Single group (repeated measures) | .0939 | 0 | 1.43 | 1 | .233 | -.07 | -.86 | .390 |
| Quasi-experimental | .0965 | 0 | 4.71 | 2 | .095 | .10 | 1.54 | .123 |
| Experimental | .0963 | 0 | 6.41 | 3 | .093 | -.15 | -1.30 | .193 |

An examination of the regression coefficients for the study methodology model showed that point estimates from studies with quasi-experimental designs ($\beta = .10, p = .123$) predicted increases in engagement point estimates when compared to point estimates from studies with correlational designs ($\beta = .32, p < .0001$). However, none of

predicted changes due to experimental design were statistically significant. Figure 8 shows a scatterplot of the regression model for study methodology.

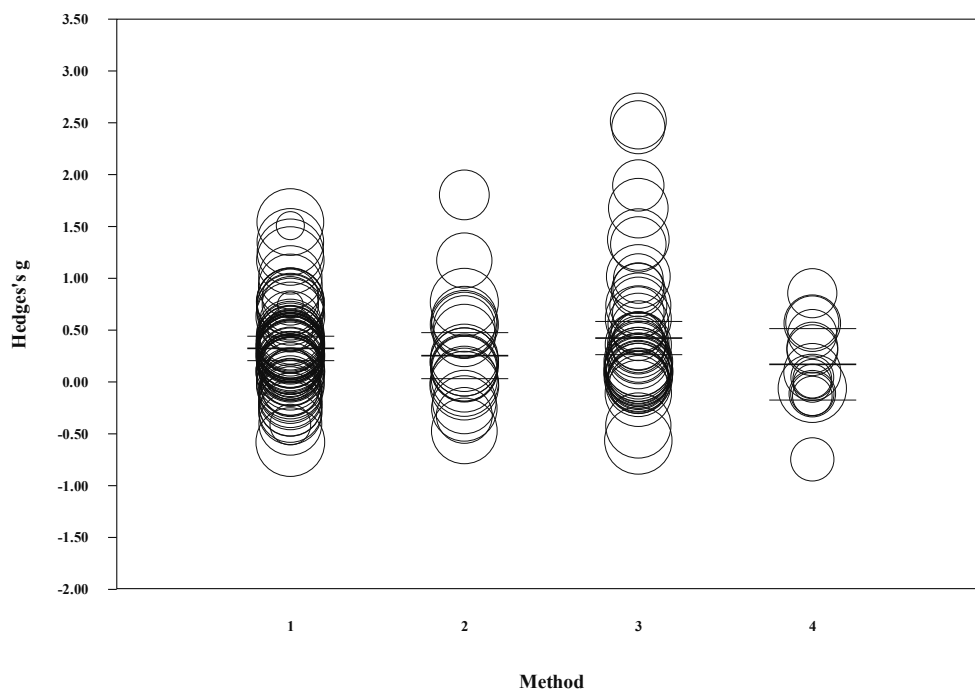


Figure 8. Regression of point estimates on study methodology. The values on the x-axis show correlational (1), single group-repeated measures (2), quasi-experimental (3), and experimental (4) codes.

Instrument validity. Point estimates from studies with references to external instrument validity showed the highest effect size ($g = .48$, 95% CI [.39, .56]), followed by face validity assessed by the investigator ($g = .42$, 95% CI [.16, .69]). Point estimates from studies with face validity assessed within the study ($g = .03$, 95% CI [-.13, .18]) and references to external instrument validity ($g = .12$, 95% CI [-.03, .26]) showed the lowest effect sizes, and were not statistically significant (see Table 21). The descriptive statistics for instrument validity included did not include a minimum of 10 point estimates per category, indicating that meta-regression was not appropriate. Furthermore, logical

combinations of validity categories (e.g., face validity vs investigator with face validity by study), would appear to combine seemingly disparate results.

Table 21

Effect Sizes and Null Tests for Instrument Validity

| Instrument validity | <i>n</i> | <i>g</i> | 95% CI | <i>Z</i> | <i>p</i> |
|--|----------|----------|-------------|----------|----------|
| Face validity (investigator) | 8 | .42 | [.16, .69] | 3.12 | .002 |
| Face validity (study) | 17 | .03 | [-.13, .18] | .31 | .757 |
| Reference external instrument | 72 | .48 | [.39, .56] | 11.52 | .000 |
| Reference external instrument validity | 20 | .12 | [-.03, .26] | 1.58 | .114 |
| Internal validity measures | 41 | .33 | [.23, .42] | 6.59 | .000 |

Instrument reliability. Point estimates from studies referencing external instruments or external instrument reliability showed the highest effect sizes ($g = .60$, 95% CI [.39, .81], and $g = .58$, 95% CI [.37, .78], respectively). Studies not reporting instrument reliabilities or reporting internal tests of validity with Cronbach's alpha less than .70 showed the lowest effect sizes ($g = .27$, 95% CI [-.039, .57], and $g = .26$, 95% CI [.12, .39], respectively). It was unlikely that any of the mean effect size point estimates was zero with the exception of studies that did not report instrument reliabilities ($Z = 1.71$, $p = .088$) (see Table 22).

Table 22

Effect Sizes and Null Tests for Instrument Reliability

| Instrument reliability | <i>n</i> | <i>g</i> | 95% CI | <i>Z</i> | <i>p</i> |
|--|----------|----------|--------------|----------|----------|
| Not reported | 6 | .27 | [-.039, .57] | 1.71 | .088 |
| References external instrument | 11 | .60 | [.39, .81] | 5.61 | .000 |
| References external instrument reliability | 14 | .58 | [.37, .78] | 5.45 | .000 |
| Internal reliability < .70 | 28 | .26 | [.12, .39] | 3.72 | .000 |
| Internal reliability > .70 | 99 | .30 | [.22, .37] | 7.79 | .000 |

The descriptive statistics for instrument reliability included a minimum of 10 point estimates in all but one category, indicating that meta-regression was not appropriate. However, as all but the base reference category were statistically significant, the investigator conducted a meta-regression of instrument reliability. A test of the regression model revealed that it was likely effect size differed by instrument reliability ($Q = 13.65, p = .008$). However, the model was incomplete, as there was unexplained variance between point estimates that shared similar instrument reliabilities ($Q = 4837, p < .0001$). The incremental changes in unexplained variance (T^2) for instrument reliabilities are presented in Table 23. The proportion of the unexplained variance that represented true variance, rather than error variance, was 96.84% ($I^2 = 96.84, I = .98$). This suggests that the observed variance around subgroup means would shrink by approximately 2% if the error variance were removed. The instrument reliability model explained a negligible amount of the total between-studies variance in effect sizes ($R^2 < .0001$).

Table 23

Meta-regression Model for Instrument Reliability

| Instrument reliability | Variance | | Test of Model | | | Regression | | |
|---------------------------------|----------|-------|---------------|------|-------|------------|-------|------|
| | T^2 | R^2 | Q | df | p | Coeff. | Z | p |
| Intercept (not reported) | .0898 | 0 | | | | .27 | 1.71 | .088 |
| References external instrument | .0933 | 0 | 8.36 | 1 | .0038 | .33 | 1.76 | .078 |
| References external reliability | .1105 | 0 | 14.11 | 2 | .0009 | .31 | 1.65 | .099 |
| Internal reliability < .70 | .1174 | 0 | 13.65 | 3 | .0034 | -.01 | -.043 | .965 |
| Internal reliability > .70 | .1178 | 0 | 13.65 | 4 | .0084 | .03 | .200 | .841 |

An examination of the regression coefficients for the instrument reliability model showed that only point estimates from studies with internal reliabilities less than .70 predicted decreases in engagement ($\beta = -.01, p = .965$). However, no category in the

regression model predicted increases or decreases that were statistically significant (see Table 23). Figure 9 shows a scatterplot of the regression model for instrument reliability.

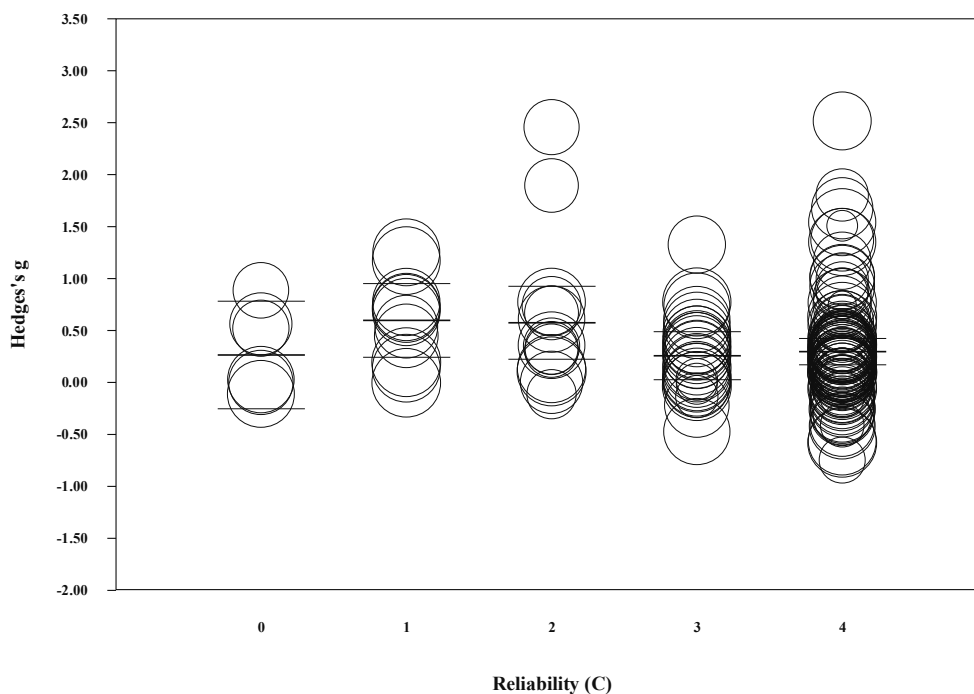


Figure 9. Regression of point estimates on instrument reliability. The values on the x-axis show unspecified (0), references external instrument (1), references external instrument reliability (2), internal reliability < .70 (3), and internal reliability > .70 (4) codes.

Repeat authors. Mean point estimates from studies with unique authors and repeat authors were similar ($g = .33$, 95% CI [.28, .39], and $g = .32$, 95% CI [.17, .48], respectively). It was unlikely that either of the mean effect size point estimates was zero (see Table 24).

The descriptive statistics for repeat authors included a minimum of 10 point estimates per category, indicating that meta-regression was appropriate. A test of the regression model revealed that it was unlikely effect size differed by whether or not they originated from studies with unique or repeat authors ($Q = .025$, $p = .873$). The model was incomplete, as there was unexplained variance between point estimates within the

unique or repeat authors category ($Q = 4985, p < .0001$). The incremental changes in unexplained variance (T^2) for repeat authors are presented in Table 25. The proportion of the unexplained variance that represented true variance, rather than error variance, was 96.87% ($I^2 = 96.87, I = .9842$). This suggests that the observed variance around subgroup means would shrink by approximately 2% if the error variance were removed. The repeat author model explained a negligible amount of the total between-studies variance in effect sizes ($R^2 < .0001$).

Table 24

Effect Sizes and Null Tests for Authors

| Authors | <i>n</i> | <i>g</i> | 95% CI | <i>Z</i> | <i>p</i> |
|---------|----------|----------|------------|----------|----------|
| Unique | 139 | .33 | [.28, .39] | 11.63 | .000 |
| Repeat | 19 | .32 | [.17, .47] | 4.26 | .000 |

Table 25

Meta-regression Model for Authors

| Authors | Variance | | Test of model | | | Regression | | |
|--------------------|----------|-------|---------------|-----------|----------|------------|----------|----------|
| | T^2 | R^2 | Q | <i>df</i> | <i>p</i> | Coeff. | <i>Z</i> | <i>p</i> |
| Unique (intercept) | .0898 | 0 | | | | .33 | 11.63 | .000 |
| Repeat | .0906 | 0 | .025 | 1 | .873 | -.01 | -.16 | .873 |

An examination of the regression coefficients for the author model showed that while point estimates from studies with repeat authors predicted decreases in engagement ($\beta = -.01, p = .873$), that change was not significant. Figure 10 shows a scatterplot of the regression model for authors.

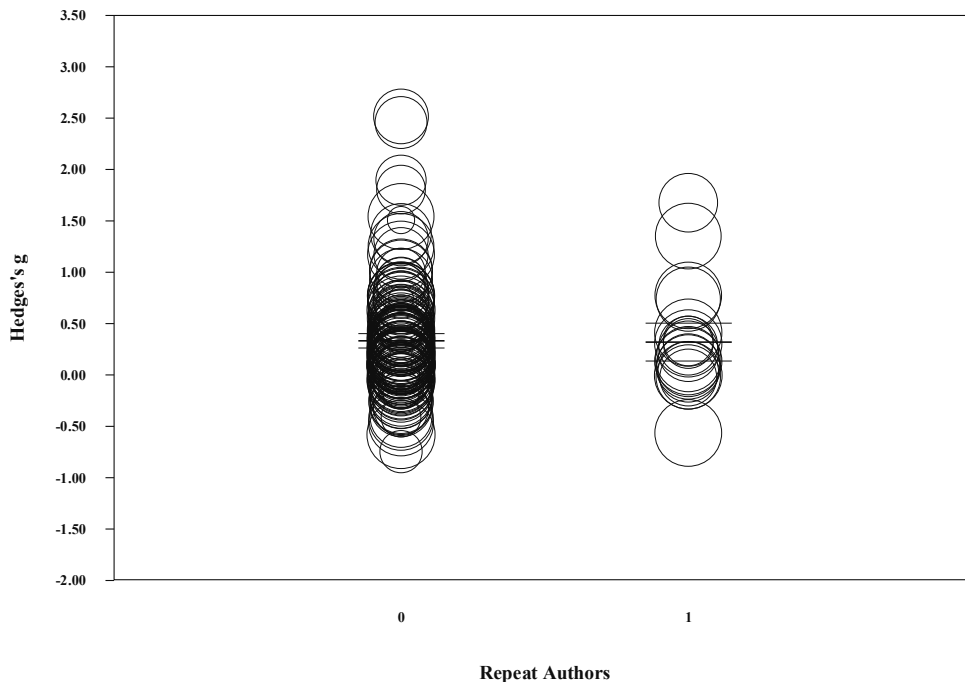


Figure 10. Regression of point estimates on author. The values on the x-axis show unique authors (0) and repeat authors (1).

Summary. Of the 12 coded moderators, five provided a minimum number of ten point estimates for each category. An additional two categories provided a minimum number of 10 point estimates for the majority of categories, and the investigator proceeded with meta-regression for those moderators (see Table 26). Omnibus tests revealed statistically significant results for four of these seven moderators—publication status, geographic location, school setting (with rural included), and instrument reliability. All of the coefficients for publication status and geographic location were statistically significant, while none of the coefficients for the instrument reliability model were significant. The school setting model that included the rural category results in three of five statistically significant coefficients.

These results suggest that publication status, geographic location, school setting, and instrument reliability were moderators of engagement in this meta-analysis. Tests of the regression models for study methodology and peer review status were just beyond the established range of statistical significance ($p = .093$ and $p = .123$, respectively). Thus, it is possible that these two variables could be moderators of engagement, though they did not manifest as statistically significant in this analysis. Clearly, whether studies originated from studies with unique or repeat authors did not moderate engagement results. For the remaining four variables, there was not enough data to make a determination about their potential moderating effects on engagement.

Table 26

Summary of Effect Sizes and Regression Models for Moderators

| Moderators | Point estimate categories | | | Regression model | | | |
|---------------------------|---------------------------|--------------------------|----------------------------|------------------|-----------|----------|---------------------------------|
| | <i>n</i> | Significant (<i>n</i>) | Minimum of 10 (<i>n</i>) | <i>Q</i> | <i>df</i> | <i>p</i> | Significant coeff. (<i>n</i>) |
| Publication status | 2 | 2 | 2 | 15.70 | 1 | .0007 | 2 |
| Geographic location | 2 | 2 | 2 | 11.28 | 1 | .0007 | 2 |
| School setting (w rural) | 5 | 4 | 4 | 14.81 | 4 | .0051 | 3 |
| Instrument reliability | 5 | 4 | 4 | 13.65 | 4 | .008 | 0 |
| School setting (no rural) | 4 | 4 | 4 | 7.17 | 3 | .067 | 2 |
| Study methodology | 4 | 3 | 4 | 6.41 | 3 | .093 | 1 |
| Peer review status | 2 | 2 | 2 | 2.38 | 1 | .123 | 1 |
| Repeat authors | 2 | 2 | 2 | .03 | 1 | .873 | 1 |
| School structure | 7 | 6 | 4 | - | - | - | - |
| School type | 4 | 3 | 3 | - | - | - | - |
| Instrument validity | 5 | 3 | 4 | - | - | - | - |
| Socioeconomic status | 5 | 3 | 3 | - | - | - | - |

Research Question 2: Practically Significant Predictors of Engagement. In

order to address the second research question: *what predictors have the largest practical effect on early adolescent's science engagement as assessed by student self-report?*, the

investigator utilized Ferguson's (2009) suggested guidelines for interpreting Hedges' g . Ferguson recommended a minimum practical effect size of $g > .41$, moderate $g > 1.15$, and strong $g > 2.70$. With these criteria, 107 point size estimates fell below the recommended minimum practical effect size, while 51 exceeded the minimum practical effect. No point estimates exceeded the guidelines for a strong practical effect ($g > 2.7$). See Table 27 for the distribution of effect sizes using Ferguson's recommended guidelines for interpreting Hedges' g .

Table 27

Distribution of Point Estimates

| Magnitude of point estimates | n | Percent |
|---|-----|---------|
| Strong ($g > 2.7$) | 0 | 0 |
| Moderate ($2.7 > g > 1.15$) | 13 | 8.2% |
| Minimum practical effect ($1.15 > g > .41$) | 38 | 24.1% |
| Below minimum practical effect ($g < .41$) | 107 | 67.7% |

The 51 practically significant effect sizes represented 32.3% of the 158 point estimates and 46.8% ($n = 37$) of included studies. Thirteen of the 51 practically significant effect sizes reflected moderate effects ($g > 1.15$), and two of those had effect sizes with magnitudes approaching classification as strong effects—a science-technology-society curriculum approach ($g = 2.5$, 95% CI [2.079, 2.947]) and project-based learning ($g = 2.5$, 95% CI [1.954, 2.953]). The difference between the project-based learning point estimate and the next was .562, confirming that the top two predictors were exceptional in terms of their engagement effects. The remaining 11 moderate effect size point estimates reflected a variety of predictors, including different instructional approaches (PBL, research, and scaffolding), self-determination theory components (autonomy and competence), and class characteristics (student-teacher

relationship and perceptions of class goals). Of the 38 small effect sizes ($1.15 > g > .41$), four point estimates had confidence intervals that spanned zero, suggesting the possibility that those predictors had no effect in the given studies. See Figure 11 for a forest plot of the small and moderate effect size point estimates.

The investigator also examined 34 negative point estimates to determine which predictors were negatively related to engagement. Four point estimates reflected predictors that would be expected to produce negative effect sizes, such as perceptions of the teacher as strict, admonishing, or dissatisfied. Of the remaining 30 negative estimates, the predictor with the most negative relationship to engagement was autonomy support comprised of procedural and cognitive components. This predictor had negative relationships with both affective ($g = -.75$) and behavioral ($g = -.14$) engagement. Additional autonomy predictors from this study also showed negative engagement effects (procedural autonomy supports on affective engagement: $g = -.12$, cognitive autonomy supports on behavioral engagement: $g = -.02$). Though this study utilized a true experimental design with random assignment (Furtak & Kunter, 2012), it is expected that the negative relationship of autonomy with engagement is spurious, as the predictor showed practically significant effects in five other point estimates.

Summary. Fifty-one of 158 point estimates showed practically significant effects on engagement. Of these, 11 showed moderate effect sizes, and reflected a variety of predictors, including different instructional approaches, self-determination theory components and class characteristics. In order to identify commonalities in the practically significant predictors, the investigator analyzed the predictors using both

descriptive and inferential statistical methods (see Research Question 3: Commonalities in Practically Significant Predictors).

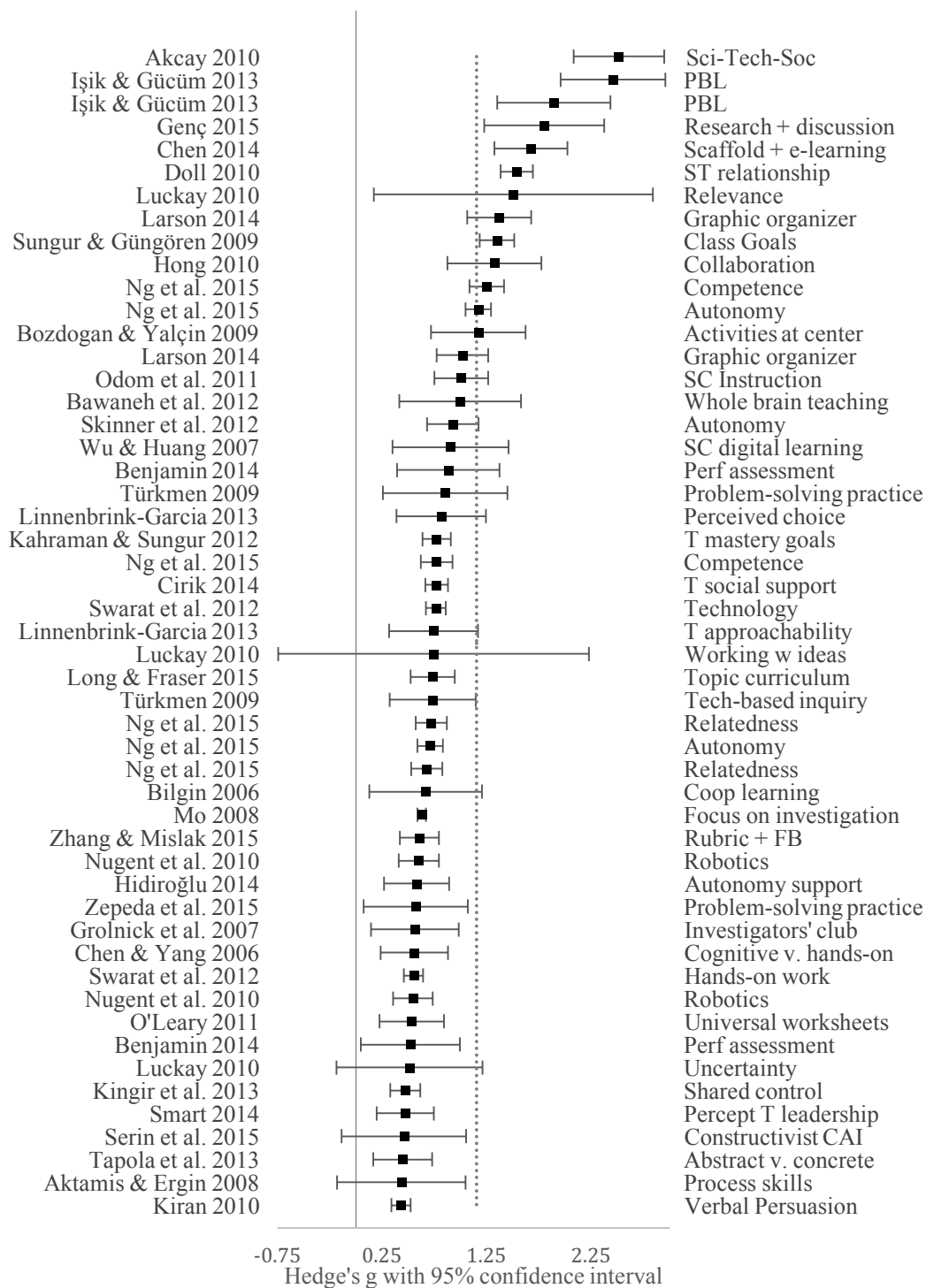


Figure 11. Forest plot of 51 engagement effect sizes with Hedges' g of greater than .41. Dashed line represents a moderate practical effect size ($g = 1.15$).

Research Question 3: Commonalities in Practically Significant Predictors. In

order to address the third research question: *what commonalities exist among predictors that have the largest practical effect on early adolescents' science engagement as assessed by student self-report?* the investigator analyzed predictors using both descriptive and inferential statistics. Mean point estimates for predictors were examined in terms of their statistical significance as well as to determine their significance in a regression model. Where possible, the effects of malleable classroom and task level predictors were considered in conjunction with statistically significant moderators of engagement in regression models.

Descriptive statistics. The investigator examined the distribution of engagement effect sizes for each predictor type. Instructional method predictors had the highest frequency of practically significant effect sizes ($n = 24$; 46%), the highest frequency of moderate effect sizes ($n = 7$, 12.8%) and the lowest frequency of negative effect sizes ($n = 9$, 15.8%). Though the other three categories of predictor types (technology, class characteristics, and social characteristics) yielded comparable frequencies of practically significant effects (26.7%, 28.3%, and 23%, respectively), technology had the highest frequency of negative effect sizes ($n = 5$, 33.3%). Further, there were no practically significant technology point estimates that represented moderate effects of greater than 1.15. Table 28 shows the distribution of effect sizes by predictor type.

Competence was the self-determination theory predictor with the highest frequency of practically significant effect sizes ($n = 17$, 58.6%), the highest frequency of moderate effect sizes ($n = 4$, 13.8%), and lowest frequency of negative effect sizes ($n = 1$, 3.4%). Autonomy and relatedness yielded similar frequencies of practically significant

point estimates ($n = 22$, 23.3% and $n = 11$, 31.4%, respectively) and negative point estimates ($n = 21$, 22.2% and $n = 8$, 22.9%, respectively).

Table 28

Distribution of Point Estimates by Predictor Classification

| Predictor Classification Type | Practically Significant Effect Sizes | | | | Practically Insignificant Effect Sizes | | | |
|----------------------------------|---|---------|------------------------------------|---------|---|---------|-------------------------|---------|
| | Moderate ($2.7 > g >$ 1.15) | | Small ($1.15 > g$ $> .41$) | | Small ($.41 > g \geq 0$) | | Negative ($g < 0$) | |
| | <i>n</i> | Percent | <i>n</i> | Percent | <i>n</i> | Percent | <i>n</i> | Percent |
| Instructional method | 7 | 12.8% | 17 | 29.8% | 24 | 42.1% | 9 | 15.8% |
| Technology | 0 | 0% | 4 | 26.7% | 6 | 40% | 5 | 33.3% |
| Class characteristics | 5 | 8.3% | 12 | 20% | 33 | 55% | 10 | 16.7% |
| Social characteristics | 1 | 3.8% | 5 | 19.2% | 15 | 57.7% | 5 | 19.2% |
| Self-determination theory | | | | | | | | |
| Autonomy | 5 | 5.3% | 17 | 18% | 51 | 54% | 21 | 22.2% |
| Competence | 4 | 13.8% | 13 | 44.8% | 11 | 37.9% | 1 | 3.4% |
| Relatedness | 4 | 11.4% | 7 | 20% | 16 | 45.7% | 8 | 22.9% |

Inferential statistics. The investigator analyzed mean effect sizes for each category and type of predictor. Meta-regression was used to analyze the predictive power of models for each predictor type, as well as combined models of predictors and moderators. Within each model, the investigator analyzed the regression coefficients in terms of both size and direction with respect to their effects on engagement.

Predictor classification: Type. Mean effect sizes were calculated for each category of predictor. Instructional method predictors showed the highest effect size ($g = .42$, 95% CI [.34, .51]), followed by class characteristics ($g = .34$, 95% CI [.25, .42], and social characteristics ($g = .25$, 95% CI [.12, .38]. For technology predictors ($g = .10$, 95% CI [-.06, .27]), it was possible that the effect size was zero ($Z = 1.23$, $p = .22$). See Table 29 for effect sizes and null tests of each predictor.

The investigator analyzed predictor type via meta-regression. A test of the predictor type regression model reveals that it was likely effect size differed by predictor type ($Q = 13.56, p = .004$). However, the model was incomplete, as there was unexplained variance between point estimates with the same predictor type ($Q = 4525, p < .0001$). The incremental changes in unexplained variance (T^2) as each category of predictor type was added to the regression model are presented in Table 30. The proportion of the unexplained variance that represents true variance, rather than error variance, was 96.60% ($I^2 = 96.60, I = .98$). This suggests that the observed variance around subgroup means would decrease by approximately 2% if the error variance were removed. The predictor type model explained 5% of the total between-studies variance in effect sizes ($R^2 = .05$).

Table 29

Effect Sizes and Null Tests for Predictor Classification: Type

| Predictor type | <i>n</i> | <i>g</i> | 95% CI | <i>Z</i> | <i>p</i> |
|------------------------|----------|----------|-------------|----------|----------|
| Instructional methods | 57 | .42 | [.34, .51] | 9.82 | .0000 |
| Technology | 15 | .10 | [-.06, .27] | 1.23 | .2201 |
| Class characteristics | 60 | .34 | [.25, .42] | 7.87 | .0000 |
| Social characteristics | 26 | .25 | [.12, .38] | 3.72 | .0002 |

Table 30

Meta-regression Model for Predictor Classification: Type

| Predictor type | Variance | | Test of model | | | Regression | | |
|----------------------------------|----------|-------|---------------|-----------|----------|------------|----------|----------|
| | T^2 | R^2 | Q | <i>df</i> | <i>p</i> | Coeff. | <i>Z</i> | <i>p</i> |
| Instructional method (intercept) | .0898 | 0 | | | | .42 | 9.82 | .000 |
| Technology | .09 | 0 | 7.92 | 1 | .005 | -.32 | -3.40 | .0006 |
| Class characteristics | .0888 | .01 | 8.42 | 2 | .015 | -.09 | -1.44 | 0.149 |
| Social characteristics | .0857 | .05 | 13.56 | 3 | .004 | -.18 | -2.22 | 0.026 |

An examination of the regression coefficients for the model suggested that technology, class, and social predictors predicted decreased engagement point estimates when compared to instructional methods. However, only the coefficients for technology ($\beta = -.32, p = .0006$) and social characteristics ($\beta = -.18, p = .027$) were statistically significant (see Table 30). Though the null test of technology ($Z = 1.23, p = .2201$) indicated that the mean effect size point estimate for technology predictors on could be zero, the regression model suggested the impact of technology predictors on the model was significant. Figure 12 shows a scatterplot of the regression model for predictor type classification.

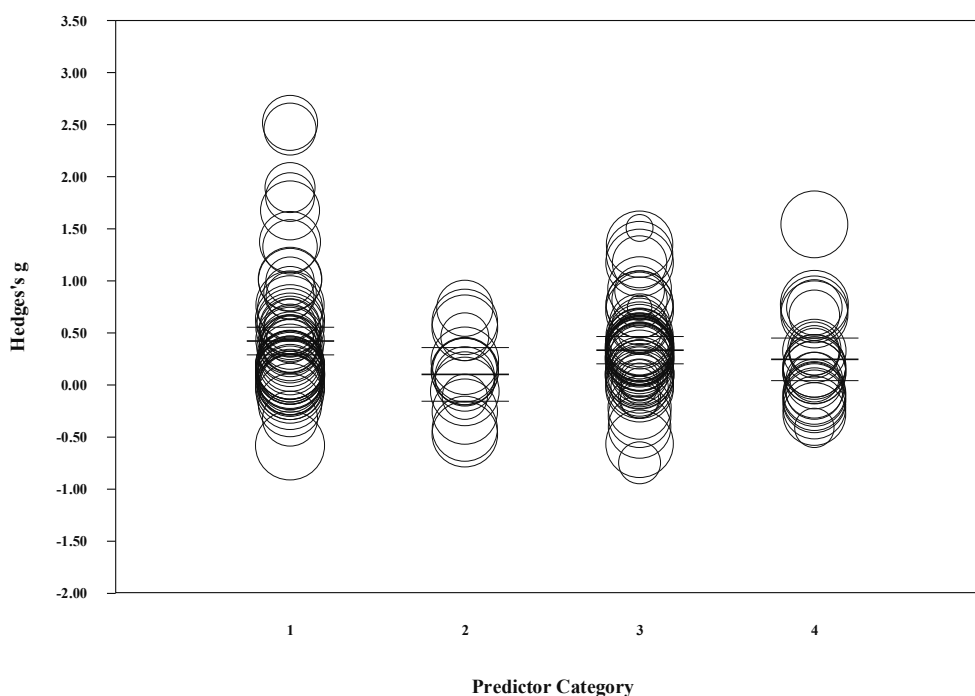


Figure 12. Regression of point estimates on predictor classification: type. The values on the x-axis show instructional methods (1), technology (2), class characteristics (3), and social characteristics (4) codes.

Predictor classification: Self-Determination Theory. Mean effect sizes were calculated for each category of SDT predictor. Competence showed the highest effect

size ($g = .56$, 95% CI [.44, .69]), and autonomy showing the lowest effect size ($g = .26$, 95% CI [.19, .33]). All of the SDT predictors were statistically significant. See Table 31 for effect sizes and null tests of each predictor.

The investigator analyzed SDT theory predictor type via meta-regression. Though it was likely that the effect size differed by SDT predictor type ($Q = 17.80$, $p = .0001$), the model explained a negligible amount of the between-studies variance in effect sizes ($R^2 < .001$). Goodness of fit tests confirmed that the predictor model left unexplained variance between point estimates within predictor subgroups ($Q = 4434$, $p < .001$). Nevertheless, an examination of the incremental changes to the model suggested that a model with just autonomy and competence explained 6% of the variance in effect sizes ($R^2 = .06$).

An examination of the regression coefficients for the model suggested that each SDT component predicted increased engagement (see Table 32). Furthermore, the coefficient for competence was statistically significant ($\beta = .31$, $p = .00002$) when compared to the intercept for autonomy. Though relatedness predicted increased engagement ($\beta = .08$), it was possible that the effect of relatedness predictors on engagement could be zero ($Z = 1.18$, $p = .236$). Figure 13 shows a scatterplot of the regression model for SDT predictor type.

Table 31

Effect Sizes and Null Tests for Predictor Classification: SDT

| SDT predictor type | <i>n</i> | <i>g</i> | 95% CI | <i>Z</i> | <i>p</i> |
|--------------------|----------|----------|------------|----------|----------|
| Autonomy | 94 | .26 | [.19, .33] | 7.31 | .0000 |
| Competence | 29 | .56 | [.44, .69] | 8.90 | .0000 |
| Relatedness | 35 | .34 | [.22, .46] | 5.74 | .0000 |

Table 32

Meta-regression Model for Predictor Classification: SDT

| SDT predictor type | Variance | | Test of model | | | Regression | | |
|----------------------|----------|-------|---------------|------|-------|------------|------|------|
| | T^2 | R^2 | Q | df | p | Coeff. | Z | p |
| Autonomy (intercept) | .0898 | 0 | | | | .26 | 7.31 | .000 |
| Competence | .0840 | 0.06 | 18.13 | 1 | .0000 | .31 | 4.22 | .000 |
| Relatedness | .0950 | 0.00 | 17.80 | 2 | .0001 | .08 | 1.18 | .236 |

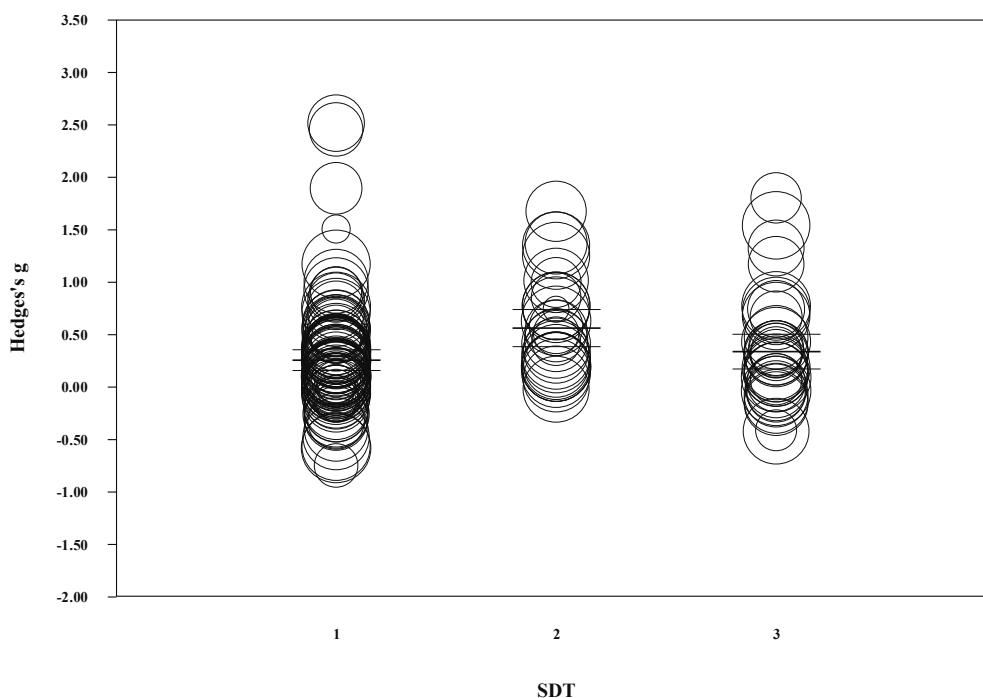


Figure 13. Regression of point estimates on predictor classification: SDT. The values on the x-axis show autonomy (1), competence (2), and relatedness (3).

Combined predictor/moderator models. The investigator considered the combined effect of predictors and statistically significant moderators through meta-regression. Four moderators were found to have statistically significant practical effects: publication status, geographic location, instrument reliability, and school setting (see Research Question 1: Moderators of Engagement). Combinations of each moderator with predictor type explained negligible variance in engagement. When the effect of each moderator was held constant, each predictor type predicted decreased engagement with respect to

instructional method. No models produced predictor categories with all statistically significant results. Though class characteristics did not have a statistically significant effect in a regression model with only predictor type ($\beta = -.09$, $p = .15$), the category was a significant predictor in regression models with geographic location ($\beta = -.19$, $p = .062$), and school setting ($\beta = -.14$, $p = .035$) (see Table 33).

Combinations of each moderator with self-determination theory predictor type explained negligible variance in engagement as well. No models produced SDT categories with all statistically significant results. None of the models increased the ability of relatedness to predict engagement to a statistically significant level (see Table 34).

Summary. Instructional methods had the highest representation of practically significant effect sizes in the predictor type classification, while competence had the highest representation for the self-determination theory predictor types. The mean point estimates for both of these predictor classification categories exceeded the minimum practical effect size and were statistically significant. Furthermore, the coefficients for instructional methods and competence predicted increases in engagement, and were statistically significant in the regression model.

The mean point estimates for these three remaining categories in the predictor type classification model—technology, class characteristics, and social characteristics—did not meet the minimum guidelines for a practically significant effect ($g > .41$). A meta-regression analysis of those categories revealed that all predicted decreases in engagement in relation to autonomy. However, the coefficient for class characteristics was not statistically significant. Though the coefficient for technology was statistically

significant, the mean point estimate was not. Thus, technology and social characteristics predicted decreases in engagement, when compared to instructional methods.

Table 33

Meta-regression Models for Combined Moderators and Predictors: Type

| Models | Test of model | | | Regression | | |
|---------------------------|---------------|-----------|----------|-------------------------------------|----------|----------|
| | <i>Q</i> | <i>df</i> | <i>p</i> | Coeff. | <i>Z</i> | <i>p</i> |
| Model 1 ^a | | | | | | |
| Intercept | | | | .23 | 3.58 | .0003 |
| Published | 15.70 | 1 | .0000 | .27 | 4.21 | .0000 |
| Technology | 22.01 | 2 | .0000 | -.30 | -3.14 | .0017 |
| Class characteristics | 22.48 | 3 | .0000 | -.08 | -1.27 | .2035 |
| Social characteristics | 30.97 | 4 | .0000 | -.23 | -2.85 | .004 |
| | | | | } <i>Q</i> = 14.75, <i>p</i> = .002 | | |
| Model 2 ^b | | | | | | |
| Intercept | | | | .34 | 6.70 | .0000 |
| Countries outside U.S. | 11.28 | 1 | .0008 | .22 | 3.43 | .0006 |
| Technology | 15.96 | 2 | .0003 | -.30 | -2.97 | .0030 |
| Class characteristics | 20.29 | 3 | .0001 | -.19 | -2.74 | .0062 |
| Social characteristics | 24.52 | 4 | .0001 | -.16 | -1.94 | .0529 |
| | | | | } <i>Q</i> = 13.42, <i>p</i> = .004 | | |
| Model 3 ^c | | | | | | |
| Intercept | | | | .26 | 1.72 | .085 |
| External instrument | 8.36 | 1 | .0038 | .41 | 2.14 | .032 |
| External reliability | 14.11 | 2 | .0009 | .41 | 2.19 | .029 |
| Internal reliability < .7 | 13.65 | 3 | .0034 | .08 | .481 | .631 |
| Internal reliability > .7 | 13.65 | 4 | .0085 | .14 | .866 | .387 |
| Technology | 18.94 | 5 | .0020 | -.32 | -2.90 | .004 |
| Class characteristics | 19.11 | 6 | .0040 | -.09 | -1.25 | .214 |
| Social characteristics | 25.30 | 7 | .0007 | -.22 | -2.38 | .017 |
| | | | | } <i>Q</i> = 11.30, <i>p</i> = .010 | | |
| Model 4 ^d | | | | | | |
| Intercept | | | | .51 | 9.15 | .000 |
| Rural | 8.81 | 1 | .0030 | -.51 | -2.99 | .003 |
| Suburban | 15.28 | 2 | .0005 | -.21 | -2.47 | .014 |
| Urban | 14.21 | 3 | .0026 | -.01 | -.162 | .871 |
| Mix | 14.81 | 4 | .0051 | -.08 | -.845 | .398 |
| Technology | 20.48 | 5 | .0010 | -.32 | -3.05 | .002 |
| Class characteristics | 23.23 | 6 | .0007 | -.14 | -2.11 | .035 |
| Social characteristics | 26.32 | 7 | .0004 | -.15 | -1.64 | .101 |
| | | | | } <i>Q</i> = 11.20, <i>p</i> = .011 | | |

^aModel 1: Publication status and prediction classification: type ($R^2 < .001$)

^bModel 2: Geographic location and predictor classification: type ($R^2 < .001$)

^cModel 3: Instrument reliability and predictor classification: type ($R^2 < .001$)

^dModel 4: School setting and predictor classification: type ($R^2 < .001$)

Table 34

Meta-regression Models for Combined Moderators and Predictors: SDT

| Model | Test of model | | | Regression | | |
|-------------------------------------|---------------|-----------|----------|------------|----------|----------|
| | <i>Q</i> | <i>df</i> | <i>p</i> | Coeff. | <i>Z</i> | <i>p</i> |
| Model 1 ^a | | | | | | |
| Intercept | | | | .13 | 2.23 | .026 |
| Published | 15.70 | 1 | .0000 | .20 | 2.98 | .003 |
| Competence | 27.40 | 2 | .0000 | .25 | 3.33 | .001 |
| Relatedness | 26.04 | 3 | .0000 | .05 | .701 | .483 |
| } <i>Q</i> = 11.13, <i>p</i> = .004 | | | | | | |
| Model 2 ^b | | | | | | |
| Intercept | | | | .16 | 3.28 | .001 |
| Countries outside U.S. | 11.28 | 1 | .0008 | .19 | 3.29 | .001 |
| Competence | 27.87 | 2 | .0000 | .30 | 4.10 | .000 |
| Relatedness | 27.98 | 3 | .0000 | .11 | 1.54 | .124 |
| } <i>Q</i> = 17.12, <i>p</i> < .001 | | | | | | |
| Model 3 ^c | | | | | | |
| Intercept | | | | .27 | 1.67 | .094 |
| External instrument | 8.36 | 1 | .0038 | .20 | .998 | .318 |
| External reliability | 14.11 | 2 | .0009 | .24 | 1.23 | .218 |
| Internal reliability < .7 | 13.65 | 3 | .0034 | -.07 | -3.88 | .699 |
| Internal reliability > .7 | 13.65 | 4 | .0085 | -.03 | -.174 | .862 |
| Competence | 18.94 | 5 | .0020 | .27 | 3.25 | .001 |
| Relatedness | 19.11 | 6 | .0040 | .06 | .752 | .452 |
| } <i>Q</i> = 10.56, <i>p</i> = .005 | | | | | | |
| Model 4 ^d | | | | | | |
| Intercept | | | | .32 | 7.19 | .000 |
| Rural | 8.81 | 1 | .003 | -.45 | -2.72 | .006 |
| Suburban | 15.28 | 2 | .000 | -.22 | -2.76 | .006 |
| Urban | 14.21 | 3 | .003 | -.002 | -.021 | .983 |
| Mix | 14.81 | 4 | .005 | -.14 | -1.39 | .165 |
| Competence | 31.09 | 5 | .0000 | .31 | 3.99 | .000 |
| Relatedness | 31.22 | 6 | .0000 | .11 | 1.53 | .127 |
| } <i>Q</i> = 16.20, <i>p</i> < .001 | | | | | | |

^aModel 1: Publication status and predictor classification: SDT ($R^2 < .001$)

^bModel 2: Geographic location and predictor classification: SDT ($R^2 < .001$)

^cModel 3: Instrument reliability and predictor classification: SDT ($R^2 < .001$)

^dModel 4: School setting and predictor classification: SDT ($R^2 < .001$)

A regression model combining publication status and predictor type improved the fit of predictor type with engagement ($Q = 14.75, p = .002$) when compared to predictor type alone ($Q = 13.56, p = .004$). In this model, publication status was the strongest predictor. When publication status was held constant, technology was the strongest

predictor, and produced a statistically significant decrease in engagement. Combinations of other statistically significant engagement moderators with predictor type did not improve the fit of predictor type with engagement. See Table 30 for the meta-regression model of predictor type and Table 33 for the meta-regression models of combined moderators and predictors.

The mean point estimates for the remaining self-determination theory predictor types—autonomy and relatedness—did not meet the minimum guidelines for a practically significant effect. In the meta-regression model, relatedness predicted an increase in engagement with respect to autonomy, though the coefficient was not statistically significant. This suggests that observed increases in engagement due to relatedness predictors could be due to chance and not true effects. Thus, competence was the only self-determination theory predictor type that reliably predicted an increase in engagement, with respect to autonomy.

Combinations of statistically significant engagement moderators with self-determination theory predictor types did not improve the fit of SDT predictor type with engagement. The values of each combination were less than the model fit of SDT predictor type alone ($Q = 17.80, p = .0001$). See Table X for the meta-regression model of SDT predictor type and Table X for the meta-regression models of combined moderators and SDT predictors.

Research Questions 4 and 5: Predictors of Engagement Types. In order to address the fourth and fifth research questions: *what predictors have the largest practical effect on early adolescents' behavioral, affective and cognitive engagement?* and *what were the commonalities in those predictors?* the investigator first examined the

descriptive statistics and mean point estimates for each coded engagement type category. The highest effect sizes resulted from studies measuring a combination of two engagement types ($g = .46, p < .001$), though these studies were small in number ($n = 13$). The lowest effect sizes resulted from studies assessing behavioral engagement or a combination of all three types of engagement ($g = .15, p = .214$, and $g = .19, p = .433$, respectively). However, both categories were small ($n = 10$ and $n = 2$, respectively) and neither mean effect size was statistically significant (see Table 35). Analyses of small categories were conducted with caution.

Table 35

Effect Sizes and Null Tests for Engagement Type

| Engagement type | <i>n</i> | <i>g</i> | 95% CI | <i>Z</i> | <i>p</i> |
|--------------------|----------|----------|-------------|----------|----------|
| Affective | 84 | .34 | [.26, .42] | 8.53 | .0000 |
| Behavioral | 10 | .15 | [-.08, .38] | .78 | .214 |
| Cognitive | 49 | .33 | [.24, .43] | 6.77 | .0000 |
| Two outcomes | 13 | .46 | [.27, .65] | 4.78 | .0000 |
| All three outcomes | 2 | .19 | [-.29, .68] | .78 | .433 |

Affective engagement predictors. Fifty-six studies generated 84 affective point size estimates. When combining across predictors, the summary mean affective engagement effect size was $g = .34$, 95% CI [.26, .42]. The 28 practically significant effect sizes represented 33.3% of the 84 affective point estimates and 41.4% ($n = 23$) of studies yielding affective point estimates. Nine of the practically significant point estimates represented moderate effects ($g > 1.15$), and one of those had an effect size with a magnitude approaching classification as a strong effect—a science-technology-society curriculum approach ($g = 2.51$, 95% CI [2.08, 2.95]). The difference between the curriculum approach point estimate and the next was .621, confirming that this predictor was exceptional in terms of its affective engagement effects. The remaining eight

moderate affective engagement effect sizes reflected a variety of predictors, including different instructional approaches (PBL, research, and field trip activities), self-determination theory components (autonomy/relevance and competence), and class characteristics (student-teacher relationships and collaboration). Of the 19 small effect sizes ($1.15 > g > .41$) three point estimates had confidence intervals that spanned zero, suggesting the possibility that those predictors had no effect in the given studies. See Table 36 for the distribution of affective point size estimates and Figure 14 for a forest plot of the small and moderate effect size point estimates.

Table 36

Distribution of Affective Point Estimates by Predictor Classification

| Predictor classification Type | Practically significant effect sizes | | | | Practically insignificant effect sizes | | | |
|----------------------------------|---|---------|-------------------------------|---------|---|---------|-------------------------|---------|
| | Moderate ($2.7 > g > 1.15$) | | Small ($1.15 > g > .41$) | | Small ($.41 > g \geq 0$) | | Negative ($g < 0$) | |
| | <i>n</i> | Percent | <i>n</i> | Percent | <i>n</i> | Percent | <i>n</i> | Percent |
| Instructional method | 4 | 12.9% | 6 | 19.4% | 15 | 48.4% | 6 | 19.4% |
| Technology | 0 | 0% | 3 | 27.3% | 5 | 45.5% | 3 | 27.3% |
| Class characteristics | 4 | 15.4% | 7 | 26.9% | 11 | 42.3% | 4 | 15.4% |
| Social characteristics | 1 | 6.3% | 3 | 18.8% | 6 | 37.5% | 6 | 37.5% |
| Self-determination theory | | | | | | | | |
| Autonomy | 4 | 8.7% | 9 | 19.6% | 22 | 47.8% | 11 | 23.9% |
| Competence | 1 | 7.1% | 7 | 50% | 6 | 42.9% | 0 | 0% |
| Relatedness | 4 | 16.7% | 3 | 12.5% | 9 | 37.5% | 8 | 33.3% |

The investigator also examined 19 negative affective effect sizes to determine which predictors were negatively related to engagement. Two point estimates reflected predictors that would be expected to produce negative effect sizes, including perceptions of the teacher as strict or dissatisfied. Interestingly, students' perceptions of their teachers as admonishing had a small, positive effect on affective engagement ($g = .18$, 95% CI [- .08, .44]). Of the remaining 17 negative estimates, the predictor with the most negative

relationship to affective engagement was autonomy support comprised of procedural and cognitive components ($g = -.75$, 95% CI[-1.49, -.01]).

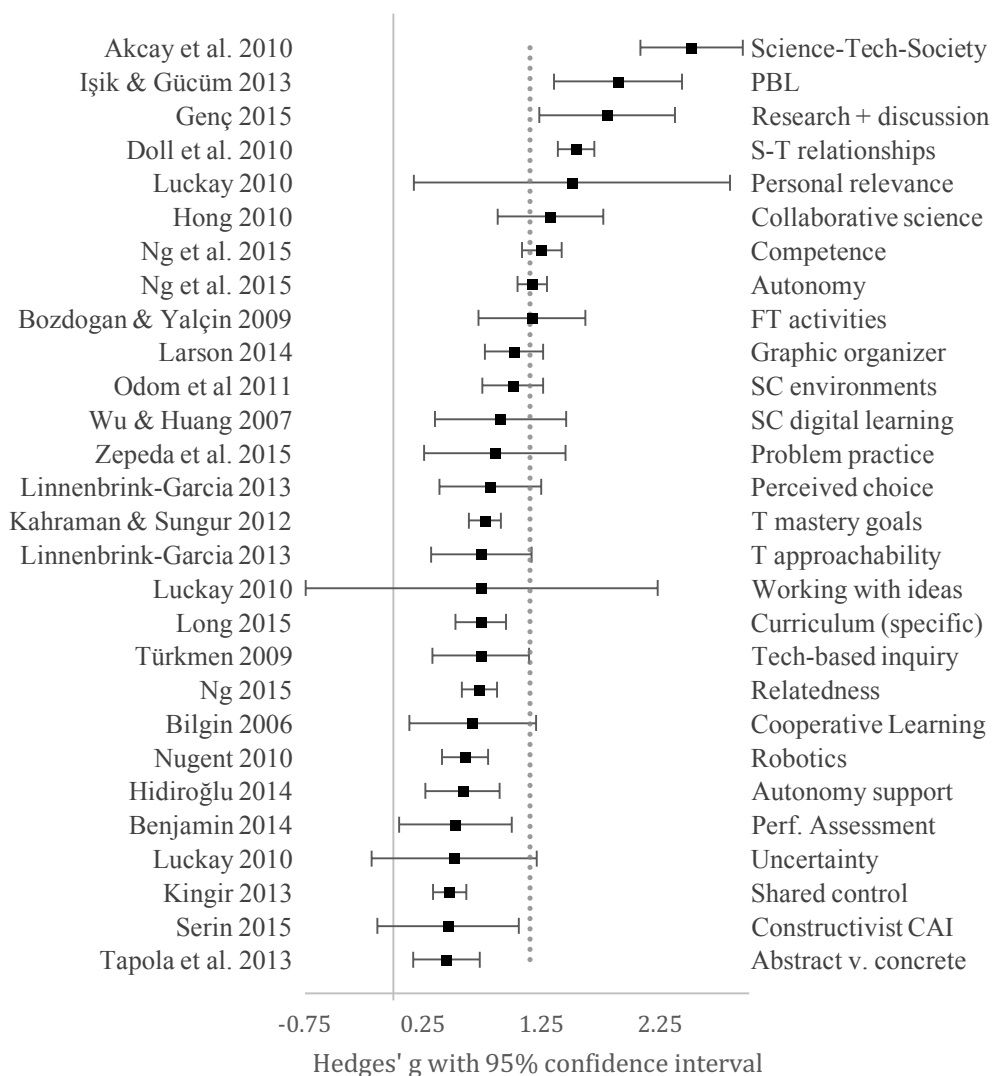


Figure 14. Forest plot of 28 affective engagement effect sizes with Hedges' g greater than .41. Dashed line represents the moderate practical effect size of $g = 1.15$.

The investigator examined the distribution of affective engagement effect sizes for each predictor type. Class characteristics had the highest frequency of practically significant effect sizes ($n = 11$; 42.3%), the highest frequency of moderate effect sizes ($n = 4$, 15.4%) and the lowest frequency of negative effect sizes ($n = 4$, 15.4%). Though the other three categories of predictor types (instructional method, technology, and social

characteristics) yielded comparable frequencies of practically significant effects (32.3%, 27.3%, and 25.1%, respectively), social characteristics had the highest frequency of negative effect sizes ($n = 6$, 37.5%). Further, there were no practically significant technology point estimates that represented moderate effects of greater than 1.15.

The investigator examined the mean effect size point estimate for each predictor type. Class characteristics and instructional methods showed the highest effect sizes ($g = .42$, 95% CI [.30, .53], and $g = .38$, 95% CI [.28, .48], respectively). Technology showed the lowest effect size ($g = .09$, 95% CI [-.08, .25]), though it was possible that the effect size was zero ($Z = 1.04$, $p = .30$). See Table 37 for effect sizes and null tests of each predictor.

The investigator analyzed predictor type via random-effects meta-regression. A test of the regression model revealed that it was likely affective engagement effect size differed by predictor type ($Q = 11.74$, $p = .008$), and the model explained 13.2% of the between-studies variance in affective engagement effect sizes ($R^2 = .132$). However, the model was incomplete, as there was unexplained variance between affective engagement point estimates with the same predictor type ($Q = 1776$, $p < .001$). The incremental changes in unexplained variance (T^2) for predictor type are presented in Table 38. The proportion of the unexplained variance that represented true variance, rather than error variance was 95.33% ($I^2 = 95.33$, $I = .98$). This suggests that the observed variance around subgroup means would shrink by approximately 2% if the error variance were removed.

An examination of the regression coefficients for the predictor type model showed that class characteristics ($\beta = .04$, $p = .622$) predicted increases in affective

engagement when compared to the intercept for instructional method ($\beta = .38$, $p = .000$). However, this predicted increase was not statistically significant. Alternatively, technology ($\beta = -.29$, $p = .003$) and social characteristics ($\beta = -.10$, $p = .273$) predicted decreases in affective engagement. Only the coefficient for technology was statistically significant when compared to the intercept for instructional method (see Table 38). Thus, though the predictor type model suggested that affective engagement differed by predictor type, only technology had a statistically significant effect, and that effect was negative with respect to affective engagement. Further, the mean point estimate for technology was not statistically significant. Figure 15 shows a scatterplot of the regression model for predictor type.

Table 37

Effect Sizes and Null Tests for Affective Engagement by Predictor Classification: Type

| Predictor type | <i>n</i> | <i>g</i> | 95% CI | <i>Z</i> | <i>p</i> |
|------------------------|----------|----------|-------------|----------|----------|
| Instructional method | 31 | .38 | [.28, .48] | 7.36 | .000 |
| Technology | 11 | .09 | [-.08, .25] | 1.04 | .30 |
| Class characteristics | 26 | .42 | [.30, .53] | 6.99 | .000 |
| Social characteristics | 16 | .28 | [.12, .43] | 3.55 | .000 |

Table 38

Meta-regression Model for Affective Predictor Classification: Type

| Predictor type | Variance | | Test of model | | | Regression | | |
|----------------------------------|-----------------------|-----------------------|---------------|-----------|----------|------------|----------|----------|
| | <i>T</i> ² | <i>R</i> ² | <i>Q</i> | <i>df</i> | <i>p</i> | Coeff. | <i>Z</i> | <i>p</i> |
| Instructional method (intercept) | .0703 | 0 | | | | .38 | 7.36 | .000 |
| Technology | .0711 | 0 | 8.49 | 1 | .004 | -.29 | -2.94 | .003 |
| Class characteristics | .0655 | .07 | 9.89 | 2 | .007 | .04 | .493 | 0.622 |
| Social characteristics | .0610 | .013 | 11.74 | 3 | .008 | -.10 | -1.10 | 0.273 |

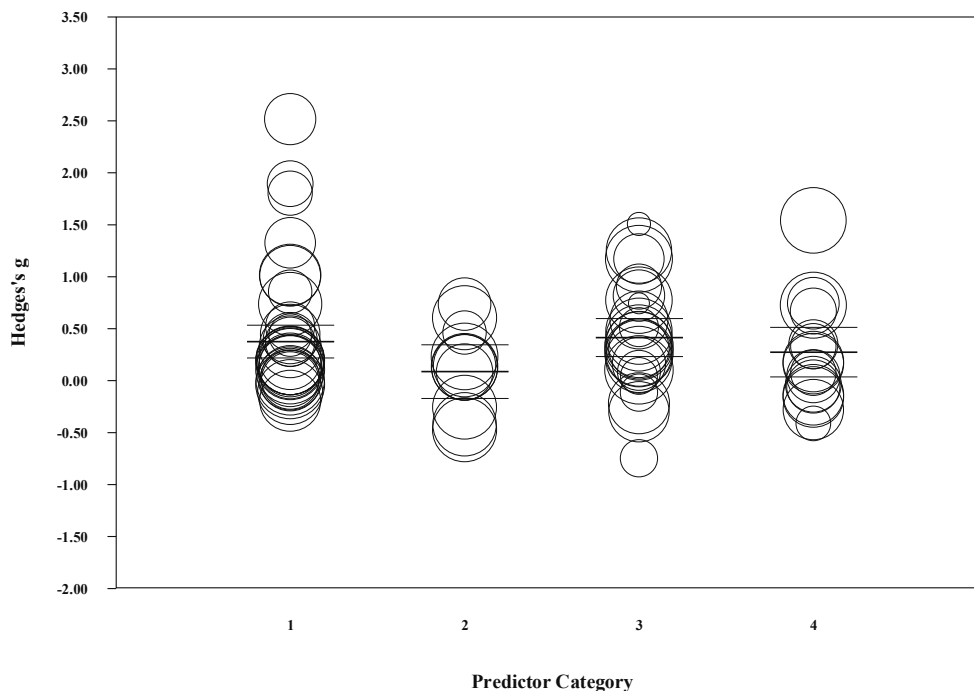


Figure 15. Regression of affective point estimates on predictor classification: type. The values on the x-axis show instructional methods (1), technology (2), class characteristics (3), and social characteristics (4).

The investigator considered the combined effect of affective engagement predictor types and statistically significant moderators through meta-regression. Four moderators produced statistically significant practical effects: publication status, geographic location, instrument reliability, and school setting. Combinations of school setting and instrument reliability with affective predictor type explained negligible variance in affective engagement. Though instrument reliability was a statistically significant moderator of engagement point estimates, it was not a statistically significant moderator of affective engagement point estimates ($Q = 7.19, p = .126$). Models which combined publication status or geographic location explained 14% and .6% of the variance in affective engagement point estimates, respectively. When the effect of publication status was held constant, the ability of predictor type to predict affective engagement increased slightly ($Q = 12, p = .007$) when compared to predictor type alone

($Q = 11.74, p = .008$). Though class characteristics did not have a statistically significant effect in a regression model with only predictor type ($\beta = -.01, p = .15$), the category was a significant predictor in regression models with geographic location ($\beta = -.19, p = .062$), and school setting ($\beta = -.14, p = .035$) No models produced predictor categories with all statistically significant results (see Table 39).

The investigator examined the distribution of affective engagement effect sizes by self-determination theory predictor type. Competence was the self-determination theory predictor with the highest frequency of practically significant affective engagement effect sizes ($n = 8, 57.1\%$) and no negative effect sizes. Relatedness had the highest frequency of moderate effect sizes ($n = 4, 16.7\%$). Autonomy predictors were largely insignificant or negative with respect to affective engagement ($n = 33, 71.7\%$).

The investigator examined the mean effect size point estimates for each SDT predictor type. Competence showed the highest effect size ($g = .53, 95\% \text{ CI } [.34, .71]$), while autonomy showed the lowest effect size ($g = .27, 95\% \text{ CI } [.17, .37]$). It was unlikely that any of the effect sizes were zero. See Table 40 for effect sizes and null tests of each SDT predictor on affective engagement.

The investigator analyzed SDT theory predictor type via meta-regression. It was unlikely that the affective engagement effect size differed by SDT predictor type ($Q = 4.49, p = .06$), and the model explained a negligible amount of the between-studies variance in effect sizes ($R^2 < .001$). Goodness of fit tests confirmed that the predictor model left unexplained variance in affective engagement within SDT subgroups ($I^2 = .07$), and that the true effect size still differed from study to study within those subgroups ($Q = 1776, p < .0001$) (see Table 41).

Table 39

Meta-regression Models for Combined Moderators and Affective Predictors: Type

| Model | Test of model | | | Regression | | |
|-------------------------------------|---------------|-----------|----------|------------|----------|----------|
| | <i>Q</i> | <i>df</i> | <i>p</i> | Coeff. | <i>Z</i> | <i>p</i> |
| Model 1 ^a | | | | | | |
| Intercept | | | | .09 | 1.15 | .251 |
| Published | 18.91 | 1 | .0000 | .36 | 4.62 | .000 |
| Technology | 25.93 | 2 | .0000 | -.26 | -2.63 | .008 |
| Class characteristics | 29.73 | 3 | .0000 | .06 | .738 | .461 |
| Social characteristics | 33.09 | 4 | .0000 | -.15 | -1.57 | .118 |
| } <i>Q</i> = 12, <i>p</i> = .007 | | | | | | |
| Model 2 ^b | | | | | | |
| Intercept | | | | .32 | 5.21 | .0000 |
| Countries outside U.S. | 8.36 | 1 | .0038 | .18 | 2.35 | .0186 |
| Technology | 14.76 | 2 | .0006 | -.29 | -2.76 | .0058 |
| Class characteristics | 14.27 | 3 | .0026 | -.05 | -.592 | .5541 |
| Social characteristics | 16.20 | 4 | .0028 | -.10 | -1.06 | .289 |
| } <i>Q</i> = 7.78, <i>p</i> = .051 | | | | | | |
| Model 3 ^c | | | | | | |
| Intercept | | | | .08 | .413 | .680 |
| External instrument | 4.34 | 1 | .0373 | .51 | 2.26 | .024 |
| External reliability | 6.37 | 2 | .0414 | .59 | 2.52 | .012 |
| Internal reliability < .7 | 5.87 | 3 | .1180 | .28 | 1.27 | .204 |
| Internal reliability > .7 | 7.19 | 4 | .1260 | .34 | -2.45 | .014 |
| Technology | 11.73 | 5 | .0386 | -.32 | -2.45 | .014 |
| Class characteristics | 12.75 | 6 | .0472 | -.01 | -.121 | .904 |
| Social characteristics | 16.34 | 7 | .0222 | -.20 | -1.68 | .094 |
| } <i>Q</i> = 8.26, <i>p</i> = .041 | | | | | | |
| Model 4 ^d | | | | | | |
| Intercept | | | | .45 | 7.01 | .000 |
| Rural | 9.89 | 1 | .0017 | -.48 | -3.06 | .002 |
| Suburban | 13.70 | 2 | .0011 | -.21 | -1.80 | .072 |
| Urban | 13.49 | 3 | .0037 | -.05 | -.555 | .579 |
| Mix | 11.90 | 4 | .0181 | .08 | .743 | .457 |
| Technology | 19.65 | 5 | .0015 | -.33 | -3.09 | .002 |
| Class characteristics | 21.83 | 6 | .0013 | -.01 | -.164 | .870 |
| Social characteristics | 24.47 | 7 | .0009 | -.11 | -1.06 | .290 |
| } <i>Q</i> = 10.66, <i>p</i> = .014 | | | | | | |

^aModel 1: Publication status and prediction classification: type ($R^2 = .14$)

^bModel 2: Geographic location and predictor classification: type ($R^2 = .006$)

^cModel 3: Instrument reliability and predictor classification: type ($R^2 < .001$)

^dModel 4: School setting and predictor classification: type ($R^2 < .001$)

Table 40

Effect Sizes and Null Tests for Affective Engagement by Predictor Classification: SDT

| SDT predictor | <i>n</i> | <i>g</i> | 95% CI | <i>Z</i> | <i>p</i> |
|---------------|----------|----------|------------|----------|----------|
| Autonomy | 46 | .27 | [.17, .37] | 5.27 | .000 |
| Competence | 14 | .53 | [.34, .71] | 5.54 | .000 |
| Relatedness | 24 | .35 | [.21, .50] | 4.73 | .000 |

Table 41

Meta-regression Model for Affective Predictor Classification: SDT

| SDT predictor | Variance | | Test of Model | | | Regression | | |
|----------------------|-----------------------|-----------------------|---------------|-----------|----------|------------|----------|----------|
| | <i>T</i> ² | <i>R</i> ² | <i>Q</i> | <i>df</i> | <i>p</i> | Coeff. | <i>Z</i> | <i>p</i> |
| Autonomy (intercept) | .0703 | 0 | | | | .27 | 5.27 | .000 |
| Competence | .0788 | 0 | 5.88 | 1 | .015 | .26 | 2.35 | .019 |
| Relatedness | .0982 | 0 | 5.59 | 2 | .06 | .08 | .876 | .381 |

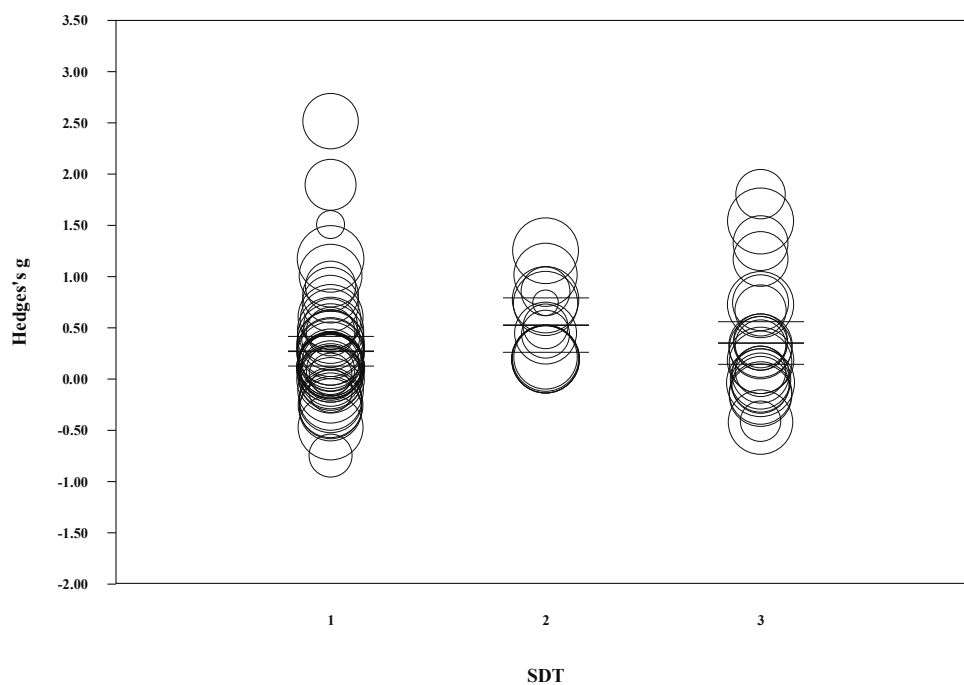


Figure 16. Regression of affective point estimates on predictor classification: SDT. The values on the x-axis show autonomy (1), competence (2), and relatedness (3).

An examination of the regression coefficients for the model suggested that both competence and relatedness predicted increased affective engagement point estimates

when compared to autonomy. However, only the coefficient for competence was statistically significant ($\beta = .26$, $p = .019$). Figure 16 shows a scatterplot of the regression model for SDT predictor classification.

Table 42

Meta-Regression Models for Combined Moderators and Affective Predictors: SDT

| Model | Test of model | | | Regression | | |
|---------------------------|---------------|-----------|----------|------------------------|----------|----------|
| | <i>Q</i> | <i>df</i> | <i>p</i> | Coeff. | <i>Z</i> | <i>p</i> |
| Model 1 ^a | | | | | | |
| Intercept | | | | .04 | .417 | .677 |
| Published | 18.91 | 1 | .000 | .32 | 3.41 | .001 |
| Competence | 20.98 | 2 | .000 | .19 | 1.72 | .085 |
| Relatedness | 17.41 | 3 | .001 | .06 | .71 | .480 |
| | | | | } $Q = 3.01, p = .022$ | | |
| Model 2 ^b | | | | | | |
| Intercept | | | | .17 | 2.46 | .014 |
| Countries outside U.S. | 8.36 | 1 | .004 | .20 | 2.51 | .012 |
| Competence | 13.36 | 2 | .001 | .26 | 2.31 | .021 |
| Relatedness | 11.70 | 3 | .008 | .10 | 1.12 | .261 |
| | | | | } $Q = 5.62, p = .060$ | | |
| Model 3 ^c | | | | | | |
| Intercept | | | | .08 | .395 | .693 |
| External instrument | 4.34 | 1 | .037 | .39 | 1.49 | .137 |
| External reliability | 6.37 | 2 | .041 | .44 | 1.67 | .094 |
| Internal reliability < .7 | 5.87 | 3 | .118 | .18 | .740 | .459 |
| Internal reliability > .7 | 7.19 | 4 | .126 | .19 | .867 | .386 |
| Competence | 8.89 | 5 | .114 | .20 | 1.47 | .141 |
| Relatedness | 8.56 | 6 | .200 | .04 | .381 | .703 |
| | | | | } $Q = 2.17, p = .034$ | | |
| Model 4 ^d | | | | | | |
| Intercept | | | | .33 | 4.88 | .000 |
| Rural | .989 | 1 | .002 | -.46 | -2.39 | .017 |
| Suburban | 13.70 | 2 | .001 | -.255 | -1.84 | .065 |
| Urban | 13.49 | 3 | .004 | -.02 | -.175 | .861 |
| Mix | 11.90 | 4 | .018 | .02 | 2.02 | .043 |
| Competence | 15.91 | 5 | .007 | .24 | 2.02 | .043 |
| Relatedness | 13.44 | 6 | .037 | .10 | .981 | .327 |
| | | | | } $Q = 4.27, p = .118$ | | |

^aModel 1: Publication status and predictor classification: SDT ($R^2 < .001$)

^bModel 2: Geographic location and predictor classification: SDT ($R^2 < .001$)

^cModel 3: Instrument reliability and predictor classification: SDT ($R^2 < .001$)

^dModel 4: School setting and predictor classification: SDT ($R^2 < .001$)

The investigator considered the combined effect of affective engagement SDT predictor types and statistically significant moderators through meta-regression. Combinations of each school setting and instrument reliability with predictor type explained negligible variance in affective engagement. Though instrument reliability was a statistically significant moderator of engagement point estimates, it was not a statistically significant moderator of affective engagement point estimates ($Q = 7.20, p = .126$). When the effect of geographic location was held constant, the ability of SDT predictor type to predict affective engagement increased slightly ($Q = 5.62, p = .06$) when compared to SDT predictor type alone ($Q = 5.59, p = .06$). However, the predicted increases were not statistically significant. No models produced predictor categories with all statistically significant results (see Table 42).

Cognitive engagement predictors. Thirty-one studies generated 49 cognitive point size estimates. When combining across predictors, the summary mean cognitive engagement effect size was $g = .60$, 95% CI [.44, .76]. The 12 practically significant effect sizes represented 24.5% of the 49 cognitive point estimates and 32.3% ($n = 10$) of studies yielding cognitive point estimates. Three of the practically significant point estimates represented moderate effects ($g > 1.15$), and one of those had an effect size with a magnitude approaching classification as a strong effect—a project-based learning approach ($g = 2.45$, 95% CI [1.954, 2.953]). The difference between the project-based learning point estimate and the next was .781, confirming that this predictor was exceptional in terms of its cognitive engagement effects. The predictors for the remaining two moderate cognitive engagement effect sizes were scaffolding with e-learning and perceptions of classroom goal structure. See Table 43 for the distribution of cognitive

point size estimates and Figure 17 for a forest plot of the small and moderate effect size point estimates.

Table 43

Distribution of Cognitive Point Estimates by Predictor Classification

| Predictor classification | Practically significant effect sizes | | | | Practically insignificant effect sizes | | | |
|---------------------------|--------------------------------------|---------|-------------------------------|---------|--|---------|-------------------------|---------|
| | Moderate ($2.7 > g > 1.15$) | | Small ($1.15 > g > .41$) | | Small ($.41 > g \geq 0$) | | Negative ($g < 0$) | |
| | <i>n</i> | Percent | <i>n</i> | Percent | <i>n</i> | Percent | <i>n</i> | Percent |
| Type | | | | | | | | |
| Instructional method | 2 | 16.7% | 3 | 25% | 6 | 50% | 1 | 8.3% |
| Technology | 0 | 0% | 1 | 33.3% | 1 | 33.3% | 1 | 33.3% |
| Class characteristics | 1 | 4.2% | 3 | 12.5% | 17 | 70.8% | 3 | 12.5% |
| Social characteristics | 0 | 0% | 2 | 20% | 5 | 50% | 3 | 30% |
| Self-determination theory | | | | | | | | |
| Autonomy | 1 | 3.3% | 3 | 10% | 20 | 66.7% | 6 | 20% |
| Competence | 2 | 25% | 2 | 25% | 3 | 37.5% | 1 | 12.5% |
| Relatedness | 0 | 0% | 4 | 36.4% | 6 | 54.5% | 1 | 9.1% |

The investigator also examined 11 negative affective effect sizes to determine which predictors were negatively related to cognitive engagement. Two point estimates reflected predictors that would be expected to produce negative effect sizes, including perceptions of the teacher as strict or admonishing. Interestingly, students' perceptions of their teachers as dissatisfied had a small, positive effect on cognitive engagement ($g = .12$, 95% CI [-.144, .384]). Of the remaining nine negative estimates, the predictor with the most negative relationship to cognitive engagement was perception of student freedom ($g = -.41$, 95% CI[-.676, -.138]).

The investigator examined the distribution of cognitive engagement effect sizes for each predictor type. Instructional method had the highest frequency of practically significant effect sizes ($n = 5$; 41.7%), the highest frequency of moderate effect sizes ($n = 2$, 16.7%) and the lowest frequency of negative effect sizes ($n = 1$, 8.3%). Class

characteristics had the highest frequency of practically insignificant or negative effect sizes ($n = 20$, 83.3%). Technology and social characteristics had similar frequencies of negative cognitive engagement predictors ($n = 1$, 33.3%, and $n = 3$, 30%, respectively), though only three point estimates reflected technology as a cognitive engagement predictor.

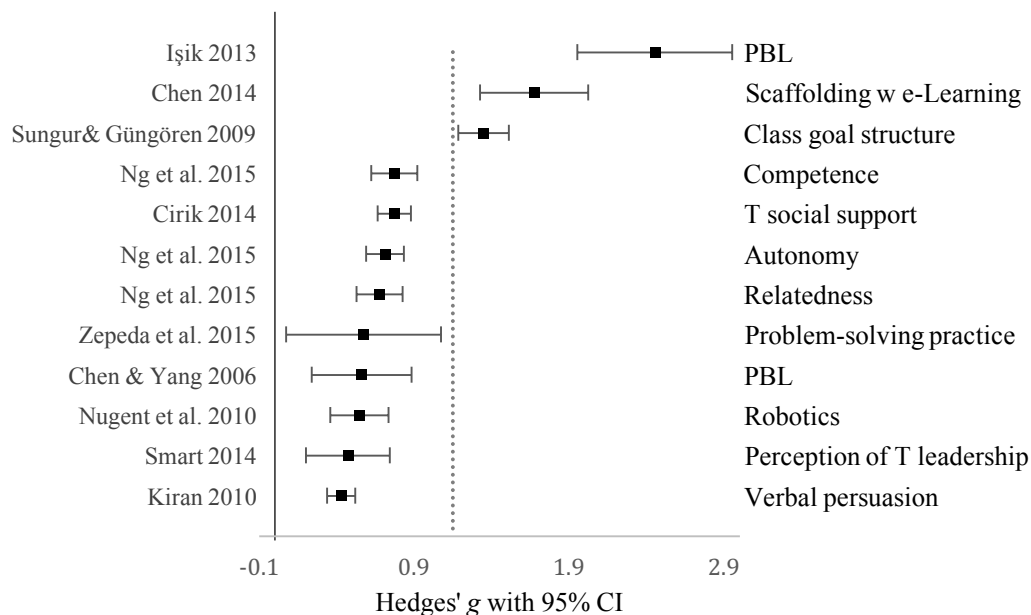


Figure 17. Forest plot of 12 cognitive engagement effect sizes with Hedges' g greater than .41. Dashed line represents a moderate practical effect size of $g = 1.15$.

Competence was the self-determination theory predictor with the highest frequency of practically significant cognitive engagement effect sizes ($n = 4$, 50%), and the highest frequency of moderate effect sizes ($n = 2$, 25%). Though relatedness produced no moderate point estimates, it did produce four (36.4%) small, practically-significant effect sizes. Autonomy predictors were largely insignificant or negative with respect to affective engagement ($n = 26$, 86.7%).

The investigator examined the mean effect size point estimates for each predictor type. Instructional methods showed the highest effect size ($g = .49$, 95% CI [.33, .66]).

The remaining three categories—technology, class characteristics, and social characteristics showed similar effect sizes (see Table 44). The effect size for technology was not significant ($Z = 1.28, p = .200$), though there were also only three cognitive engagement point estimates for that predictor.

Table 44

Effect Sizes and Null Tests for Cognitive Engagement by Predictor Classification: Type

| Predictor type | <i>n</i> | <i>g</i> | 95% CI | <i>Z</i> | <i>p</i> |
|------------------------|----------|----------|-------------|----------|----------|
| Instructional method | 12 | .49 | [.33, .66] | 5.83 | .000 |
| Technology | 3 | .24 | [-.13, .60] | 1.28 | .200 |
| Class characteristics | 24 | .31 | [.20, .41] | 5.51 | .000 |
| Social characteristics | 10 | .23 | [.05, .41] | 2.47 | .014 |

The investigator analyzed predictor type via random-effects meta-regression, though such an analysis was conducted with caution, as there were only three point estimates reflecting technology as a predictor. The predictor type model explained a negligible portion of between-studies variance in cognitive engagement effect sizes ($R^2 < .0001$), and it was unlikely that the cognitive engagement effect sizes differed by predictor type ($Q = 5.52, p = .138$). The model was incomplete, as there was unexplained variance between cognitive engagement point estimates with the same predictor type ($Q = 791, p = 0.0000$). The incremental changes in unexplained variance (T^2) for predictor type are presented in Table 45. The proportion of the unexplained variance that represented true variance, rather than error variance was 94.31% ($I^2 = 94.31, I = .9711$). This suggests that the observed variance around subgroup means would shrink by approximately 3% if the error variance were removed.

An examination of the regression coefficients for the model suggested that technology, class, and social predictors predicted decreased cognitive engagement point

estimates when compared to instructional methods. However, only the coefficient for social characteristics ($\beta = -.27$, $p = .031$) was statistically significant (see Table 45).

Figure 18 shows a scatterplot of the regression model for predictor type classification.

Table 45

Meta-regression Model for Cognitive Predictor Classification: Type

| Predictor type | Variance | | Test of model | | | Regression | | |
|----------------------------------|----------|-------|---------------|------|------|------------|-------|------|
| | T^2 | R^2 | Q | df | p | Coeff. | Z | P |
| Instructional method (intercept) | .0579 | 0 | | | | .49 | 5.83 | .000 |
| Technology | .0580 | 0 | .251 | 1 | .616 | -.26 | -1.27 | .203 |
| Class characteristics | .0684 | 0 | .884 | 2 | .643 | -.19 | -1.87 | .062 |
| Social characteristics | .0656 | 0 | 5.52 | 3 | .138 | -.27 | -2.15 | .031 |

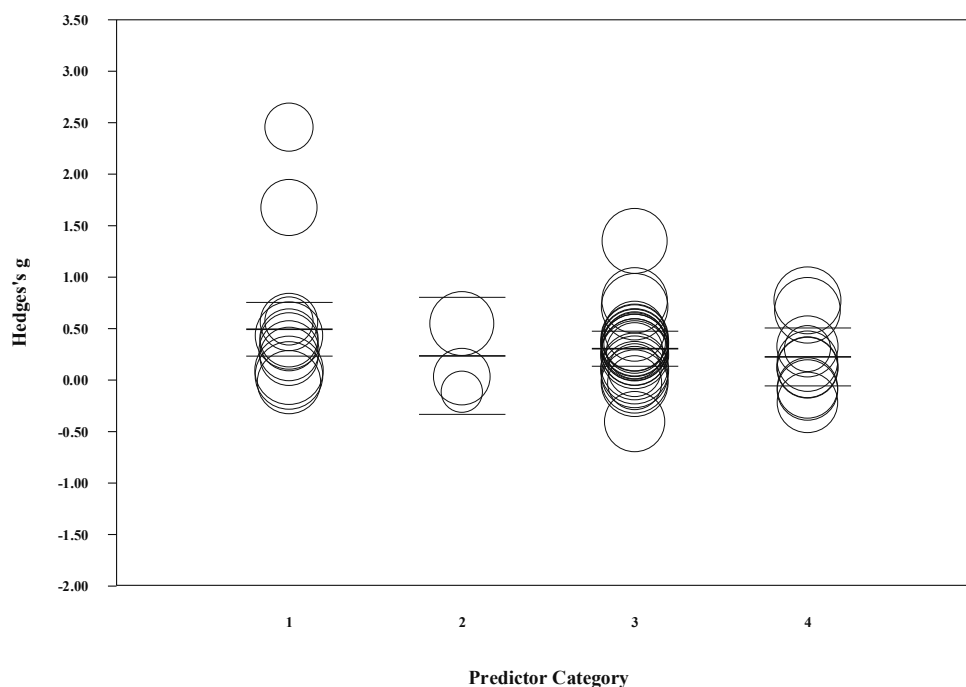


Figure 18. Regression of cognitive point estimates on predictor classification: type. The values on the x-axis show instructional methods (1), technology (2), class characteristics (3), and social characteristics (4).

The investigator examined the mean effect size point estimates for each SDT predictor type. Competence showed the highest effect size ($g = .61$, 95% CI [.41, .81]. Autonomy and relatedness predictors showed similar mean cognitive effect sizes (see

Table 46). All of the point estimates for all the SDT predictors were statistically significant.

The investigator analyzed SDT predictor type via random-effects meta-regression, though such an analysis was conducted with caution, as there were only eight point estimates reflecting competence as a cognitive engagement predictor. The predictor type model explained a negligible portion of between-studies variance in cognitive engagement effect sizes ($R^2 < .0001$), though it was likely that the cognitive engagement effect sizes differed by predictor type ($Q = 9.45, p = .009$). The model was incomplete, as there was unexplained variance between cognitive engagement point estimates with the same predictor type ($Q = 839, p = 0.0000$). The incremental changes in unexplained variance (I^2) for predictor type are presented in Table 47. The proportion of the unexplained variance that represented true variance, rather than error variance was 94.51% ($I^2 = 94.51, I = .97$). This suggests that the observed variance around subgroup means would shrink by approximately 3% if the error variance were removed.

An examination of the regression coefficients for the model suggested that both competence and relatedness predicted increased cognitive engagement point estimates when compared to autonomy. However, only the coefficient for competence ($\beta = .35, p = .002$) was statistically significant (see Table 47). Figure 19 shows a scatterplot of the regression model for SDT predictor type classification.

Table 46

| <i>Effect Sizes and Null Tests for Cognitive Engagement by Predictor Classification: SDT</i> | | | | | |
|--|----------|----------|------------|----------|----------|
| Predictor type | <i>n</i> | <i>g</i> | 95% CI | <i>Z</i> | <i>p</i> |
| Autonomy | 30 | .26 | [.15, .36] | 4.81 | .000 |
| Competence | 8 | .61 | [.41, .81] | 5.98 | .000 |
| Relatedness | 11 | .32 | [.15, .49] | 3.63 | .000 |

Table 47

Meta-regression Model for Cognitive Predictor Classification: SDT

| Predictor type | Variance | | Test of model | | | Regression | | |
|----------------------|----------|-------|---------------|------|------|------------|------|------|
| | T^2 | R^2 | Q | df | p | Coeff. | Z | p |
| Autonomy (intercept) | .06 | 0 | | | | .26 | 4.81 | .000 |
| Competence | .05 | .07 | 11.06 | 1 | .001 | .35 | 3.07 | .002 |
| Relatedness | .07 | 0 | 9.45 | 2 | .009 | .06 | .615 | .539 |

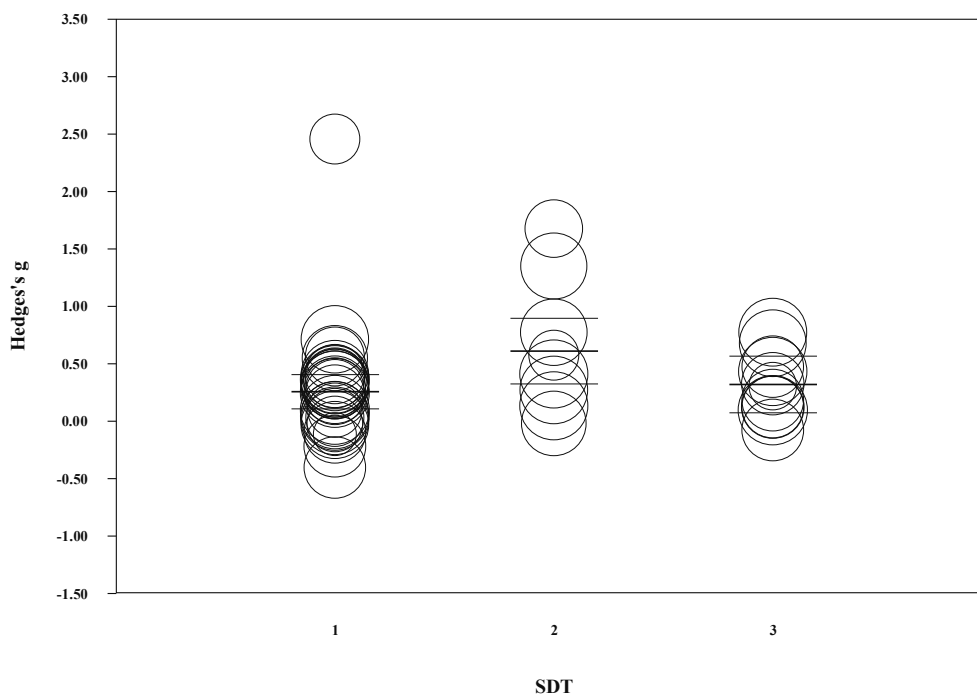


Figure 19. Regression of cognitive point estimates on predictor classification: SDT. The values on the x-axis show instructional methods (1), technology (2), class characteristics (3), and social characteristics (4).

Behavioral engagement predictors. Seven studies generated ten behavioral engagement point size estimates. When combining across predictors, the summary mean behavioral engagement effect size was $g = .23$, 95% CI [-.13, .59]. Seven behavioral engagement point estimates were positive, with two of those representing practically significant effects. The predictors for the two practically significant effects were a focus on investigation and universally-designed worksheets. Three behavioral point estimates

were negative, representing cognitive autonomy support ($g = -.016$, 95% CI [-.920, .888]), procedural and cognitive autonomy support ($g = -.14$, 95% CI[-.997, .72]), and a focus on science and society ($g = -.58$, 95% CI [-.63, -.54]). However, two negative effect sizes and three practically insignificant effect sizes had confidence intervals that spanned zero, suggesting the possibility that those predictors had no effect in the given studies. See Table 48 for the distribution of behavioral point size estimates and Figure 20 for a forest plot of behavioral engagement point estimates.

Table 48

Distribution of Behavioral Point Estimates by Predictor Classification

| Predictor classification | Practically significant effect sizes | | | | Practically insignificant effect sizes | | | |
|---------------------------|--------------------------------------|---------|-------------------------------|---------|--|---------|-------------------------|---------|
| | Moderate ($2.7 > g > 1.15$) | | Small ($1.15 > g > .41$) | | Small ($.41 > g \geq 0$) | | Negative ($g < 0$) | |
| | <i>n</i> | Percent | <i>n</i> | Percent | <i>n</i> | Percent | <i>n</i> | Percent |
| Type | | | | | | | | |
| Instructional method | 0 | 0% | 2 | 33.3% | 3 | 50% | 1 | 16.7% |
| Technology | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| Class characteristics | 0 | 0% | 0 | 0% | 2 | 50% | 2 | 50% |
| Social characteristics | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| Self-determination theory | | | | | | | | |
| Autonomy | 0 | 0% | 0 | 0% | 4 | 57.1% | 3 | 42.9% |
| Competence | 0 | 0% | 2 | 66.7% | 1 | 33.3% | 0 | 0% |
| Relatedness | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |

The investigator examined the distribution of behavioral engagement effect sizes for each predictor type. Instructional method was the only predictor type with practically significant effect sizes ($n = 2$; 33.3%). Class characteristics had the highest frequency of negative effect sizes ($n = 2$, 50%). None of the studies investigated the relationship of technology with behavioral engagement. Competence was the only self-determination theory predictor with practically significant behavioral engagement effect sizes ($n = 2$, 66.7%). Three of the four autonomy predictors yielded negative effect sizes. There were

not enough point estimates to enable the investigator to conduct a meta-regression of behavioral engagement predictors.

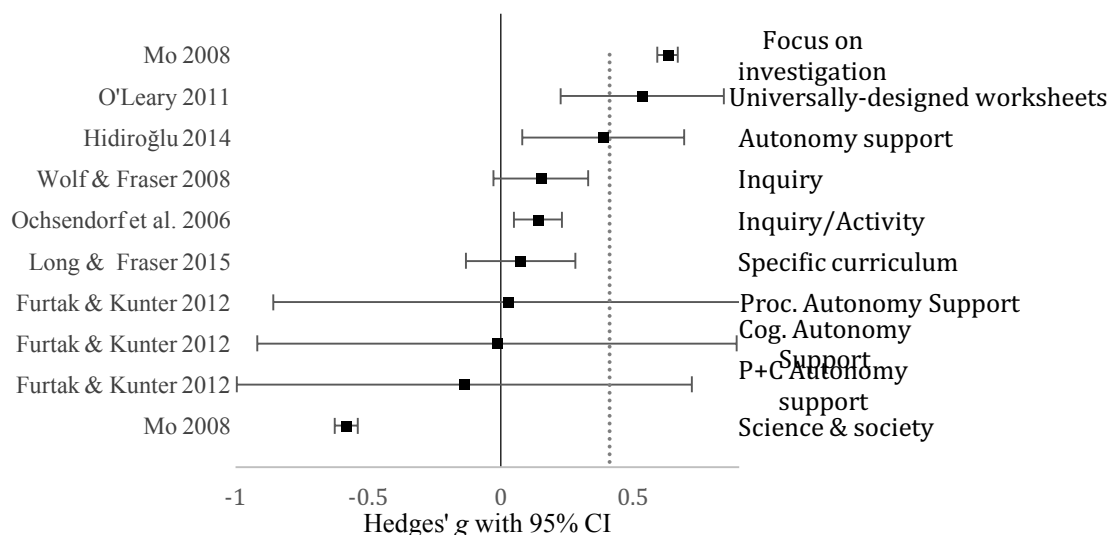


Figure 20. Forest plot of 10 behavioral engagement effect sizes. Dashed line represents a minimum practical effect size of $g = .41$.

The investigator examined the mean behavioral engagement effect size point estimate for each predictor type. The mean point estimates for instructional method and class characteristics were similar in magnitude ($g = .15$, 95% CI [-0.42, .73], and $g = .09$, 95% CI [-0.70, .88], respectively), and neither was statistically significant ($Z = .528$, $p = .60$, and $Z = .220$, $p = .83$, respectively). There were no technology or social characteristic predictors for behavioral engagement, and there were not enough estimates in each category to examine behavioral engagement predictor type through meta-regression. See Table 49 for a summary of effect sizes and null tests for behavior engagement by predictor type.

The investigator examined the mean behavioral engagement effect size point estimate for each SDT predictor type. The mean point estimate for autonomy showed the lower effect size ($g = -.004$, 95% CI [-0.40, .39]), and competence showed the higher

effect size ($g = .41$, 95% CI [-.13, .96]). Neither was statistically significant ($Z = -.018$, $p = .99$, and $Z = 1.49$, $p = .14$). There were no relatedness predictors for behavioral engagement, and there were not enough estimates in each category to examine behavioral engagement SDT predictor type through meta-regression. See Table 50 for a summary of effect sizes and null tests for behavior engagement by SDT predictor type.

Table 49

Effect Sizes and Null Tests for Behavioral Engagement by Predictor Classification: Type

| Predictor type | <i>n</i> | <i>g</i> | 95% CI | <i>Z</i> | <i>P</i> |
|-----------------------|----------|----------|-------------|----------|----------|
| Instructional method | 6 | .15 | [-.42, .73] | .528 | .597 |
| Class characteristics | 4 | .09 | [-.70, .88] | .220 | .826 |

Table 50

Effect Sizes and Null Tests for Behavioral Engagement by Predictor Classification: SDT

| SDT predictor type | <i>n</i> | <i>g</i> | 95% CI | <i>Z</i> | <i>P</i> |
|--------------------|----------|----------|-------------|----------|----------|
| Autonomy | 7 | -.004 | [-.40, .39] | -.018 | .985 |
| Competence | 3 | .41 | [-.13, .96] | 1.49 | .137 |

Summary. Class characteristics had the highest representation of practically significant affective engagement effect sizes in the predictor type classification, while competence had the highest representation for the self-determination theory predictor types. The mean affective point estimates for both of these SDT predictor classification categories exceeded the minimum practical effect size and were statistically significant. While the coefficients for class characteristics and competence predicted increases in engagement, though only the coefficient for competence was statistically significant in the SDT predictor type regression model.

The mean affective point estimates for these three remaining categories in the predictor type classification model—instructional method, technology, and social

characteristics—did not meet the minimum guidelines for a practically significant effect ($g > .41$). A meta-regression analysis of those categories revealed that technology and social characteristics predicted decreases in affective engagement in relation to instructional method. However, the coefficient for social characteristics characteristics was not statistically significant. Though the coefficient for technology was statistically significant, the mean point estimate was not. Thus, technology predicted decreases in affective engagement when compared to instructional methods.

One combination of statistically significant engagement moderators with predictor types slightly improved the fit of predictor type with affective engagement in meta-regression. A combination of publication status and predictor type improved the fit of predictor type with affective engagement ($Q = 12, p = .007$) when compared to predictor type alone ($Q = 11.74, p = .008$). In this model, publication status was the strongest predictor. When publication status was held constant, technology was the strongest predictor, and produced a statistically significant decrease in engagement. Combinations of other statistically significant engagement moderators with predictor type did not improve the fit of predictor type with affective engagement. See Table 38 for the meta-regression model of affective engagement predictor type and Table 39 for the meta-regression models of combined moderators and predictors.

The mean point estimates for the remaining self-determination theory predictor types—autonomy and relatedness—did not meet the minimum guidelines for a practically significant effect. In the meta-regression model, relatedness predicted an increase in engagement with respect to autonomy, though the coefficient was not statistically significant. This suggests that observed increases in affective engagement

due to relatedness predictors could be due to chance and not true effects. Thus, competence was the only self-determination theory predictor type that reliably predicted an increase in engagement, with respect to autonomy.

No combinations of other statistically significant engagement moderators with predictor type significantly improved the fit of predictor type with affective engagement. Though a combination of geographic location and SDT predictor type rendered SDT predictor type slightly more effective at predicting affective engagement ($Q = 5.62, p = .06$) than SDT predictor type alone ($Q = 5.59, p = .06$), neither model was statistically significant, suggesting that the predicted change in affective engagement could be due to chance alone. See Tables 38 and 41 for the meta-regression model of affective engagement predictor type and Table 39 and 42 for the meta-regression models of combined moderators and predictors.

Research Questions 6 & 7: Underrepresented Engagement Predictors and Types. The investigator considered research questions six and seven together in order to more comprehensively identify underrepresentation of engagement predictors and types. Studies that produced combinations of two or more engagement outcomes had low representation ($n = 15$) in the 158 point estimates, with combinations of all three engagement types producing the fewest point estimates ($n = 2$). Of the three main engagement types, behavioral engagement outcomes had the lowest number of point estimates ($n = 10$). Affective engagement outcomes were most numerous ($n = 84$). Thus, engagement types representing two or more engagement outcomes, as well as behavioral engagement, were underrepresented in self reports of early adolescents' science engagement (see Table 51).

Table 51

Distribution of Point Estimates by Engagement Type and Predictor

| Predictor classification | A | B | C | Two Types | Three Types | Total |
|---------------------------|----|----|----|-----------|-------------|-------|
| Type | | | | | | |
| Instructional method | 31 | 6 | 12 | 8 | 0 | 57 |
| Technology | 11 | 0 | 3 | 0 | 1 | 15 |
| Class characteristics | 26 | 4 | 24 | 5 | 1 | 60 |
| Social characteristics | 16 | 0 | 10 | 0 | 0 | 26 |
| Self-determination theory | | | | | | |
| Autonomy | 46 | 7 | 30 | 9 | 2 | 94 |
| Competence | 14 | 3 | 8 | 4 | 0 | 20 |
| Relatedness | 24 | 0 | 11 | 0 | 0 | 35 |
| Total | 84 | 10 | 49 | 13 | 2 | 158 |

An examination of predictor classification revealed that class characteristics and instructional methods were well represented in point estimates ($n = 60$ and $n = 57$, respectively). Technology and social characteristics produced fewer point estimates ($n = 15$, and $n = 26$, respectively). Autonomy was the self-determination theory predictor with the highest representation ($n = 94$), while competence and relatedness had lower representations ($n = 20$ and $n = 35$, respectively). Technology and competence were the most underrepresented predictor classifications in the point estimates.

Combinations of predictor and engagement types were analyzed in order to identify specific areas of underrepresentation. As engagement comprised of two or more outcomes and behavioral engagement were previously identified as areas with significant underrepresentation, they were excluded from further analysis by predictor classification. Though studies measuring affective engagement reflected instructional method and class characteristic predictors relatively equally ($n = 31$ and $n = 26$, respectively), the representation was not as equal for the predictors' relationship with cognitive

engagement ($n = 12$ and $n = 24$, respectively). Instructional methods appear to be underrepresented in cognitive engagement studies (see Table 51).

Publication Bias Analysis

The investigator analyzed possible publication bias through comparisons of mean effect sizes for published and unpublished studies, funnel plots, the fail-safe N , and Duval and Tweedie's trim and fill. A comparison of the mean effect size for unpublished studies ($g = .15$, 95% CI [.04, .25]) with the mean effect size for published studies ($g = .40$, 95% CI [.33, .46]) indicated that publication bias could be a concern. There were fewer point estimates from unpublished studies in the analysis ($n = 39$) than from published studies ($n = 119$). Further, the regression model indicated that while publication status explained a negligible portion of the variance in effect sizes ($R^2 < .0001$), the model was statistically significant. These results suggest a further examination of publication bias.

A visual examination of the funnel plot for all studies showed a high dispersion of point estimates not only for the small studies, but also for the studies with a larger sample size (see Figure 21). Though there was one small study on the right side of the graph that did not have correlates on the left side of the graph, it appeared as if there were also a cluster of larger-sized studies on the left side of the graph that did not have correlates on the right side. Thus, a visual inspection of the funnel plot was inconclusive.

An analysis of the classic fail-safe N revealed that 9197 studies would be required to bring the mean Hedges' g to a value that would no longer be statistically significant. Though this value appeared to indicate that publication bias was not a concern, there were two studies that produced particularly high point estimates (Akçay et al. 2010, Işık & Gücüm, 2013). Operating under the assumption that these point estimates might be

anomalous, the investigator recalculated the fail-safe N without those two studies. Without the two studies, the mean effect size dropped ($g = .33$, 95% CI[.27, .39]) with respect to the mean effect size with all studies included ($g = .37$, 95% CI[.30, .43]). The fail-safe N did decrease to 6,531 studies needed to bring the mean Hedges' g to a value that would no longer be statistically significant. As both values were exceedingly large in comparison to the number of studies in this analysis ($n = 79$), the fail-safe N indicates that publication bias was not a concern. As the classic fail-safe N is sometimes criticized for its focus on an effect size of zero and on statistical significance, Orwin's fail-safe N was also considered by the investigator. However, as the mean effect size for the studies was below the guidelines for a minimum practical effect size ($g = .41$), the investigator concluded that Orwin's fail-safe N was not an appropriate analysis of publication bias.

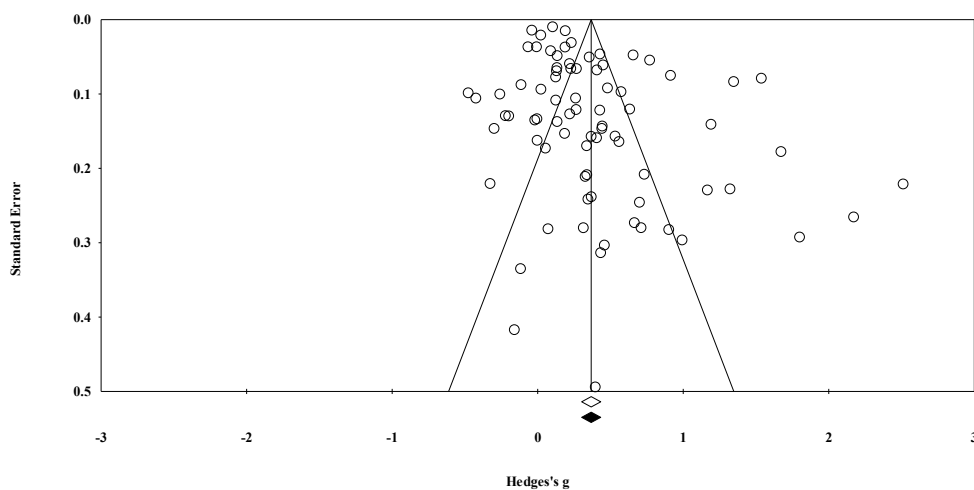


Figure 21. Funnel plot of 79 studies.

The investigator used Duval and Tweedie's trim and fill procedure to locate potential missing studies to the left and to the right of the mean effect size. No potential studies were found missing to the left of the mean, though there were studies found missing to the right of mean (see Figure 22 for the funnel plot with ten imputed studies included). The adjusted mean effect size increased ($g = .42$, 95% CI [.35,.48]) from the

mean effect size before the trim and fill procedure was applied ($g = .37$, 95% CI [.30, .42]). These results suggest that there was no publication bias in terms of failing to find studies with insignificant or negative effect sizes. In fact, though the investigator found more published than unpublished studies, more of the resulting point estimates were to the right of the mean effect size (see Figure 22).

Though the mean effect size for published studies was higher than that of unpublished studies, neither the fail-safe N nor the trim and fill procedure indicated a publication bias concern. As engagement was assessed as an ancillary outcome in many studies, and as the determination of the direction of an engagement effect size was largely an arbitrary one, the lack of bias toward positive, statistically significant results is not surprising. The bias in this particular study was in the overrepresentation of point estimates below the mean. When the trim and fill procedure was used to fill in ten studies, the mean engagement effect size increased to a practically significant level.

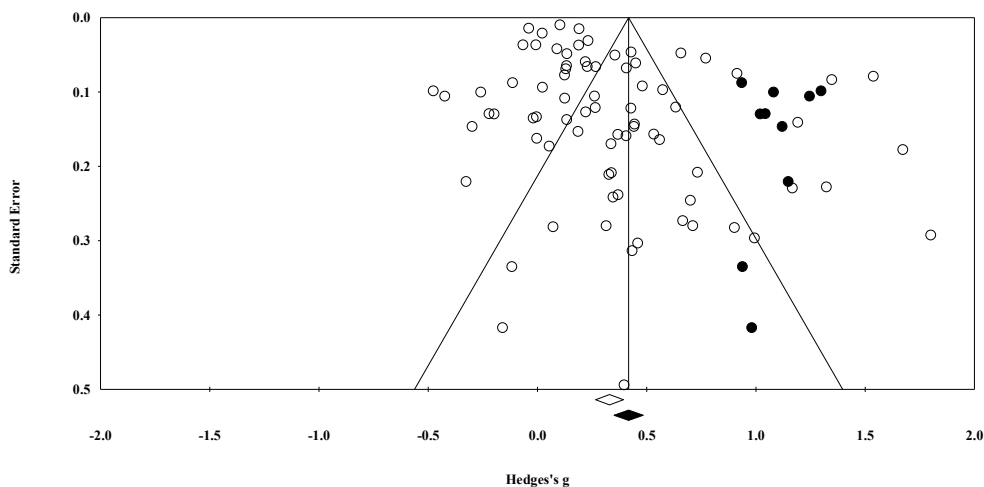


Figure 22. Funnel plot of 79 studies including ten imputed studies right of the mean

Chapter 5: Summary of Findings

The mean effect size generated from 79 studies and 158 point estimates in this meta analysis was $g = .37$, 95% CI [.30, .43]. When adjusted by Duval and Tweedie's trim and fill procedure, the mean increased slightly to $g = .42$, 95% CI [.35, .48]. The combination of a practically significant mean effect size and a robust number of point estimates suggests that this meta-analysis contained information about effective predictors of early adolescents' engagement with science.

Research Question 1: Moderators of Engagement

Statistically significant moderators. When analyzed via meta-regression, four moderators were found to be statistically significant with respect to science engagement: publication status, instrument reliability, school setting, and geographic location. A summary is provided for each moderator, including the statistical findings from the study, an analysis of possible reasons for group differences, and recommendations for considering the moderator in future research.

Publication status. The meta-regression model with this moderator was significant ($Q = 15.70$, $p = .001$), the coefficient for publication status in the regression model was positive and statistically significant ($\beta = .25$, $p = .00007$), and the mean effect size for published studies ($g = .40$, 95% CI [.33, .46]) was higher than the mean for unpublished studies ($g = .15$, 95% CI [.04, .25]). Similarly, regression models combining publication status with either predictor classification—type or SDT—were statistically significant ($Q = 30.97$, $p < .0001$, and $Q = 26.04$, $p < .0001$, respectively), but explained negligible variance in engagement. Regression models of affective engagement that combined publication status with either predictor type classification or SDT predictor

classification were likewise, statistically significant ($Q = 33.09, p < .0001$, and $Q = 17.41, p = .001$, respectively). The model of publication status and predictor type explained 14% of the variance in effect sizes ($R^2 = .14$), while the model for publication status and SDT predictor type explained negligible variance.

There are a number of explanations for the higher effect size observed in the published studies within this meta-analysis. One explanation is that the investigator failed to find relevant gray literature, and that the findings represented publication bias, rather than a true effect. Of the 21 unpublished studies included in the analysis, 18 of those represented dissertations or theses, with only three studies originating from conference proceedings or other sources. Though the mean effect size was higher for published studies, other findings suggest that publication bias was not a concern—the regression model explained a negligible amount of the variance in engagement point estimates ($R^2 < .0001$), the classic fail-safe N was 9,197 with outliers included and 6,531 without outliers, and Duval and Tweedie's trim and fill procedure actually found missing studies to the *right* of the mean effect size. Thus, the investigator concluded that negative bias due to underrepresentation of unpublished studies did not exist.

Another explanation for the statistical significance of publication status on effect size is that published studies differed from unpublished studies on other variables. Though a correlation matrix is often used to identify possible confounding variables, that approach was not appropriate in this study, as variables were categorical. Thus, the investigator compared the descriptive statistics for published and unpublished studies to determine if other moderators were over or underrepresented.

Geographic location. The meta-regression model with this moderator was significant ($Q = 11.28, p = .0007$), the coefficient for studies outside the United States was positive and statistically significant ($\beta = .18, p = .0008$), and the mean effect size for studies outside the United States ($g = .42, 95\% \text{ CI } [.04, .49]$) was higher than the mean for studies inside the United States ($g = .24, 95\% \text{ CI } [.16, .31]$). Similarly, regression models combining geographic location with either predictor classification—type or SDT—were statistically significant ($Q = 24.52, p = .0001$, and $Q = 27.98, p < .0001$, respectively), but explained negligible variance in engagement. Regression models of affective engagement that combined publication status with either predictor type classification or SDT predictor classification were likewise, statistically significant ($Q = 16.20, p = .0028$, and $Q = 11.70, p = .008$, respectively), but negligible in terms of affective engagement variance explained ($R^2 = .006$ and $R^2 < .0001$, respectively).

Though the amount of engagement variance explained by geographic location was negligible, the fact that the mean engagement effect size for studies within the United States was lower than that of studies outside the United States warrants further examination. While one conclusion is that approaches to science instruction in countries outside the United States are fundamentally more effective in terms of increasing student engagement, a more likely conclusion lies in a closer examination of the distribution of countries represented in studies outside the United States. Eighteen of the 44 studies from countries outside the United States originated from Turkey, where a K-8 school structure is common. In fact, each of 16 studies providing engagement point estimates from K-8 schools was also from Turkey. Though there were not enough point estimates for each school structure category to allow for meta-regression of this moderator, there were

sufficient point estimates for middle school and K-8 structures to allow for comparison. The mean science engagement effect size for point estimates from middle schools was $g = .16$, 95% CI [.06, .25), and from K-8 schools was $g = .42$, 95% CI [.31, .52]. These results suggest that the observed differences in science engagement due to geographic location might also be explained by school structure.

The observed difference in science engagement between studies from the United States and studies from other countries suggests that geographic location is a moderator to evaluate in future research. Geographic location is a complex identification, grouping different cultural approaches to schooling and teacher authority, school structures, and political pressures. As this meta-analysis was focused on a small subset of school engagement data within a certain grade range and content area, a more in-depth analysis of differences in educational variables between countries was not feasible. The recommendation is for future researchers to continue to report geographic location in addition to other educational variables that might differ by country. Future meta-analyses of engagement with broader inclusion criteria, such as a greater age range or multiple content areas, could then potentially elucidate reasons for engagement differences by country.

School setting. School setting was reported for fewer than half of the point estimates ($n = 75$) within the study. Of those 75 point estimates, 18 reflected a mix of school settings (e.g., rural and suburban), and thus could be analyzed no further with respect to the effect of school setting on science engagement. Of the remaining 58 point estimates, those from urban schools reflected the highest effect size ($g = .40$, 95% CI [.25, .54]), and rural schools reflected the lowest effect size ($g = -.11$, 95% CI [-.42, .21]).

Though the effect size for rural schools was not significant, the coefficient for rural schools was significant in the meta-regression ($\beta = -.50, p = .003$). This suggests that science engagement is expected to be lower in rural settings than in suburban or urban settings. However, an analysis of the lower mean science engagement effect size in rural schools was conducted with caution, as there were only five point estimates originating from schools in rural settings.

One possible explanation for the difference in science engagement by school setting is that national content standards and published curricula are not well-suited to the needs of students in rural schools, resulting in disengagement (Schafft & Jackson, 2011). One would expect that autonomy-supportive predictors emphasizing relevance would positively impact the mean science engagement in rural schools (Avery & Kassam, 2011). An analysis that included only autonomy-supportive predictors showed an increase in the mean engagement effect size for rural schools from $g = -.11, 95\% \text{ CI } [-.42, .21]$ to $g = -.016, 95\% \text{ CI } [-.38, .34]$. However, based on the limited number of point estimates for rural schools ($n = 5$), drawing conclusions from means is not advisable.

Instrument reliability. Instrument reliability was reported for all but six point estimates within the study. Point estimates from studies referencing an external instrument produced the highest mean effect size ($g = .60, 95\% \text{ CI } [.39, .81]$), followed closely by point estimates from studies referencing external instrument reliabilities ($g = .58, 95\% \text{ CI } [.37, .78]$). Though the effect sizes for both categories were statistically significant, the coefficients for each category within the regression model were not ($\beta = .33, p = .078$, and $\beta = .31, p = .099$, respectively). Point estimates from studies providing measures of internal reliability produced lower mean effect sizes, regardless of

whether the internal measure was less or greater than .70 ($g = .26$, 95% CI [.12, .39], and $g = .30$, 95% CI [.22, .37], respectively). Neither coefficient was statistically significant in the regression model ($\beta = -.01$, $p = .965$, and $\beta = .03$, $p = .841$, respectively).

The mean effect sizes for instrument reliability suggest that the use of vetted psychometric instruments yields higher mean effect sizes, and that studies reporting internal instrument reliabilities do not show increased engagement measures. However, a higher mean effect size is not necessarily more accurate than a lower mean effect size, so it is possible that while the use of well-researched psychometric instruments produced higher mean science engagement effect sizes, that higher effect size could be less accurate. Furthermore, there were fewer point estimates produced from studies referencing external instruments and external instrument reliabilities ($n = 11$ and $n = 14$, respectively), than those reporting internal reliabilities less than .70 or greater than .70 ($n = 28$, and $n = 99$). Thus, the observed effect could be a function of a greater spread of point estimates in the larger categories.

Though categories of point estimates from studies reporting internal instrument reliability yielded lower mean effect sizes than those only referencing external measures, there is value in including internal instrument reliabilities in further science engagement studies. One recommendation is to use the reported reliabilities as a continuous variable for meta-regression, rather than collapsing the reliability measures into categories. Such an analysis would have been problematic in this study, as the investigator often created composite variables of engagement from two or more instruments with different reliabilities. Another recommendation would be to distinguish between studies reporting internal instrument reliabilities from external, vetted instruments, and studies reporting

internal instrument reliabilities with no reference to an external instrument. Such an analysis would allow for a clearer determination of how the use of a vetted psychometric instrument affects mean engagement effect sizes.

Statistically nonsignificant moderators. Three moderators that were analyzed via meta-regression were found to be statistically nonsignificant with respect to science engagement: peer review status, study methodology, and repeat authors. A summary is provided for each moderator, including the statistical findings from the study, an analysis of possible reasons for the lack of statistically-significant group differences, and recommendations for considering the moderator in future research.

Peer-review status. While point estimates from studies originating from peer-reviewed journals yielded a higher mean effect size ($g = .36$, 95% CI [.30, .42]) than from journals that were not peer-reviewed ($g = .27$, 95% CI [.17, .37]), the meta-regression model was not significant ($Q = 2.38$, $p = .123$) This suggests that the increase in effect size due to a journal's peer review status could be due to chance, rather than a true effect.

The presence of three high outliers in the nonpeer-reviewed category could have skewed the mean, rendering the difference between peer-reviewed and nonpeer-reviewed point estimates nonsignificant (see Figure 4). Without the outliers, it is possible that the differences between mean effect sizes from peer-reviewed studies and nonpeer-reviewed studies would be significant. Because the investigator did not locate many studies that were published but not peer-reviewed ($n = 8$), the analysis of peer-review status and publications status yielded similar information. However, publication status was a statistically significant moderator, while peer review status was not. The regression model for publication status did not have the outliers that the peer review status model

did (see Figure 3). For further engagement research, locating more gray literature would provide information about whether peer-review status significantly moderates science engagement.

Study methodology. Studies utilizing a quasi-experimental design yielded the highest mean effect size ($g = .42$, 95% CI [.32, .53]), while studies utilizing an experimental design yielded the lowest mean effect size ($g = .17$, 95% CI[-.05, .39]). The mean for experimental studies, however, was not significant ($p = .127$). The regression model for study methodology was not statistically significant ($Q = 6.41$, $p = .093$), and no coefficients were statistically significant with respect to correlational studies.

Though the regression model for study methodology was not statistically significant, an examination of observed versus expected effect sizes was warranted. Despite statistical nonsignificance, a lower mean effect size was expected from experimental studies, as such a design attempts to remove effects of confounding variables on observed outcomes. The effect size for correlational designs ($g = .32$, 95% CI [.25, .40]), fell between that of quasi-experimental and experimental designs. Correlational studies would be expected to produce a high mean effect, as simple bivariate correlations do not parcel out the effect of confounding variables. This meta-analysis included not only simple bivariate correlations, but also coefficients from regression and path analysis studies. The effect sizes that result from regression and path analysis studies are smaller, as the coefficients from these studies tend to be smaller than simple bivariate correlations. Thus, a recommendation for future meta-analyses is to analyze potential differences between different correlational designs.

Repeat authors. The mean effect size for point estimates originating from studies with repeat authors was very similar to the mean from unique authors ($g = .32$, 95% CI [.17, .47], and $g = .33$, 95% CI [.28, .39], respectively). Thus, the regression model was not statistically significant ($Q = .03$, $p = .873$). The analysis of repeat authors was conducted to ensure that point estimates from studies sharing the same author did not skew the mean in either direction. The lack of effect of repeat authors was expected, and supported by a representative case of three studies with the same author (Vedder-Weiss & Fortus, 2011; Vedder-Weiss & Fortus, 2012; Vedder-Weiss & Fortus, 2013). The point estimates from these studies ranged from $g = -.57$ to $g = .41$, indicating that the use of similar measures, samples, methods, and/or theoretical lenses among the three studies did not produce three similar effect sizes that could skew the mean engagement effect size. Nevertheless, future researchers conducting meta-analyses are encouraged to consider the effect of repeat versus unique authors on the mean effect.

Moderators not analyzed via meta-regression. Four possible moderators of science engagement were not analyzed via meta-regression due to insufficient sample size within individual categories—school structure, school type, instrument validity and socioeconomic status. For each moderator, the investigator provides a summary, analysis, and recommendations for future science engagement researchers.

School structure. Excluding categories with fewer than 10 point estimates, point estimates from studies of K-8 school structures produced the highest mean effect ($g = .42$, 95% CI [.31, .52]), while point estimates from middle school structures produced the lowest mean effect ($g = .16$, 95% CI [.06, .25]). These results suggest that a K-8 school structure could be more effective at engaging early adolescents in science than a middle

school structure. This is concerning, particularly given that the middle school structure evolved as a solution to the perceived developmental mismatch between early adolescent needs and the junior high school structure. However, as all point estimates from K-8 studies also originated from Turkey, it is not possible to disentangle the effects of school structure from geographic location. The difference between the mean science engagement effect between the two school structures suggests that further research about the features of the school structure is warranted.

School type. Though the mean effect size of point estimates originating from studies of public schools was higher than the mean from private schools ($g = .32$, 95% CI [.25, .39], and $g = -.02$, 95% CI [-.31, .27], respectively), the small number of point estimates originating from studies of private school science engagement ($n = 6$) precluded a comparison of the two school types. Furthermore, the mean effect size for point estimates originating from studies including a mix of school types was $g = .36$, 95% CI [.16, .55]. These results suggest that the mean effect from public schools and the mean effect from a combination of public and private schools were similar. Though these results seem contradictory, the most likely reason for the uninterpretable results was that there were not enough point estimates to provide a valid comparison of public vs. private school science engagement.

Instrument validity. Point estimates from studies that referenced an external instrument or achieved face validity by the investigator showed the highest means ($g = .48$, 95% CI [.39, .56], and $g = .42$, 95% CI [.16, .69], respectively). Conversely, point estimates from studies that reference external instrument validities or achieved face validity as assessed by the study showed the lowest means ($g = .12$, 95% CI [-.03, .26],

and $g = .03$, 95% CI [-.13, .18], respectively). Though it may appear surprising that instruments achieving face validity as assessed by the investigator would show a high mean engagement effect, instruments without robust validity assessments could show either exceedingly high or low results. What is more contradictory about the results for this analysis is that studies with face validity assessed by the investigator were high, while studies with face validity assessed by the study authors were lower. Likewise, it is contradictory that studies referencing external instruments produced high mean effects while studies referencing external instrument validity produced lower mean effects.

In order to explain seemingly disparate results concerning the potential moderating effect of instrument validity on engagement outcomes, it is necessary to examine what measures of instrument validity in this study communicate about those instruments. A wide variety of instruments were represented by the included studies, and while many of those instruments explicitly assessed engagement, some did not. For example, the MSLQ was used by researchers in many of the studies to provide a measure of cognitive engagement. The MSLQ, however, was originally designed as an instrument to assess motivation by way of strategy use, metacognition, and task value. Though the MSLQ has been shown to have instrument validity in other studies, that instrument validity communicates the MSLQ is a valid measure of *motivation*. It does not necessarily follow that the MSLQ is a valid measure of engagement. As a wide variety of indicators are accepted as measures of engagement, a wide variety of instruments and constructs were included in this study. Thus, while an instrument may show validity, it may or may not show validity with the engagement construct. This is the likely

explanation for the contradictory moderating effects of instrument validity on engagement.

Recommendations concerning future research of the potential moderating effects of instrument validity on science engagement include calls to increase the validity of the engagement construct in existing and future psychometric instruments. For existing measures of engagement, such recommendations demand concurrent and predictive validity measures with respect to other explicit engagement measures. Likewise, for existing measures of constructs that are considered proxies for engagement, assessments of concurrent and predictive validity with vetted engagement instruments is warranted. Though clarity about the engagement construct may seem a prerequisite for such psychometric evaluation, it is through the psychometric evaluation that clarifying discussions about the construct may be grounded.

Socioeconomic status. The vast majority of point estimates ($n = 101$) originated from studies that did not report the socioeconomic status of students in the sample. Because of this, there were not enough point estimates in each SES category to allow for a valid analysis. Though the mean effect size from studies of students with average SES was highest ($g = .49$, 95% CI [.21, .77]), followed by low SES ($g = .25$, 95% CI [- .001, .50]), and then high SES ($g = .02$, 95% CI [- .15, .19]), there were minimal point estimates in each category ($n = 7$, $n = 9$, and $n = 19$, respectively). Thus, no conclusions could be drawn about potential moderating effects of SES on early adolescents' science engagement.

Research Question 2 & 3: Commonalities in Engagement Predictors

In order to determine commonalities in science engagement predictors, the investigator analyzed 51 practically significant point estimates and conducted meta-regression analyses of engagement predictor classifications. Of the practically significant point estimates, instructional methods had the highest representation of the predictor type classifications ($n = 24$, 46%). Instructional methods also had the highest mean effect size ($g = .42$, 95% CI [.34, .51]), and the only positive coefficient in the regression model of predictor type classification ($\beta = .43$, $p < .0001$). The mean for instructional methods was the only one that reflected a minimum practical effect size. Thus, instructional methods are better predictors of engagement than technology, class characteristics, or social characteristics.

However, the results from the other three predictor type classifications are perhaps more useful than the result showing instructional methods to be the best predictor of early adolescents' science engagement. Though technology, class characteristics, and social characteristics all generated positive mean effect sizes, they all predicted decreases in science engagement in the regression model, with respect to instructional methods. Technology predicted the greatest decreases in engagement ($\beta = -.32$, $p = .0006$) when compared to instructional methods, and had the highest representation of negative point estimates of all of the predictors ($n = 5$, 33%). Class characteristics and social characteristics predicted smaller decreases ($\beta = -.09$, $p = .149$, and $\beta = -.18$, $p = .026$, respectively), and the predicted decrease for class characteristics was not statistically significant.

Though causality was not established by this study, these results suggest that interventions focusing on technology, class characteristics, and social characteristics could be less effective at increasing science engagement than interventions focusing on instructional methods. The fact that technology predictors showed the lowest mean effect size and predicted the greatest decrease in engagement with respect to instructional methods runs counter to rationales given for technology integration in science classrooms—authenticity with the scientific discipline, equity, novelty, and autonomy support (Guillén-Nieto & Aleson-Carbonell, 2012; Zucker, Tinker, Staudt, Mansfield, & Metcalf, 2008). A common rationale given for the incorporation of technology games into the curriculum is that students receive more immediate feedback on their progress in a gaming situation (Garris, Ahlers, & Driskell, 2002). One explanation for the disconnect between rationales for technology integration and the relationship of technology with engagement in this study is that technology is one of many conduits through which authenticity, equity, novelty, autonomy, and feedback can be enhanced. The mere integration of technology does not ensure that any of the aforementioned desired qualities are implemented, or implemented effectively.

The predicted decrease in engagement from social characteristics when compared to instructional methods is also contradictory to educational research. Examples of social characteristics within this study included perceptions of teacher characteristics—approachability, social support, and strictness—as well as more holistic social characteristics, such as perceptions of belonging, cooperative learning, and respect for differences. Research supports the efficacy of social interventions such as cooperative learning (Slavin, Hurley, & Chamberlain, 2003). Further, extensive research on the

middle school transition suggests that students report their teachers to be more controlling and less nurturing, and also that social comparison and competition increases (Eccles & Midgley, 1989; Eccles et al., 1993; Lepper et al., 2005; Midgley et al., 1989; Roeser & Eccles, 1998). Thus, perceptions of social characteristics should predict students' engagement.

There are a number of possible explanations for the incongruity between the observed relationship of social characteristics with engagement in this study and other educational research findings. One is that the vast majority of social characteristics point estimates ($n = 22$) reflected correlations between perceptions of those characteristics and engagement; only four of the point estimates in this category involved an intervention. Thus, it is possible that a student could report being engaged, while also reporting that his or her teacher was not approachable—in a correlational study there is no reason for one to explain the other. While the social characteristics category reflected 26 point estimates, they originated from only ten studies. In fact, one study produced 10 of the 26 point estimates (Smart, 2014). Additionally, six of the 26 point estimates reflected predictors that would be expected to have a negative relationship with engagement: perceptions of the teacher as admonishing, strict, or dissatisfied. When considering these different explanations in concert, a more likely explanation for the incongruity between observed and expected relationships between social characteristics and students' science engagement is that there were not enough point estimates to draw a definitive conclusion.

The class characteristics category, which predicted a statistically nonsignificant decrease in engagement with respect to instructional methods, was comprised of a variety of predictors, such as relevance, critical voice, autonomy support, and democratic versus

traditional environments. The mere variety of predictors in this category could explain why there is no definitive effect of class characteristics on engagement. The duration of more abstract interventions such as autonomy support could impact their efficacy, with students experiencing some discord with the intervention at early stages, and becoming more comfortable and benefitting from such interventions over time. Alternately, the novelty of such interventions could cause positive initial effects, with decreases over time as the intervention becomes more routine. In studies with multiple measures of engagement over time, the investigator selected the most proximal measure of engagement to the intervention. Thus, it is possible that longer-duration measures of the relationship between class characteristics and science engagement could show higher or lower point estimates than the more proximal measures within this study. Despite a statistically nonsignificant coefficient for class characteristics, the mean effect for the predictor was $g = .34$, 95% CI [.25, .42], which, though lower than the mean for instructional methods, is still just below the threshold for a practically modest effect.

Meta-regression models that combined predictor type classification with either geographic location or school setting rendered the decrease in engagement due to class characteristics statistically significant ($\beta = -.191$, $p = .006$, and $\beta = -.143$, $p = .035$, respectively). Interestingly, these models rendered the decrease in engagement due to social characteristics statistically nonsignificant ($\beta = -.16$, $p = .053$, and $\beta = -.15$, $p = .101$). So when geographic location or school setting was held constant in the models, the decrease in science engagement due to class characteristics became statistically significant, while the decrease in science engagement due to social characteristics became statistically nonsignificant. This suggests that the observed decrease in science

engagement due to class characteristics is likely not due to chance when taking into account geographic location or school setting, but the observed decrease in science engagement due to social characteristics may be due to chance, when taking geographic location or school setting into account.

To further complicate the analysis of predictor type classification, many instructional methods can incorporate aspects of technology, class characteristics, or social characteristics. For example, project-based learning (instructional method) can include cooperative learning (social characteristic), and/or relevance (class characteristic) components. Thus, while one can conclude that a broad focus on technology, class characteristics, and social characteristics predicts decreases in science engagement, one cannot conclude that instructional methods incorporating these other components would be less effective than instructional methods that do not.

Because the instructional methods category is a broad one—encompassing varied predictors such as project-based learning, graphic organizers, and whole brain teaching—further analysis is needed to fully answer the research question about commonalities in practically significant science engagement predictors. An analysis of self-determination theory predictor type revealed that competence point estimates yielded the greatest mean effect ($g = .56$, 95% CI [.44, .69]), and autonomy point estimates yielded the lowest mean effect ($g = .26$, 95% CI [.19, .33]). Both competence and relatedness predicted increases in relation to autonomy, but only the increase in competence was statistically significant ($\beta = .31$, $p < .0001$). Regression models combining SDT predictor type with other significant moderators did not change the predicted increases due to competence or relatedness, nor did they change the statistical significance of the SDT coefficients.

Despite the research on the middle school transition that shows students report negative perceptions of their teachers as more controlling, and their classrooms as more heavily focused on social comparison (Eccles & Midgley, 1989; Eccles et al., 1993; Lepper et al., 2005; Midgley et al., 1989; Roeser & Eccles, 1998), competence was the best predictor of increased science engagement over autonomy and relatedness. This finding is not entirely unexpected, as another defining characteristic of the middle school transition is an increased focus on academic content standards (Ryan & Patrick, 2001). Such a finding could suggest that student engagement benefits more from explicit attention to competence as science content becomes more complex during middle school than engagement benefits from attention to autonomy or relatedness concerns. Though competence was the only SDT predictor that achieved a minimum practical mean effect, the mean effect sizes for autonomy and relatedness were positive and statistically significant ($g = .26$, 95% CI [.19, .33], and $g = .34$, 95% CI [.22, .46], respectively).

Though instructional methods and competence produced the highest mean effect sizes, both predictor type and SDT predictor type regression models left a large amount of engagement variance unexplained. This finding parallels research that suggests only a small portion of engagement variance was explained by teacher and class-level variables, with the majority of variance occurring between and within individuals (Uekawa et al., 2007). Though this meta-analysis examined classroom and task level science engagement predictors, it did not capture between individual and within individual variance.

In summary, instructional methods and competence were the two predictors with the highest mean engagement effect sizes, and both produced statistically significant

coefficients in the regression models. Although technology, class characteristics, and social characteristics predicted decreases in engagement with respect to instructional methods, only technology and social characteristics predicted a statistically significant decrease. Thus, class characteristics and instructional methods were similar in their ability to predict engagement. When statistically significant moderators were added into the regression models, only technology remained a statistically significant, negative predictor of engagement with respect to instructional methods. The addition of statistically significant moderators did not fundamentally change the coefficients or regression model for SDT predictor type.

Research Question 4 & 5: Commonalities in Affective Engagement Predictors

Affective engagement. In order to determine commonalities in affective science engagement predictors, the investigator analyzed 28 practically significant point estimates and conducted meta-regressions of affective engagement predictor classifications. Of the practically significant point estimates, class characteristics had the highest representation of the predictor type classifications ($n = 11$, 42.3%). Class characteristics also had the highest mean effect size ($g = .42$, 95% CI [.30, .53]), though the mean effect for instructional method was similar ($g = .38$, 95% CI [.28, .48]). This similarity in mean effects for the two categories explains why the model for predictor type classification was significant ($Q = 11.74$, $p = .008$), though the coefficient for class characteristics was small and not statistically significant ($\beta = .04$, $p = .662$). Thus, class characteristics and instructional methods were relatively equivalent predictors of early adolescents' affective engagement in science.

The affective engagement results did not differ substantially from the holistic engagement results. Though class characteristics showed a higher mean effect for affective engagement, and instructional methods showed a higher mean effect for holistic engagement, the results from the regression models suggest that the differential predictive power of the two categories could be due to chance for both affective and holistic engagement. The coefficients for class characteristics were not statistically significant in the holistic engagement regression model ($\beta = -.09, p = .149$) or the affective engagement model ($\beta = .04, p = .622$), though the direction of the effect was different. These results suggest that while the mean effect of class characteristics was higher than that of instructional methods, the power of each category to predict differences in affective engagement was minimal. Both class characteristic and instructional method predictors yielded practically significant or nearly practically significant affective engagement effects.

Similarly, the self-determination theory predictors yielded similar results for affective and holistic engagement. Competence predictors produced the highest mean effect ($g = .56, 95\% \text{ CI } [.44, .69]$), and were the only statistically significant affective engagement predictor when compared to autonomy ($\beta = .26, p = .019$). Though it is somewhat unexpected that competence would yield higher affective engagement effects than either autonomy or relatedness, there are two possible explanations for this finding. One is that over half of the point estimates in the study represented affective engagement ($n = 84$), and this high representation of affective engagement skewed the overall results. Alternatively, though it may seem intuitive to increase early adolescent's affective engagement through predictors that most directly parallel dimensions of affect, students'

success and perceptions of competence more effectively generate positive emotions toward science. The latter explanation is supported by self-efficacy research that suggests mastery experiences are most effective at increasing learners' feelings about their ability to complete a task (Bandura, 1977).

The ability of competence to predict affective engagement became statistically nonsignificant when publication status or instrument reliability was included in the regression models. This suggests that some of the difference between the effects of autonomy and competence predictors was explained by publication status and instrument reliability. This change in statistical significance of competence predictors on affective engagement was not seen in the holistic engagement regression models that included moderators. This suggests that studies reporting affective engagement differ from studies reporting holistic or other types of engagement in terms of publication and instrument reliability.

Cognitive engagement. In order to determine commonalities in cognitive science engagement predictors, the investigator analyzed 12 practically significant point estimates and conducted meta-regressions of cognitive engagement predictor classifications. Of the practically significant point estimates, instructional methods had the highest representation of the predictor type classifications ($n = 5$, 41.7%), and the highest mean effect size ($g = .49$, 95% CI [.33, .66]). Only the predicted decrease in cognitive engagement due to social characteristics was statistically significant ($\beta = -.27$, $p = .031$) with respect to instructional methods.

These results are similar to both the holistic and affective engagement outcomes for predictor type. Though technology yielded a higher mean effect for cognitive

engagement than for the other two types of engagement, the category's ability to predict changes in cognitive engagement was not statistically significant ($\beta = -.26, p = .203$). Only the coefficient for social characteristics was statistically significant, predicting a decrease in engagement with respect to instructional methods ($\beta = -.27, p = .031$). These results suggest that while instructional methods predicted the highest cognitive engagement effects; technology and class characteristics were comparable predictors, as their coefficients were not statistically significant with respect to instructional methods. However, this result should be interpreted with caution, as there were only three technology point estimates reflecting cognitive engagement. Further, there were two potential outliers for cognitive engagement in the instructional methods category that could render the differences between instructional method and the other predictor type categories greater than they actually were. The cognitive engagement regression model for predictor type provides tentative confirmation that cognitive engagement predictors do not differ substantially from predictors of affective or holistic engagement.

Similarly, the cognitive engagement results for SDT predictor type were comparable to those for holistic and affective engagement. Competence produced the highest mean engagement effect size ($g = .61, 95\% \text{ CI } [.41, .81]$), and its coefficient was statistically significant in the regression model ($\beta = .35, p = .002$). This result is expected, as predictors that explicitly address competence would intuitively be expected to have a larger effect on cognitive engagement than autonomy or relatedness predictors would. Again, this regression model is interpreted with caution, as competence only produced eight cognitive engagement point estimates. This model provides tentative confirmation

that SDT predictors do not differ substantially from predictors of affective or holistic engagement.

In summary, instructional methods and competence yielded the highest mean cognitive effect sizes. Though the regression models were interpreted with caution due to some categories having fewer than ten point estimates, the results indicate that commonalities among predictors of cognitive engagement do not differ substantially from predictors of affective or holistic engagement. One small difference is that the mean effect sizes for predictor types did not vary as much for cognitive engagement as for the other types of engagement. It is possible that this difference is due to the smaller number of cognitive point estimates in general, and the smaller number of point estimates for each predictor type category.

Behavioral engagement. There were only 10 point estimates for behavioral engagement that originated from seven studies. Five of 10 point estimates had confidence intervals that spanned zero, suggesting that the effect of that predictor could be zero. The top four predictors of behavioral engagement that did not have confidence intervals spanning zero were a focus on investigation, universally-designed worksheets, autonomy support, and inquiry. As there were no point estimates for technology or social characteristics, nor for relatedness, a meta-regression was not run. Thus, it is difficult to conclude with any certainty what the commonalities were for predictors of behavioral engagement, though it is intuitive to suggest that interventions that allow for more students to participate, such as inquiry, or that would allow for more students to access the material, such as universally-designed worksheets, would increase behavioral engagement.

Summary. Instructional methods and competence were the two predictors with the highest effect sizes for engagement as a whole, as well as for affective and cognitive engagement. Class characteristic predictors also showed moderate engagement effects, with coefficients that did not differ substantially from instructional method predictors in the regression models. The fact that predictor commonalities did not differ by engagement type runs seems to suggest a unidimensionality to engagement that runs counter to the three-faceted model in the research literature. However, as previously discussed, variance in engagement was previously found to be explained by between and within-person variables, more than it was explained by classroom and teacher-level variables (Uekawa et al., 2007). It follows that variance in engagement types is likely also explained more by between and within-person variables. Further, the person-centered analysis by Lau and Roeser (2008) intimated that some types of students would benefit from affective interventions to increase engagement while others would benefit more from cognitive interventions. As a large amount of variance was left unexplained in the regression models for each engagement type, it is possible that the similarity in predictors for different engagement types observed in this study could be explained by the pooling of engagement effects for multiple student types.

Research Questions 6 & 7: Underrepresented Engagement Predictors and Types

Engagement types. Despite recommendations to consider the three types of engagement holistically (Fredricks et al., 2004), combinations of two or more engagement types were underrepresented in this study. Though a review of K-12 appropriate engagement instruments identified five student self-report instruments that included measures of all three engagement types, and five with measures of two

engagement types, those instruments were not well-represented in this study. In some studies that used multidimensional instruments, the authors utilized only a portion of the instrument (e.g., Little, 2015). Another reason for the underrepresentation of multidimensional engagement measures is that many of the included studies did not purport to measure engagement, but rather an indicator that has been considered an acceptable measure of a particular facet of engagement (see Table 3). For example, studies that assessed students' self-report mastery goal orientations were included in this meta-analysis as measures of students' cognitive engagement. While broad inclusion criteria allowed for a robust number of point estimates for analysis, those point estimates tended to be unidimensional in nature.

Affective engagement was well-represented in the included studies; 84 affective engagement point estimates resulted from 56 studies. This type of engagement was well represented due to ongoing interest in students' attitudes toward science. Though this field of research is extensive, there is less research connecting students' attitudes toward science to classroom and task-level variables. Affective point estimates in this study included measures of attitudes, interest, situational interest, enjoyment, and valuation. The affective point estimates included in the study far outnumbered the cognitive and behavioral point estimates.

Cognitive engagement was relatively well-represented, with 31 studies generating 49 cognitive engagement point estimates. The vast majority of the cognitive engagement effect sizes originated from studies assessing goal orientation, self-regulated learning, or strategy use, vis-à-vis the MSLQ, AGQ, or PALS. Agentic engagement was underrepresented as an indicator of cognitive engagement. Though researchers have

recently proposed agentic engagement as a fourth facet of engagement, it was considered to be an indicator of cognitive engagement in this analysis (Reeve & Tseng, 2011).

Though there were a fair number of cognitive point estimates in the study, the proportional representation of cognitive to affective engagement seems negatively skewed. Science engagement research is certainly concerned with how students *feel* about science and science activities, but also with how to encourage students to *do* science at deep cognitive levels. Some research suggests that while science activities increase students' affect toward science, that increased affect does not correlate to increased cognitive engagement or achievement (Finn & Zimmer, 2012). Other research suggested that affective engagement is either an antecedent or an outcome of deeper levels of engagement (Eccles & Wang, 2012; Pekrun & Linnenbrink-Garcia, 2013; Reschly & Christenson, 2006). These findings suggested at least an equivalent focus on affective and cognitive engagement. Thus, the representation of cognitive engagement within this analysis was fair, but not proportional to affective engagement.

Behavioral engagement was underrepresented in this study, yielding only 10 point estimates from seven studies. Because this meta-analysis focused on student self-report, a number of studies assessing students' behavioral engagement through external observation were excluded. The dearth of student self reports of behavioral science engagement compared to external observations of behavioral science engagement suggest an analysis of which assessment method is more valid and useful to engagement research. Some indicators of behavioral engagement are more directly assessed through external observation, such as time on task, compliance with teacher requests, and completion. Recall errors are likely to result from asking a student to self-report on such variables.

However, some indicators are better assessed through self-report, such as effort and participation. Thus, more self-report measures of behavioral engagement are needed, but external observation can be used as an effective method of triangulation to ensure validity.

Engagement predictors. Of predictor type classification, instructional methods and class characteristics were well represented in point estimates ($n = 57$ and $n = 60$, respectively). Technology was the most underrepresented predictor type ($n = 15$). The technology point estimates were unequally distributed among the engagement types, with 11 of the 15 producing affective engagement point estimates. Thus, while technology was underrepresented as a whole within the study, it was more underrepresented in cognitive and behavioral engagement. However, this discrepancy in representation between affective and other engagement types was found in the other predictor type categories as well. Further, a comparison of the proportional representation of instructional method with class characteristics indicated that instructional method is underrepresented for cognitive engagement. A possible explanation for this is that many studies of instructional methods assess cognitive engagement as an ancillary outcome while focusing on achievement as the primary outcome.

Of the self-determination theory predictor types, autonomy was overrepresented with respect to competence and relatedness. One reason for this difference is that autonomy is a broad category, comprised of a number of indicators—relevance, choice, perceived control, negotiation, voice. Another reason is that predictors that primarily reflect autonomy could also include elements of competence or relatedness. For example, project-based learning was categorized as primarily an autonomy predictor, as relevance is a key feature of PBL. However, some project-based learning implementations involve

group work (relevance) or scaffolded components (competence). In identifying a primary SDT categorization, the investigator necessarily collapsed additional information into the primary category.

Recommendations

Though the three-faceted model of engagement proposed by Fredricks et al. (2004) has permeated the research literature, clarity about the construct is still evolving. Similarly, clarity in the assessment of engagement via psychometric instruments is nascent. Engagement researchers and the investigator in this study utilized a number of existing measures of related and overlapping constructs that yielded information congruent with recommendations for indicators of each engagement type (see Table 3), as well as measures that were reviewed and determined to be congruent with engagement or a combination of its facets (Fredricks et al., 2004; Fredricks et al., 2011; Skinner & Pitzer, 2012). Despite these attempts to enhance the validity of engagement research or compilations of engagement research, existing measures of engagement and related constructs varied in terms of comprehensiveness and intended grain size. Few instruments assessed all three facets of engagement, and even fewer included subscale measures of the three facets. Many instruments assessed school or classroom-level engagement, with little attention to finer-grained, task-level variables.

One recommendation that emerged from this study is to design more comprehensive engagement instruments that assess all three facets of engagement. Such instruments would afford researchers the ability to distinguish differential effects of predictors on the facets of engagement in a more systematic fashion. For example, instead of utilizing the MSLQ as a measure of cognitive engagement and the TOSRA as a

measure of affective engagement, a researcher could utilize a comprehensive engagement instrument with cognitive and affective components that were determined to be distinct through factor analyses. Though the MSLQ was identified as an appropriate measure of cognitive engagement, it included not only cognitive engagement components such as strategy use and self-regulated learning, but also task value, which has been suggested as an affective engagement indicator (refs). The creation of a comprehensive instrument would afford clarification about what indicators map most closely to each facet of engagement, as well as affording a more systematic analysis of the differential effects of predictors on each engagement type.

More comprehensive instruments should then be utilized to examine trajectories of engagement for individual students. This recommendation is supported by the finding that within or between person variables explained more engagement variance than classroom or teacher-level variables (Lau & Roeser, 2008; Uekawa et al., 2007). The Experience Sampling Method (ESM) is a promising technique to examine these changes in student engagement. When self-reports of engagement through ESM are matched to the characteristics of tasks and activities occurring at the time of the self-reports, researchers can analyze nuanced changes in engagement for individuals. The Uekawa et al. (2007) study provided an exemplar of how students' self-reports of engagement, gathered through ESM, can be matched with temporally-immediate reports of class activities to produce a complete picture of students' changing engagement and possible antecedents of those changes.

Another benefit to assessing engagement longitudinally through ESM is the identification of possible engagement trajectories. Some research suggests that affective

engagement is a precursor or regulator of other types of engagement (Ajzen, 1991; Ajzen & Fishbein, 1977; Eccles & Wang, 2012; Pekrun & Linnenbrink-Garcia, 2013; Schank, 1979). Other researchers suggest that cognitive and affective engagement predict behavioral changes (Reschly & Christenson, 2006). The use of ESM could afford the kind of detailed observation necessary to elucidate temporal changes and trajectories of engagement changes. Such information could inform decisions of which engagement types are appropriate targets in initial engagement interventions, compared with interventions that would better be targeted later in the sequence.

Another recommendation is to purposefully sample disengaged students in order to determine what practices *change* engagement for those students. In other words, though the results from this study may indicate that certain predictors have a more positive relationship with engagement than others, the study cannot inform conclusions about which predictors show the largest changes in engagement, nor can the study inform conclusions about which predictors show the largest changes in engagement for specific groups. As an implicit purpose of this study was to identify practices that engage or re-engage students with science coursework, an analysis of predictors that improve engagement for disengaged students is critical to inform best engagement practices in science classrooms.

The results from this meta-analysis suggest the inclusion of certain predictors in future studies. Categories that predicted the largest mean engagement effects included instructional methods, class characteristics, and competence. The finding that instructional methods best predict science engagement bears further examination. Do some instructional methods work better for disengaged students? Does the order in

which instructional method interventions are implemented matter? What types of instructional methods work best? Similar questions emerge for class characteristics and competence predictors. Further analyses of effective engagement predictors will also be enhanced by the aforementioned use of longitudinal methods and purposeful sampling.

Though effective predictors of early adolescents' science engagement were identified in this study, it would be premature to eliminate less effective predictor categories from consideration in future science engagement studies. For example, though technology predicted a statistically significant decrease in engagement, the mean effect of technology on each engagement type was positive, and there were limited numbers of technology point estimates. Thus, the results of this study might inform hypotheses about expected results in future studies, but would not be cause for exclusion of particular predictors. Simple models with only predictor type or predictor SDT type did not predict a great deal of engagement variance, and there were also four statistically significant moderators of engagement—publication status, instrument reliability, school setting, and geographic location. These variables deserve further elucidation before definitive conclusions about predictors worthy of inclusion in future studies can be made.

Conclusion

Early adolescence is a time period marked by declining engagement with science coursework. Stage environment fit theory and self-determination theory were two theoretical lenses utilized by the investigator to both predict and interpret the results of this study. Thus, the investigator hypothesized that the observed engagement decline was due to a developmental mismatch between middle school science classrooms and the needs of early adolescents. Though much of the literature concerning early adolescents'

perceptions about the middle school transition suggested that autonomy and relatedness are the most prevalent unmet needs, the results of this study suggest that academic predictors, such as instructional methods and competence, were more effective predictors of science engagement.

Though these results are somewhat surprising, they do not fundamentally contradict interpretations through the lens of SEF and SDT. Cognitive mismatches between science classroom tasks and the changing early adolescent brain were not a neglected component of students' self-reports of their middle school classrooms in general, and their science classes in particular (Anderman & Mueller, 2010; Mahatmya et al., 2012; Piaget, 1972; Ryan & Patrick, 2001). Though students become more capable of abstract thought and considering multiple perspectives during early adolescence, middle school students reported declines in the cognitive demand of classroom tasks during this time (Walberg et al., 1973; Uekawa et al. 2007). Deci and Ryan's (2002) suggestion that developmental characteristics can change the importance of one self-determination theory need relative to another is particularly relevant to this analysis. Though research suggested the developmental mismatch between early adolescents and their middle school classrooms was greatest in the areas of autonomy and relatedness, this study found that competence predictors yielded the highest mean effects. This discrepancy could suggest that issues in science engagement are different than issues in engagement in other content areas.

Alternatively, this discrepancy could suggest that an unintuitive solution is most effective—though autonomy and relatedness may be the most prevalent unmet needs of early adolescents in their science classrooms, competence predictors could be most

effective at meeting those autonomy and relatedness needs. Because the investigator included the most proximal measure of engagement to the predictor or intervention, the recommendation to establish trajectories of engagement in future research is particularly important. The relative effectiveness of competence predictors could be an indicator that competence predictors are more effective at increasing engagement in the early stages of engagement interventions, or that competence predictors are more effective overall. Without more longitudinal studies in the research literature, it is not possible to make this determination.

Though many overlapping areas of research were included in this meta-analysis, Fredricks et al. (2004) suggested that this combination of different aspects of how learners interact with their classroom and its tasks was worthwhile. The investigator's broad inclusion of varied components such as goal structure, task value, and metacognitive strategy use ensured that a number of point estimates were available for this analysis. Further, though this broad inclusion seems to confirm the lack of operational clarity about engagement as a construct, another interpretation is that broad studies such as this one will help to ensure that the development of operational clarity is authentic. Beginning with a broad characterization and a number of psychometric instruments ensures that there is an extensive pool of assessment items from which to generate a psychometrically-valid engagement instrument.

Operational clarity will emerge as longitudinal studies with purposeful sampling are conducted, and as psychometric instruments are developed and refined. The establishment of engagement trajectories of both potential orders of engagement types and the changes in engagement for particular individuals or groups will help to clarify

relevant indicators and their temporal relationship with one another. These types of studies will also provide valuable practical recommendations for practitioners in classrooms.

Science engagement research is still in its infancy, and this study attempted to collect a broad range of data about engagement predictors and outcomes with the intent of identifying further areas of research in this area. Instructional methods, class characteristics, and competence emerged as particularly effective engagement predictors in this study. Though engagement has varied operationalizations in the current research literature, it still strongly predicts achievement (Bresó et al., 2011; Chang et al., 2007; Finn & Zimmer, 2012; Fredricks et al., 2004; Nolen, 2003). Thus, science educators are encouraged to broadly consider the assessment of engagement alongside the assessment of achievement or growth in content understanding.

References

- Ainley, M. (2012). Students' interest and engagement in classroom activities. In *Handbook of research on student engagement* (pp. 283-302). New York, NY: Springer US. http://dx.doi.org/10.1007/978-1-4614-2018-7_13
- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179-211. [http://dx.doi.org/10.1016/0749-5978\(91\)90020-T](http://dx.doi.org/10.1016/0749-5978(91)90020-T)
- Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*, 84(5), 888. <http://dx.doi.org/10.1037/0033-2909.84.5.888>
- Akcay, H., Yager, R. E., Iskander, S. M., & Turgut, H. (2010). Change in student beliefs about attitudes toward science in grades 6-9. In *Asia-Pacific Forum on Science Learning and Teaching*, 11(1).
- Ames, C., & Archer, J. (1988). Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of Educational Psychology*, 80(3), 260-267. <http://dx.doi.org/10.1037/0022-0663.80.3.260>
- Anderman, E. M., & Maehr, M. L. (1994). Motivation and schooling in the middle grades. *Review of Educational Research*, 64(2), 287-309. <http://dx.doi.org/doi:10.3102/00346543064002287>
- Anderman, E. M., & Mueller, C. (2010). Middle school transitions and adolescent development. In J. L. Meece & J. S. Eccles (Eds.), *Handbook of research on schools, schooling, and human development* (pp. 198-215). New York, NY: Routledge.

- Anderson, L. W. (1975). Student involvement in learning and school achievement. *California Journal of Educational Research*, 26(2), 53-62.
- Appleton, J. J., Christenson, S. L., Kim, D., & Reschly, A. L. (2006). Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument. *Journal of School Psychology*, 44(5), 427-445.
<http://dx.doi.org/10.1016/j.jsp.2006.04.002>
- Assor, A., Kaplan, H., & Roth, G. (2002). Choice is good, but relevance is excellent: Autonomy-enhancing and suppressing teacher behaviours predicting students' engagement in schoolwork. *British Journal of Educational Psychology*, 72(2), 261-278. <http://dx.doi.org/10.1348/000709902158883>
- Avery, L. M. & Kassam, K. A. (2011). Phronesis: Children's local rural knowledge of science and engineering. *Journal of Research in Rural Education*, 26, 1-18.
- Azevedo, R. (2015). Defining and measuring engagement and learning in science: Conceptual, theoretical, methodological, and analytical issues. *Educational Psychologist*, 50(1), 84-94. <http://dx.doi.org/10.1080/00461520.2015.1004069>
- Bandura, A. (1977). Self-efficacy: A unifying theory of behavioral change. *Psychological Review*, 94(2), 191-215. <http://dx.doi.org/10.1037/0033-295X.84.2.191>
- Becker, B. J., & Wu, M. J. (2007). The synthesis of regression slopes in meta-analysis. *Statistical Science*, 414-429.
- Betts, J. E., Appleton, J. J., Reschly, A. L., Christenson, S. L., & Huebner, E. S. (2010). A study of the factorial invariance of the Student Engagement Instrument (SEI): Results from middle and high school students. *School Psychology Quarterly*, 25(2), 84.

- Biostat. (2015). *Comprehensive Meta-Analysis, Version 3*. [Software] Englewood, NJ.
- Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. West Sussex, United Kingdom: Wiley and Sons.
- Bowman, N. A. (2012). Effect sizes and statistical methods for meta-analysis in higher education. *Research in Higher Education, 53*(3), 375-382.
<http://dx.doi.org/10.1007/s11162-011-9232-5>
- Braund, M., & Driver, M. (2005). Pupils' perceptions of practical science in primary and secondary school: Implications for improving progression and continuity of learning. *Educational Research, 47*(1), 77-91.
<http://dx.doi.org/10.1080/0013188042000337578>
- Bresó, E., Schaufeli, W. B., & Salanova, M. (2011). Can a self-efficacy-based intervention decrease burnout, increase engagement, and enhance performance? A quasi-experimental study. *Higher Education, 61*(4), 339-355.
<http://dx.doi.org/10.1007/s10734-010-9334-6>
- Britner, S. L., & Pajares, F. (2006). Sources of science self-efficacy beliefs of middle school students. *Journal of Research in Science Teaching, 43*(5), 485-499.
<http://dx.doi.org/10.1002/tea.20131>
- Bronfenbrenner, U. (1976). The experimental ecology of education. *Educational Researcher, 5*-15.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton Mifflin Company.

- Caraway, K., & Tucker, C. M. (2003). Self-efficacy, goal orientation, and fear of failure as predictors of school engagement in high school students. *Psychology in the Schools, 40*(4), 417. <http://dx.doi.org/10.1002/pits.10092>
- Chan, D. (2009). So why ask me? Are self-report data really that bad. In C. E. Lance and R. J. Vanderberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 309-336). New York, NY: Taylor and Francis.
- Chang, M., Singh, K., & Mo, Y. (2007). Science engagement and science achievement: Longitudinal models using NELS data. *Educational Research and Evaluation, 13*(4), 349-371. <http://dx.doi.org/10.1080/13803610701702787>
- Cheung, M. W. L. (2015). Three-level meta-analysis. In M. W. L. Cheung (Ed.), *Meta-Analysis* (pp. 179-213). West Sussex, United Kingdom: Wiley and Sons. <http://dx.doi.org/10.1002/9781118957813.ch6>
- Connell, J. P., & Wellborn, J. G. (1991). Competence, autonomy, and relatedness: A motivational analysis of self-system processes. In M. Gunnar & A. L. Sroufe (Eds.), *Self processes and development. The Minnesota symposia on child psychology* (Vol. 23, pp. 43-77). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Cook-Sather, A. (2002). Authorizing students' perspectives: Toward trust, dialogue, and change in education. *Educational Researcher, 31*(4), 3-14. <http://dx.doi.org/10.3102/0013189X031004003>
- Cook-Sather, A. (2006). Sound, presence, and power: "Student voice" in educational research and reform. *Curriculum Inquiry, 36*(4), 359-390. <http://dx.doi.org/10.1111/j.1467873X.2006.00363.x>

- Cooper, H. M. (1998). *Synthesizing research: A guide for literature reviews* (Vol. 2). Thousand Oaks, CA: Sage.
- Coursol, A., & Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates: A note on meta-analysis bias. *Professional Psychology: Research and Practice, 17*(2), 136-137. <http://dx.doi.org/10.1037/0735-7028.17.2.136>
- Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York, NY: Harper & Row.
- Curtin, T. R., Ingels, S. J., Wu, S., & Heuer, R. (2002). *National education longitudinal study of 1988: Base-year to fourth follow-up data file user's manual* (NCES 2002-323). Washington, DC: US Department of Education.
- DeBacker, T. K., & Nelson, R. M. (2000). Motivation to learn science: Differences related to gender, class type, and ability. *The Journal of Educational Research, 93*(4), 245-254. <http://dx.doi.org/10.1080/00220670009598713>
- Deci, E. L., & Ryan, R. M. (1985). The general causality orientations scale: Self-determination in personality. *Journal of Research in Personality, 19*(2), 109-134. [http://dx.doi.org/10.1016/0092-6566\(85\)90023-6](http://dx.doi.org/10.1016/0092-6566(85)90023-6)
- Deci, E. L., & Ryan, R. M. (2002). *Handbook of self-determination research*. Rochester, NY: University Rochester Press.
- Doğan, U. (2014). Validity and reliability of student engagement scale. *Journal of Faculty of Education, 3*(2), 390-403. <http://dx.doi.org/10.14686/BUEFAD.201428190>

- Duval, S., & Tweedie, R. (2000). Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455-463. <http://dx.doi.org/10.1111/j.0006-341X.2000.00455.x>
- Eccles, J. S., & Midgley, C. (1989). Stage-environment fit: Developmentally appropriate classrooms for young adolescents. *Research on Motivation in Education*, *3*, 139-186.
- Eccles, J. S., Midgley, C., Wigfield, A., Buchanan, C. M., Reuman, D., Flanagan, C., & MacIver, D. (1993). Development during adolescence: The impact of stage-environment fit on young adolescents' experiences in schools and in families. *American Psychologist*, *48*(2), 90. <http://dx.doi.org/10.1037/0003-066X.48.2.90>
- Eccles, J. S., & Roeser, R. W. (2010). An ecological view of schools and development. In J. L. Meece, & J. S. Eccles (Eds.), *Handbook of research on schools, schooling, and human development* (pp. 6-21). New York, NY: Routledge.
- Eccles, J., & Wang, M. T. (2012). Part I commentary: So what is student engagement anyway? In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 133-145). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-2018-7_6
- Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge, UK: Cambridge University Press.
- Engle, R. A., & Conant, F. R. (2002). Guiding principles for fostering productive disciplinary engagement: Explaining an emergent argument in a community of

learners classroom. *Cognition and Instruction*, 20(4), 399-483.

http://dx.doi.org/10.1207/S1532690XCI2004_1

Feinstein, A. R. (1995). Meta-analysis: Statistical alchemy for the 21st century. *Journal of Clinical Epidemiology*, 48(1), 71-79. [http://dx.doi.org/10.1016/0895-](http://dx.doi.org/10.1016/0895-4356(94)00110-C)

4356(94)00110-C

Feldlaufer, H., Midgley, C., & Eccles, J. S. (1988). Student, teacher, and observer perceptions of the classroom environment before and after the transition to junior high school. *The Journal of Early Adolescence*, 8(2), 133-156.

<http://dx.doi.org/10.1177/0272431688082003>

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, 40(5), 532-538.

<http://dx.doi.org/10.1037/a001580>

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. London, England: Sage.

Finn, J. D. (1989). Withdrawing from school. *Review of Educational Research*, 59(2), 117-142. <http://dx.doi.org/10.3102/00346543059002117>

Finn, J. D., & Rock, D. A. (1997). Academic success among students at risk for school failure. *Journal of Applied Psychology*, 82(2), 221. <http://dx.doi.org/10.1037/0021-9010.82.2.221>

Finn, J. D., & Zimmer, K. S. (2012). Student engagement: What is it? Why does it matter? In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 97-131). New York, NY: Springer.

http://dx.doi.org/10.1007/978-1-4614-2018-7_5

- Ford, M. E. (1992). *Motivating humans: Goals, emotions, and personal agency beliefs*. Beverly Hills, CA: Sage.
- Fraser, B. J. (1982). Differences between student and teacher perceptions of actual and preferred classroom learning environment. *Educational Evaluation and Policy Analysis*, 511-519.
- Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74(1), 59-109. <http://dx.doi.org/10.3102/00346543074001059>
- Fredricks, J., McColskey, W., Meli, J., Montrosse, B., Mordica, J., & Mooney, K. (2011). *Issues & Answers Report: Measuring student engagement in upper elementary through high school: A description of 21 instruments* (REL Report No. 098).
- Fredricks, J. A., & McColskey, W. (2012). The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 763-782). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-2018-7_37
- 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. <http://dx.doi.org/10.1037/0022-0663.95.1.148>
- Furtak, E. M., & Kunter, M. (2012). Effects of autonomy-supportive teaching on student learning and motivation. *The Journal of Experimental Education*, 80(3), 284-316. <http://dx.doi.org/10.1080/00220973.2011.573019>

- Garris, R., Ahlers, R., Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. *Simulation & Gaming, 33*(4), 441-467.
<http://dx.doi.org/10.1177/1046878102238607>
- Glass, G. V., & McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.
- Greene, B. A. (2015). Measuring cognitive engagement with self-report scales: Reflections from over 20 years of research. *Educational Psychologist, 50*(1), 14-30. <http://dx.doi.org/10.1080/00461520.2014.989230>
- Greene, B. A., & Miller, R. B. (1996). Influences on achievement: Goals, perceived ability, and cognitive engagement. *Contemporary Educational Psychology, 21*(2), 181-192. <http://dx.doi.org/10.1006/ceps.1996.0015>
- Greene, B. A., Miller, R. B., Crowson, H. M., Duke, B. L., & Akey, K. L. (2004). Predicting high school students' cognitive engagement and achievement: Contributions of classroom perceptions and motivation. *Contemporary Educational Psychology, 29*(4), 462-482.
<http://dx.doi.org/10.1016/j.cedpsych.2004.01.006>
- Guillén-Nieto, V., & Aleson-Carbonell, M. (2012). Serious games and learning effectiveness: The case of It's a Deal! *Computers & Education, 58*, 1, 435-448.
<http://dx.doi.org/10.1016/j.compedu.2011.07.015>
- Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science, 246*-255.

- Hidi, S. (1990). Interest and its contribution as a mental resource for learning. *Review of Educational Research*, 60(4), 549-571.
<http://dx.doi.org/10.3102/00346543060004549>
- Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, 327(7414), 557-560.
<http://dx.doi.org/10.1136/bmj.327.7414.557>
- Hoy, W. K. (2001). The pupil control studies. A historical, theoretical and empirical analysis. *Journal of Educational Administration*, 39(5), 424-441.
- Işık, Ö., & Gücüm, B. (2013). The effect of project based learning approach on elementary school students' motivation toward science and technology course. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 28(28-3).
- Jaber, L. Z. (2014). *Affective dynamics of students' disciplinary engagement in science* (Doctoral dissertation). Retrieved from Dissertations Abstracts International. (Order No. 3624707)
- Jaber, L. Z., & Hammer, D. (2016). Learning to feel like a scientist. *Science Education*, 100(2), 189-220. <http://dx.doi.org/10.1002/sce.21202>
- Jenkins, E. W., & Pell, R. G. (2006). *The Relevance of Science Education Project (ROSE) in England: A summary of findings*. Leeds, UK: Centre for Studies in Science and Mathematics Education, University of Leeds.
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5) 524–532. <http://dx.doi.org/10.1177/0956797611430953>

- Kahraman, N., & Sungur, S. (2013). Antecedents and consequences of middle school students' achievement goals in science. *The Asia-Pacific Education Researcher*, 22(1), 45-60. <http://dx.doi.org/10.1007/s40299-012-0024-2>
- Kintsch, W. (1980). Learning from text, levels of comprehension, or: Why anyone would read a story anyway. *Poetics*, 9(1), 87-98.
[http://dx.doi.org/10.1016/0304-422X\(80\)90013-3](http://dx.doi.org/10.1016/0304-422X(80)90013-3)
- Kumar, D. D. (1991). A meta-analysis of the relationship between science instruction and student engagement. *Educational Review*, 43(1), 49-61.
<http://dx.doi.org/10.1080/0013191910430105>
- Lam, S. F., Wong, B. P., Yang, H., & Liu, Y. (2012). Understanding student engagement with a contextual model. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 403-419). New York, NY: Springer.
- Larson, R., & Csikszentmihalyi, M. (1983). The experience sampling method. *New Directions for Methodology of Social & Behavioral Science*, 15, 41-56.
- Lau, S., & Roeser, R. W. (2008). Cognitive abilities and motivational processes in science achievement and engagement: A person-centered analysis. *Learning and Individual Differences*, 18(4), 497-504.
- Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, UK: Cambridge University Press.
- Lee, O., & Anderson, C. W. (1993). Task engagement and conceptual change in middle school science classrooms. *American Educational Research Journal*, 30(3), 585-610. <http://dx.doi.org/10.3102/00028312030003585>

- Lee, W., & Reeve, J. (2012). Teachers' estimates of their students' motivation and engagement: being in synch with students. *Educational Psychology, 32*(6), 727-747. <http://dx.doi.org/10.1080/01443410.2012.732385>
- Lepper, M. R., Corpus, J. H., & Iyengar, S. S. (2005). Intrinsic and extrinsic motivational orientations in the classroom: age differences and academic correlates. *Journal of Educational Psychology, 97*(2), 184-196. <http://dx.doi.org/10.1037/0022-0663.97.2.184>
- Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.
- Linnenbrink, E. A., & Pintrich, P. R. (2003). The role of self-efficacy beliefs in student engagement and learning in the classroom. *Reading & Writing Quarterly, 19*(2), 119-137. <http://dx.doi.org/10.1080/10573560308223>
- Linnenbrink-Garcia, L., Patall, E. A., & Messersmith, E. E. (2013). Antecedents and consequences of situational interest. *British Journal of Educational Psychology, 83*(4), 591-614. <http://dx.doi.org/10.1111/j.2044-8279.2012.02080.x>
- Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*(12), 1181. <http://dx.doi.org/10.1037/0003-066X.48.12.1181>
- Little, T. W. (2015). *Effects of digital game-based learning on student engagement and academic achievement* (Doctoral dissertation). Retrieved from Dissertation Abstracts database. (Order No. 3721273)

- Liu, L. R. (2014). ¿Cómo aumentar el interés por la Ciencia?: Una propuesta didáctica para alumnos de 12-15 años. *Boletín de la Real Sociedad Española de Historia Natural. Sección aula, museos y colecciones, 1*, 139-157.
- Logan, M., & Skamp, K. (2008). Engaging students in science across the primary secondary interface: Listening to the students' voice. *Research in Science Education, 38*(4), 501-527. <http://dx.doi.org/10.1007/s11165-007-9063-8>
- Mahatmya, D., Lohman, B. J., Matjasko, J. L., & Farb, A. F. (2012). Engagement across developmental periods. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 45-63). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-2018-7_3
- Marks, H. M. (2000). Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. *American Educational Research Journal, 37*(1), 153-184. <http://dx.doi.org/10.3102/00028312037001153>
- Martin, A. J. (2007). Examining a multidimensional model of student motivation and engagement using a construct validation approach. *British Journal of Educational Psychology, 77*(2), 413-440. <http://dx.doi.org/10.1348/000709906X118036>
- McCombs, B. (2010). Learner-centered practices. In J. L. Meece & J. S. Eccles (Eds.), *Handbook of research on schools, schooling, and human development* (pp. 60-74). New York, NY: Routledge.
- McConney, A., Oliver, M. C., Woods-McConney, A., Schibeci, R., & Maor, D. (2014). Inquiry, engagement, and literacy in science: A retrospective, cross-national analysis using PISA 2006. *Science Education, 98*(6), 963-980. <http://dx.doi.org/10.1002/sce.21135>

- Meece, J. L., Blumenfeld, P. C., & Hoyle, R. H. (1988). Students' goal orientations and cognitive engagement in classroom activities. *Journal of Educational Psychology, 80*(4), 514-523. <http://dx.doi.org/10.1037/0022-0663.80.4.514>
- Mergendoller, J. R., Marchman, V. A., Mitman, A. L., & Packer, M. J. (1988). Task demands and accountability in middle-grade science classes. *The Elementary School Journal, 251*-265.
- Midgley, C., Feldlaufer, H., & Eccles, J. S. (1989). Student/teacher relations and attitudes toward mathematics before and after the transition to junior high school. *Child Development, 981*-992.
- Mo, Y. (2008). *Opportunity to learn, engagement, and science achievement: Evidence from TIMSS 2003 Data* (Doctoral dissertation). Retrieved from Dissertation Abstracts database. (Order No. DP19739)
- Moote, J. K., Williams, J. M., & Sproule, J. (2013). When students take control: Investigating the impact of the crest inquiry-based learning program on self-regulated processes and related motivations in young science students. *Journal of Cognitive Education and Psychology, 12*(2), 178-196.
<http://dx.doi.org/10.1891/1945-8959.12.2.178>
- Mosher, R., & McGowan, B. (1985). *Assessing student engagement in secondary schools: Alternative conceptions, strategies of assessing, and instruments*. (Resource Paper for the University of Wisconsin Research and Development Center). Madison, WI.
- Newmann, F. M. (1981). Reducing student alienation in high schools: Implications of theory. *Harvard Educational Review, 51*(4), 546-564.

- Newmann, F. M. (1992). *Student engagement and achievement in American secondary schools*. New York, NY: Teachers College Press.
- Nolen, S. B. (2003). Learning environment, motivation, and achievement in high school science. *Journal of Research in Science Teaching*, 40(4), 347-368.
<http://dx.doi.org/10.1002/tea.10080>
- Nunnally, J. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.
- Nystrand, M., & Gamoran, A. (1991). Instructional discourse, student engagement, and literature achievement. *Research in the Teaching of English*, 25(3), 261-290.
- Olitsky, S. (2007). Promoting student engagement in science: Interaction rituals and the pursuit of a community of practice. *Journal of Research in Science Teaching*, 44(1), 33-56. <http://dx.doi.org/10.1002/tea.20128>
- Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, 157-159.
- Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education*, 25(9), 1049-1079. <http://dx.doi.org/10.1080/0950069032000032199>
- Pekrun, R., & Linnenbrink-Garcia, L. (2013). Academic emotions and student engagement. In S. L. Christensen, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on school engagement* (pp. 259-282). New York, NY: Springer.
http://dx.doi.org/10.1007/978-1-4614-2018-7_12
- Pell, T., & Jarvis, T. (2001). Developing attitude to science scales for use with children of ages from five to eleven years. *International Journal of Science Education*, 23(8), 847-862. <http://dx.doi.org/10.1080/09500690010016111>

- Peterson, R. A., & Brown, S. P. (2005). On the use of beta coefficients in meta-analysis. *Journal of Applied Psychology, 90*(1), 175-181.
<http://dx.doi.org/10.1037/0021-9010.90.1.175>
- Piaget, J. (1972). *The psychology of intelligence*. Totowa, NJ: Littlefield.
- Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*(1), 33-40. <http://dx.doi.org/10.1037/0022-0663.82.1.33>
- Raphael, L. M., Pressley, M., & Mohan, L. (2008). Engaging instruction in middle school classrooms: An observational study of nine teachers. *The Elementary School Journal, 109*(1), 61-81.
- Reeve, J. (2012). A self-determination theory perspective on student engagement. In S.L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 149-172). New York, NY: Springer.
http://dx.doi.org/10.1007/978-1-4614-2018-7_7
- Reeve, J., & Tseng, C. M. (2011). Agency as a fourth aspect of students' engagement during learning activities. *Contemporary Educational Psychology, 36*(4), 257-267.
<http://dx.doi.org/10.1016/j.cedpsych.2011.05.002>
- Reschly, A. L., & Christenson, S. L. (2006). Prediction of dropout among students with mild disabilities: A case for the inclusion of student engagement variables. *Remedial and Special Education, 27*(5), 276-292.
<http://dx.doi.org/10.1177/07419325060270050301>
- Reschly, A. L., & Christenson, S. L. (2012). Jingle, jangle, and conceptual haziness: Evolution and future directions of the engagement construct. In S. L. Christenson,

- A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 3-19). New York, NY: Springer.
http://dx.doi.org/10.1007/9781-4614-2018-7_1
- Richard, F. D., Bond Jr, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, 7(4), 331.
<http://dx.doi.org/10.1037/1089-2680.7.4.331>
- Roeser, R. W., & Eccles, J. S. (1998). Adolescents' perceptions of middle school: Relation to longitudinal changes in academic and psychological adjustment. *Journal of Research on Adolescence*, 8(1), 123-158.
http://dx.doi.org/10.1207/s15327795jra0801_6
- Rosenthal, R. (1984). *Meta-analysis procedures for social research*. Beverly Hills, CA: Sage.
- Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, 52(1), 59- 82.
- Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44(10), 1276.
<http://dx.doi.org/10.1037/0003-066X.44.10.1276>
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68. <http://dx.doi.org/10.1037/0003-066X.55.1.68>
- Ryan, A. M., & Patrick, H. (2001). The classroom social environment and changes in adolescents' motivation and engagement during middle school. *American*

Educational Research Journal, 38(2), 437-460.

<http://dx.doi.org/10.3102/00028312038002437>

Schafft, K. A. and Jackson, A. (2011). *Rural education for the twenty-first century:*

Identity, place, and community in a globalizing world. University Park, PA: Penn State University Press.

Schank, R. C. (1979). Interestingness: Controlling inferences. *Artificial*

Intelligence, 12(3), 273-297. [http://dx.doi.org/10.1016/0004-3702\(79\)90009-2](http://dx.doi.org/10.1016/0004-3702(79)90009-2)

Schiefele, U. (1999). Interest and learning from text. *Scientific Studies of Reading*, 3(3),

257-279. http://dx.doi.org/10.1207/s1532799xssr0303_4

Schlechty, P. C. (2011). *Engaging students: The next level of working on the work.* San

Francisco, CA: Jossey-Bass.

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and*

bias in research findings. Thousand Oaks, CA: Sage.

Schmidt, F. L., Oh, I. S., & Hayes, T. L. (2009). Fixed-versus random-effects models in

meta-analysis: Model properties and an empirical comparison of differences in

results. *British Journal of Mathematical and Statistical Psychology*, 62(1), 97-128.

<http://dx.doi.org/10.1348/000711007X255327>

Schmidt, F. L., Pearlman, K., Hunter, J. E., & Shane, G. S. (1979). Further tests of the

Schmidt-Hunter Bayesian validity generalization procedure. *Personnel*

Psychology, 32(2), 257-281. <http://dx.doi.org/10.1111/j.1744->

6570.1979.tb02134.x

Schraw, G. (1998). Processing and recall differences among selective details. *Journal of*

Educational Psychology, 90(1), 3. <http://dx.doi.org/10.1037/0022-0663.90.1.3>

- Schraw, G., & Lehman, S. (2001). Situational interest: A review of the literature and directions for future research. *Educational Psychology Review, 13*(1), 23-52.
<http://dx.doi.org/10.1023/A:1009004801455>
- Shapiro, S. (1994). Meta-analysis/Shmeta-analysis. *American Journal of Epidemiology, 140*(9), 771-778.
- Shemoff, D. J., Csikszentmihalyi, M., Schneider, B., & Shernoff, E. S. (2003). Student engagement in high school classrooms from the perspective of flow theory. *School Psychology Quarterly, 18*(2), 158-176.
<http://dx.doi.org/10.1521/scpq.18.2.158.21860>
- Shin, I. S. (2009). *Same author and same data dependence in meta-analysis* (Doctoral dissertation). Retrieved from Dissertation Abstracts database. (Order No. 3385307).
- Sinatra, G. M., Heddy, B. C., & Lombardi, D. (2015). The challenges of defining and measuring student engagement in science. *Educational Psychologist, 50*(1), 1-13.
<http://dx.doi.org/10.1080/00461520.2014.1002924>
- Singh, K., Granville, M., & Dika, S. (2002). Mathematics and science achievement: Effects of motivation, interest, and academic engagement. *The Journal of Educational Research, 95*(6), 323-332.
<http://dx.doi.org/10.1080/00220670209596607>
- Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology, 85*(4), 571-580. <http://dx.doi.org/10.1037/0022-0663.85.4.571>

- Skinner, E. A., Kindermann, T. A., & Furrer, C. J. (2008). A motivational perspective on engagement and disaffection: Conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. *Educational and Psychological Measurement*.
<http://dx.doi.org/10.1177/0013164408323233>
- Skinner, E. A., & Pitzer, J. R. (2012). Developmental dynamics of student engagement, coping, and everyday resilience. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 21-44). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-2018-7_2
- Slavin, R. E., Hurley, E. A., & Chamberlain, A. (2003). Cooperative learning and achievement: Theory and research. *Handbook of psychology*, 3(9), 177-198.
<http://dx.doi.org/10.1002/0471264385.wei0709>
- Smart, J. B. (2014). A mixed methods study of the relationship between student perceptions of teacher-student interactions and motivation in middle level science. *RMLE Online*, 38(4), 1.
- Snow, R. E. (1992). Aptitude theory: Yesterday, today, and tomorrow. *Educational Psychologist*, 27(1), 5-32. http://dx.doi.org/10.1207/s15326985ep2701_3
- Spearman, J., & Watt, H. M. (2013). Perception shapes experience: The influence of actual and perceived classroom environment dimensions on girls' motivations for science. *Learning Environments Research*, 16(2), 217-238.
<http://dx.doi.org/10.1007/s10984-013-9129-7>

- Stallings, J. (1980). Allocated academic learning time revisited, or beyond time on task. *Educational Researcher*, 9(11), 11-16.
<http://dx.doi.org/10.3102/0013189X009011011>
- Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, 22(13), 2113-2126.
<http://dx.doi.org/10.1002/sim.1461>
- Uekawa, K., Borman, K., & Lee, R. (2007). Student engagement in US urban high school mathematics and science classrooms: Findings on social organization, race, and ethnicity. *The Urban Review*, 39(1), 1-43. <http://dx.doi.org/10.1007/s11256-006-0039-1>
- Vedder-Weiss, D., & Fortus, D. (2011). Adolescents' declining motivation to learn science: Inevitable or not? *Journal of Research in Science Teaching*, 48(2), 199-216. <http://dx.doi.org/10.1002/tea.20398>
- Vedder-Weiss, D., & Fortus, D. (2012). Adolescents' declining motivation to learn science: A follow-up study. *Journal of Research in Science Teaching*, 49(9), 1057-1095. <http://dx.doi.org/10.1002/tea.21049>
- Vedder-Weiss, D., & Fortus, D. (2013). School, teacher, peers, and parents' goals emphases and adolescents' motivation to learn science in and out of school. *Journal of Research in Science Teaching*, 50(8), 952-988.
<http://dx.doi.org/10.1002/tea.21103>
- Veiga, F., Reeve, J., Wentzel, K., & Robu, V. (2014). Assessing students' engagement: A review of instruments with psychometric qualities. In F. Veiga (Coord.), *Students'*

engagement in school: International perspectives of psychology and education (pp. 38-57). Lisboa, Portugal: Instituto de Educação da Universidade de Lisboa.

- Veiga, F. H., & Robu, V. (2014). Measuring student engagement with school across cultures: Psychometric findings from Portugal and Romania. *Romanian Journal of School Psychology, 14*, 57-72.
- Vygotsky, L. (1978). Interaction between learning and development. *Readings on the Development of Children, 23*(3), 34-41.
- Walberg, H. J., & Haertel, G. D. (1980). Evaluation reflections: Validity and use of educational environment assessments. *Studies in Educational Evaluation, 6*(3), 225-238. [http://dx.doi.org/10.1016/0191-491X\(80\)90026-7](http://dx.doi.org/10.1016/0191-491X(80)90026-7)
- Walberg, H. J., House, E. R., & Steele, J. M. (1973). Grade level, cognition, and affect: A cross-section of classroom perceptions. *Journal of Educational Psychology, 64*(2), 142. <http://dx.doi.org/10.1037/h0034614>
- Wang, M. T., & Holcombe, R. (2010). Adolescents' perceptions of school environment, engagement, and academic achievement in middle school. *American Educational Research Journal, 47*(3), 633-662. <http://dx.doi.org/10.3102/0002831209361209>
- Wang, M. T., Willett, J. B., & Eccles, J. S. (2011). The assessment of school engagement: Examining dimensionality and measurement invariance by gender and race/ethnicity. *Journal of School Psychology, 49*(4), 465-480. <http://dx.doi.org/10.1016/j.jsp.2011.04.001>
- Wehlage, G. G., Rutter, R. A., Smith, G. A., Lesko, N., & Fernandez, R. R. (1989). *Reducing the risk: Schools as communities of support*. London, England: Falmer Press.

- Wentzel, K. R. (2010). Students' relationships with teachers. In J. L. Meece, & J. S. Eccles (Eds.), *Handbook of research on schools, schooling, and human development* (pp. 75-91). New York, NY: Routledge.
- Wilson, D. B. (2015). Practical meta-analysis effect size calculator [Web software]. Fairfax, VA: George Mason University.
- Wilson, G. T., & Rachman, S. J. (1983). Meta-analysis and the evaluation of psychotherapy outcome: Limitations and liabilities. *Journal of Consulting and Clinical Psychology, 51*(1), 54-64. <http://dx.doi.org/10.1037/0022-006X.51.1.54>
- Wolf, S. J., & Fraser, B. J. (2008). Learning environment, attitudes and achievement among middle-school science students using inquiry-based laboratory activities. *Research in Science Education, 38*(3), 321-341. <http://dx.doi.org/10.1007/s11165-007-9052-y>
- Yazzie-Mintz, E., & McCormick, K. (2012). Finding the humanity in the data: Understanding, measuring, and strengthening student engagement. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 743-761). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-2018-7_36
- Zheng, M., & Spires, H. A. (2014). Fifth graders' flow experience in a digital game-based science learning environment. *International Journal of Virtual and Personal Learning Environments, 5*(2), 69-86. <http://dx.doi.org/10.4018/ijvple.2014040106>
- Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist, 25*(1) 3-17. http://dx.doi.org/10.1207/s15326985ep2501_2

Zucker, A., Tinker, R., Staudt, C., Mansfield, A., & Metcalf, S. (2008). Learning science in grades 3–8 using probeware and computers: Findings from the TEEMSS II project. *Journal of Science Education and Technology, 17*, 42–48.
<http://dx.doi.org/10.1007/s10956-007-9086-y>

Appendices

Appendix A

Coding Scheme

Table A1

Coding Scheme for Included Studies

| | Source Characteristics |
|------------------------------|---|
| Publication | 0: no 1: yes |
| Peer-reviewed | 0: no 1: yes |
| | Study Characteristics |
| Predictor Type | 1: instructional method 2: technology 3: class characteristics 4: social characteristics |
| Engagement Conceptualization | 1: behavioral 2: affective 3: cognitive 4: behavioral and affective 5: behavioral and cognitive 6: affective and cognitive 7: all three |
| Grade Level | 1: 5 th grade (ages 10-11) 2: 6 th -8 th grades (ages 11-14) 3: 9 th grade (ages 14-15) 4: mix |
| School Structure | 0: not specified 1: elementary 2: middle school 3: junior high 4: K-8 5: high school 6: other/mix |
| School Type | 0: not specified 1: public 2: charter 3: private/independent 4: alternative 5: mix |
| School Location | 0: not specified 1: rural 2: suburban 3: urban 4: mix |

| | |
|--------------------------|--|
| Socio-economic Status | <ul style="list-style-type: none"> 0: not specified 1: low (greater than 60% FRL), 2: average (35-59% FRL) 3: high (<35%) 4: mix |
| Experimental Design | <ul style="list-style-type: none"> 1: correlation/regression/SEM 2: single group (repeated measures) 3: quasi-experimental 4: experimental |
| Reliability ^a | <ul style="list-style-type: none"> 0: not reported 1: references external instrument, no measure 2: references external instrument reliability 3: internal reliability < .70 4: internal reliability > .70 |
| Validity | <ul style="list-style-type: none"> 0: not reported or some face/content validity (assessed by investigator) 1: face/content validity (assessed by investigator) 2: face/content validity (assessed by study) 3: references external instrument 4: references external instrument validity 5: internal validity measures (EFA, CFA, etc.) |
| Repeat Authors | <ul style="list-style-type: none"> 0: study with unique authors 1: study with authors duplicated in another study |

^a Actual reliability recorded and then collapsed into categories

Appendix B

Statistics and Moderators by Point Estimate

Table B1

Statistics and Moderators by Point Estimate

| First Author & Year | Statistics | | | Specific | PV | | CV | | | Study | | | Inst. | | School Variables | | | | |
|---------------------|------------|------|--------|------------------------------------|----|-----|----|---|---|-------|----|---|-------|----|------------------|----|----|-----|-----|
| | <i>g</i> | SE | Var | | T | SDT | E | M | A | P | PR | R | V | GL | T | St | Se | SES | Geo |
| Akcay 2010 | 2.51 | 0.22 | 0.0490 | STS | 1 | 1 | 2 | 3 | 0 | 1 | 1 | 4 | 3 | 4 | 0 | 0 | 0 | 0 | 0 |
| Aktamis 2008 | 0.43 | 0.31 | 0.0984 | Science Process Skills | 1 | 2 | 2 | 3 | 0 | 1 | 1 | 4 | 2 | 2 | 0 | 4 | 0 | 0 | 1 |
| Bathgate 2013 | 0.19 | 0.04 | 0.0014 | Analyzing v Action | 1 | 2 | 2 | 2 | 0 | 1 | 1 | 4 | 2 | 4 | 0 | 0 | 0 | 0 | 0 |
| Bathgate 2013 | 0.19 | 0.04 | 0.0014 | Consuming new knowledge v action | 1 | 2 | 2 | 2 | 0 | 1 | 1 | 4 | 2 | 4 | 0 | 0 | 0 | 0 | 0 |
| Bawaneh 2012 | 0.99 | 0.30 | 0.0880 | Whole Brain Teaching | 1 | 2 | 6 | 3 | 0 | 1 | 1 | 4 | 4 | 2 | 0 | 0 | 0 | 0 | 1 |
| Bilgin 2006 | 0.67 | 0.27 | 0.0747 | Cooperative learning v demo | 4 | 3 | 2 | 3 | 0 | 1 | 1 | 2 | 3 | 2 | 1 | 4 | 3 | 0 | 1 |
| Blanchard 2015 | 0.00 | 0.13 | 0.0178 | Innovation Club; inquiry | 1 | 1 | 2 | 3 | 0 | 1 | 1 | 0 | 3 | 2 | 1 | 2 | 1 | 1 | 0 |
| Bowling 2013 | 0.13 | 0.07 | 0.0048 | Microbiology game | 2 | 1 | 2 | 3 | 0 | 1 | 1 | 4 | 3 | 2 | 1 | 2 | 0 | 4 | 1 |
| Bozdogan 2009 | 1.17 | 0.23 | 0.0526 | Exhibit/Activ. at Learning Center | 3 | 3 | 2 | 2 | 0 | 1 | 1 | 4 | 5 | 2 | 1 | 4 | 0 | 0 | 1 |
| Brown 2013 | 0.13 | 0.08 | 0.0060 | PBL-Web | 2 | 1 | 2 | 2 | 0 | 1 | 1 | 4 | 4 | 2 | 0 | 0 | 4 | 4 | 0 |
| Cetin-Dindar 2015 | -0.22 | 0.13 | 0.0168 | Perceptions of a Constructivist LE | 3 | 1 | 2 | 1 | 0 | 1 | 1 | 3 | 3 | 2 | 1 | 4 | 0 | 0 | 1 |
| Chen 2010 | -0.01 | 0.04 | 0.0014 | Live Simulation | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 3 | 2 | 1 | 2 | 0 | 0 | 0 |
| Chen 2014 | 1.67 | 0.18 | 0.0316 | Scaffolding w e-learning | 1 | 2 | 3 | 3 | 1 | 1 | 0 | 4 | 3 | 2 | 1 | 2 | 3 | 0 | 1 |

| | | | | | | | | | | | | | | | | | | | |
|------------------|-------|------|--------|--|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chen 2015 | 0.33 | 0.28 | 0.0786 | Collaborative vs. Ind. Game Play | 4 | 3 | 2 | 4 | 1 | 1 | 1 | 2 | 4 | 2 | 0 | 2 | 0 | 0 | 0 |
| Chen 2015 | 0.30 | 0.28 | 0.0784 | Collaborative vs. Ind. Game play | 4 | 3 | 3 | 4 | 1 | 1 | 1 | 2 | 4 | 2 | 0 | 2 | 0 | 0 | 0 |
| Cheng 2014 | 0.08 | 0.17 | 0.0300 | Humunology Game | 2 | 1 | 2 | 3 | 0 | 1 | 1 | 4 | 3 | 4 | 0 | 0 | 0 | 0 | 0 |
| Cheng 2014 | 0.03 | 0.17 | 0.0300 | Humunology Game | 2 | 1 | 3 | 3 | 0 | 1 | 1 | 4 | 4 | 4 | 0 | 0 | 0 | 0 | 0 |
| Cirik 2014 | 0.77 | 0.05 | 0.0030 | Teacher Social Support | 4 | 3 | 3 | 1 | 0 | 1 | 1 | 3 | 5 | 4 | 0 | 3 | 0 | 4 | 1 |
| Dettweiler 2015 | 0.37 | 0.16 | 0.0248 | Outdoor vs. Indoor Education | 1 | 1 | 3 | 3 | 0 | 1 | 1 | 3 | 5 | 4 | 5 | 6 | 0 | 0 | 1 |
| Doll et al. 2010 | 1.54 | 0.08 | 0.0063 | Student-Teacher Relationship | 4 | 3 | 2 | 1 | 0 | 1 | 1 | 4 | 3 | 4 | 5 | 6 | 4 | 0 | 0 |
| Furtak 2012 | 0.04 | 0.39 | 0.1491 | Cog. Autonomy Support | 3 | 1 | 2 | 4 | 0 | 1 | 1 | 4 | 3 | 2 | 0 | 0 | 0 | 0 | 1 |
| Furtak 2012 | -0.02 | 0.46 | 0.2127 | Cog. Autonomy Support | 3 | 1 | 1 | 4 | 0 | 1 | 1 | 3 | 1 | 2 | 0 | 0 | 0 | 0 | 1 |
| Furtak 2012 | -0.75 | 0.38 | 0.1438 | Proc. + Cog. Autonomy Support | 3 | 1 | 2 | 4 | 0 | 1 | 1 | 4 | 3 | 2 | 0 | 0 | 0 | 0 | 1 |
| Furtak 2012 | -0.14 | 0.44 | 0.1916 | Proc. + Cog. Autonomy Support | 3 | 1 | 1 | 4 | 0 | 1 | 1 | 3 | 1 | 2 | 0 | 0 | 0 | 0 | 1 |
| Furtak 2012 | -0.12 | 0.38 | 0.1435 | Proc. Autonomy Support | 3 | 1 | 2 | 4 | 0 | 1 | 1 | 4 | 3 | 2 | 0 | 0 | 0 | 0 | 1 |
| Furtak 2012 | 0.03 | 0.45 | 0.2041 | Proc. Autonomy Support | 3 | 1 | 1 | 4 | 0 | 1 | 1 | 3 | 1 | 2 | 0 | 0 | 0 | 0 | 1 |
| Genc 2015 | 1.80 | 0.29 | 0.0857 | Scientific study (research + discussion) | 1 | 3 | 2 | 2 | 0 | 1 | 0 | 4 | 3 | 2 | 0 | 0 | 0 | 0 | 1 |
| Glenn 2015 | 0.13 | 0.11 | 0.0118 | Junior Master Gardener AS Program: | 1 | 1 | 2 | 2 | 0 | 1 | 1 | 4 | 3 | 2 | 0 | 0 | 1 | 0 | 1 |
| Grolnick 2007 | 0.56 | 0.21 | 0.0454 | Investigators' Club | 3 | 1 | 7 | 4 | 0 | 1 | 1 | 4 | 3 | 2 | 0 | 0 | 3 | 1 | 0 |
| Grolnick 2007 | 0.09 | 0.21 | 0.0437 | AS Program: | 3 | 1 | 3 | 4 | 0 | 1 | 1 | 4 | 4 | 2 | 0 | 0 | 3 | 1 | 0 |

| | | | | | | | | | | | | | | | | | | | |
|--------------------|-------|------|--------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Hong 2010 | 1.32 | 0.23 | 0.0520 | Investigators' Club Collaborative science investigation | 1 | 3 | 2 | 3 | 0 | 1 | 1 | 3 | 3 | 2 | 1 | 3 | 0 | 0 | 1 |
| Isik 2013 | 1.89 | 0.28 | 0.0762 | Project-Based Learning | 1 | 1 | 2 | 3 | 0 | 1 | 0 | 2 | 3 | 2 | 1 | 4 | 0 | 0 | 1 |
| Isik 2013 | 2.45 | 0.25 | 0.0650 | Project-Based Learning | 1 | 1 | 3 | 3 | 0 | 1 | 0 | 2 | 3 | 2 | 1 | 4 | 0 | 0 | 1 |
| Kahraman 2012 | 0.77 | 0.07 | 0.0047 | Teacher's Mastery Goals | 3 | 2 | 2 | 1 | 1 | 1 | 0 | 2 | 3 | 2 | 1 | 4 | 0 | 0 | 1 |
| Kahraman 2012 | 0.13 | 0.05 | 0.0028 | Teacher's Mastery Goals | 3 | 2 | 3 | 1 | 1 | 1 | 0 | 2 | 3 | 2 | 1 | 4 | 0 | 0 | 1 |
| Kaloti-Hallak 2015 | -0.20 | 0.13 | 0.0168 | Lego Robotics Competition | 1 | 1 | 2 | 2 | 0 | 1 | 0 | 4 | 3 | 4 | 1 | 2 | 0 | 0 | 1 |
| Kanter 2010 | 0.21 | 0.12 | 0.0147 | Support analyzing data | 1 | 2 | 2 | 1 | 0 | 1 | 1 | 4 | 3 | 2 | 1 | 0 | 3 | 1 | 0 |
| Kanter 2010 | 0.32 | 0.12 | 0.0147 | Support explaining to others | 1 | 3 | 2 | 1 | 0 | 1 | 1 | 4 | 3 | 2 | 1 | 0 | 3 | 1 | 0 |
| Kingir 2013 | 0.26 | 0.07 | 0.0051 | Critical Voice | 3 | 1 | 2 | 1 | 0 | 1 | 1 | 3 | 5 | 2 | 1 | 4 | 0 | 0 | 1 |
| Kingir 2013 | 0.05 | 0.06 | 0.0040 | Critical Voice | 3 | 1 | 3 | 1 | 0 | 1 | 1 | 3 | 5 | 2 | 1 | 4 | 0 | 0 | 1 |
| Kingir 2013 | 0.34 | 0.07 | 0.0051 | Relevance | 3 | 1 | 2 | 1 | 0 | 1 | 1 | 3 | 5 | 2 | 1 | 4 | 0 | 0 | 1 |
| Kingir 2013 | 0.00 | 0.06 | 0.0035 | Relevance | 3 | 1 | 3 | 1 | 0 | 1 | 1 | 3 | 5 | 2 | 1 | 4 | 0 | 0 | 1 |
| Kingir 2013 | 0.47 | 0.07 | 0.0053 | Shared Control | 3 | 1 | 2 | 1 | 0 | 1 | 1 | 3 | 5 | 2 | 1 | 4 | 0 | 0 | 1 |
| Kingir 2013 | 0.36 | 0.06 | 0.0035 | Shared Control | 3 | 1 | 3 | 1 | 0 | 1 | 1 | 3 | 5 | 2 | 1 | 4 | 0 | 0 | 1 |
| Kingir 2013 | 0.41 | 0.07 | 0.0052 | Student Negotiation | 3 | 1 | 2 | 1 | 0 | 1 | 1 | 3 | 5 | 2 | 1 | 4 | 0 | 0 | 1 |
| Kingir 2013 | 0.31 | 0.06 | 0.0035 | Student Negotiation | 3 | 1 | 3 | 1 | 0 | 1 | 1 | 3 | 5 | 2 | 1 | 4 | 0 | 0 | 1 |
| Kingir 2013 | 0.20 | 0.07 | 0.0050 | Uncertainty | 3 | 2 | 2 | 1 | 0 | 1 | 1 | 3 | 5 | 2 | 1 | 4 | 0 | 0 | 1 |
| Kingir 2013 | 0.27 | 0.06 | 0.0035 | Uncertainty | 3 | 2 | 3 | 1 | 0 | 1 | 1 | 3 | 5 | 2 | 1 | 4 | 0 | 0 | 1 |
| Kose 2010 | 0.35 | 0.24 | 0.0584 | Cooperative learning | 4 | 3 | 2 | 3 | 0 | 1 | 1 | 2 | 3 | 2 | 1 | 4 | 3 | 0 | 1 |

| | | | | | | | | | | | | | | | | | | | |
|-------------------------|-------|------|--------|--------------------------------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Kurbanoglu 2015 | 0.37 | 0.24 | 0.0569 | Context-based Questions | 1 | 3 | 2 | 3 | 0 | 1 | 1 | 4 | 3 | 2 | 0 | 4 | 0 | 0 | 1 |
| Larson 2014 | 1.01 | 0.13 | 0.0157 | Graphic Organizer | 1 | 2 | 2 | 3 | 0 | 1 | 1 | 4 | 3 | 3 | 1 | 5 | 2 | 4 | 0 |
| Larson 2014 | 1.37 | 0.16 | 0.0241 | Graphic Organizer | 1 | 2 | 6 | 3 | 0 | 1 | 1 | 4 | 3 | 3 | 1 | 5 | 2 | 4 | 0 |
| Linnenbrink-Garcia 2013 | 0.74 | 0.22 | 0.0468 | Teacher Approachability | 4 | 3 | 2 | 1 | 0 | 1 | 1 | 4 | 3 | 4 | 0 | 0 | 0 | 0 | 0 |
| Linnenbrink-Garcia 2013 | 0.16 | 0.20 | 0.0413 | Relevance | 3 | 1 | 2 | 1 | 0 | 1 | 1 | 4 | 3 | 4 | 0 | 0 | 0 | 0 | 0 |
| Linnenbrink-Garcia 2013 | -0.06 | 0.20 | 0.0411 | Group Work | 4 | 3 | 2 | 1 | 0 | 1 | 1 | 4 | 3 | 4 | 0 | 0 | 0 | 0 | 0 |
| Linnenbrink-Garcia 2013 | 0.82 | 0.22 | 0.0479 | Perceived Choice | 3 | 1 | 2 | 1 | 0 | 1 | 1 | 4 | 3 | 4 | 0 | 0 | 0 | 0 | 0 |
| Linnenbrink-Garcia 2013 | 0.04 | 0.20 | 0.0410 | Involvement Supports | 4 | 3 | 2 | 1 | 0 | 1 | 1 | 4 | 3 | 4 | 0 | 0 | 0 | 0 | 0 |
| Liu 2014 | 0.13 | 0.06 | 0.0042 | Inquiry Activities | 1 | 1 | 2 | 2 | 0 | 1 | 0 | 1 | 3 | 4 | 0 | 0 | 3 | 0 | 1 |
| Long 2015 | 0.73 | 0.11 | 0.0117 | Curriculum: Spec. v Gen. | 1 | 2 | 2 | 3 | 1 | 1 | 1 | 3 | 5 | 2 | 1 | 2 | 2 | 2 | 0 |
| Long 2015 | 0.07 | 0.10 | 0.0110 | Curriculum: Spec. v Gen. | 1 | 2 | 1 | 3 | 1 | 1 | 1 | 3 | 5 | 2 | 1 | 2 | 2 | 2 | 0 |
| Long 2015 | -0.02 | 0.10 | 0.0109 | Curriculum: Spec. v Gen. | 1 | 2 | 3 | 3 | 1 | 1 | 1 | 3 | 5 | 2 | 1 | 2 | 2 | 2 | 0 |
| McConney 2014 | 0.19 | 0.02 | 0.0002 | Level of inquiry learning | 1 | 2 | 2 | 3 | 0 | 1 | 1 | 1 | 3 | 3 | 0 | 0 | 4 | 4 | 1 |
| Moote 2013 | 0.39 | 0.20 | 0.0385 | CREST-Student run sci. project | 1 | 1 | 2 | 3 | 0 | 1 | 1 | 4 | 3 | 2 | 3 | 0 | 3 | 0 | 1 |
| Moote 2013 | 0.28 | 0.14 | 0.0192 | CREST-Student run sci. project | 1 | 1 | 3 | 3 | 0 | 1 | 1 | 4 | 3 | 2 | 3 | 0 | 3 | 0 | 1 |
| Nelson 2008 | 0.32 | 0.13 | 0.0163 | Belongingness | 4 | 3 | 3 | 1 | 0 | 1 | 1 | 4 | 4 | 4 | 1 | 6 | 2 | 3 | 0 |
| Nelson 2008 | 0.12 | 0.13 | 0.0160 | Class Involvement | 4 | 3 | 3 | 1 | 0 | 1 | 1 | 4 | 4 | 4 | 1 | 6 | 2 | 3 | 0 |
| Ng et al. 2015 | 1.17 | 0.06 | 0.0040 | Autonomy (perc. of T and S) | 3 | 1 | 2 | 1 | 0 | 1 | 1 | 1 | 3 | 4 | 0 | 0 | 0 | 0 | 1 |
| Ng et al. 2015 | 0.71 | 0.06 | 0.0038 | Autonomy (perc. of T and S) | 3 | 1 | 3 | 1 | 0 | 1 | 1 | 1 | 3 | 4 | 0 | 0 | 0 | 0 | 1 |

| | | | | | | | | | | | | | | | | | | | |
|--------------|-------|------|--------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ng 2015 | 0.77 | 0.08 | 0.0059 | Competence | 3 | 2 | 3 | 1 | 0 | 1 | 1 | 1 | 3 | 4 | 0 | 0 | 0 | 0 | 1 |
| Ng 2015 | 1.25 | 0.08 | 0.0071 | Competence | 3 | 2 | 2 | 1 | 0 | 1 | 1 | 1 | 3 | 4 | 0 | 0 | 0 | 0 | 1 |
| Ng 2015 | 0.72 | 0.08 | 0.0058 | Relatedness | 4 | 3 | 2 | 1 | 0 | 1 | 1 | 1 | 3 | 4 | 0 | 0 | 0 | 0 | 1 |
| Ng 2015 | 0.67 | 0.08 | 0.0057 | Relatedness | 4 | 3 | 3 | 1 | 0 | 1 | 1 | 1 | 3 | 4 | 0 | 0 | 0 | 0 | 1 |
| Nugent 2010 | 0.60 | 0.10 | 0.0096 | Robotics | 2 | 1 | 2 | 2 | 0 | 1 | 1 | 3 | 5 | 2 | 1 | 0 | 3 | 0 | 0 |
| Nugent 2010 | 0.55 | 0.10 | 0.0093 | Robotics | 2 | 1 | 3 | 2 | 0 | 1 | 1 | 3 | 5 | 2 | 1 | 0 | 3 | 0 | 0 |
| O'Leary 2011 | 0.53 | 0.16 | 0.0246 | Universally- designed worksheets | 1 | 2 | 1 | 2 | 0 | 1 | 1 | 1 | 3 | 2 | 5 | 5 | 3 | 0 | 1 |
| Odom 2011 | 0.14 | 0.12 | 0.0137 | Computer Usage | 2 | 1 | 1 | 1 | 0 | 1 | 1 | 4 | 5 | 2 | 5 | 2 | 4 | 2 | 0 |
| Odom 2011 | 1.01 | 0.13 | 0.0171 | Student centered instruction | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 4 | 5 | 2 | 5 | 2 | 4 | 2 | 0 |
| Odom 2011 | 0.14 | 0.12 | 0.0137 | Teacher centered instruction | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 3 | 5 | 2 | 5 | 2 | 4 | 2 | 0 |
| Osborne 2013 | 0.18 | 0.07 | 0.0047 | Scientific Argumentation | 1 | 3 | 2 | 3 | 0 | 1 | 1 | 4 | 4 | 2 | 0 | 0 | 3 | 3 | 1 |
| Osborne 2013 | 0.09 | 0.01 | 0.0001 | Scientific Argumentation | 1 | 3 | 3 | 3 | 0 | 1 | 1 | 4 | 4 | 2 | 0 | 0 | 3 | 3 | 1 |
| Park 2009 | 0.23 | 0.07 | 0.0044 | CAI | 2 | 1 | 2 | 2 | 0 | 1 | 1 | 4 | 3 | 2 | 1 | 2 | 0 | 0 | 1 |
| Serin 2015 | 0.46 | 0.30 | 0.0921 | Constructivist CAI | 2 | 1 | 2 | 4 | 0 | 1 | 1 | 4 | 3 | 2 | 1 | 4 | 0 | 0 | 1 |
| Skinner 2012 | 0.92 | 0.13 | 0.0157 | Autonomy | 3 | 1 | 4 | 1 | 0 | 1 | 1 | 4 | 5 | 2 | 1 | 2 | 0 | 1 | 0 |
| Skinner 2012 | 0.34 | 0.12 | 0.0134 | Competence | 3 | 2 | 4 | 1 | 0 | 1 | 1 | 4 | 5 | 2 | 1 | 2 | 0 | 1 | 0 |
| Smart 2014 | 0.18 | 0.13 | 0.0182 | Perception of teacher as admonishing | 4 | 1 | 2 | 1 | 0 | 1 | 1 | 4 | 2 | 2 | 1 | 2 | 2 | 3 | 0 |
| Smart 2014 | -0.22 | 0.14 | 0.0183 | Perception of teacher as admonishing | 4 | 1 | 3 | 1 | 0 | 1 | 1 | 4 | 2 | 2 | 1 | 2 | 2 | 3 | 0 |
| Smart 2014 | 0.12 | 0.13 | 0.0181 | Perception of teacher as dissatisfied | 4 | 3 | 3 | 1 | 0 | 1 | 1 | 4 | 2 | 2 | 1 | 2 | 2 | 3 | 0 |
| Smart 2014 | -0.14 | 0.13 | 0.0181 | Perception of teacher as | 4 | 3 | 2 | 1 | 0 | 1 | 1 | 4 | 2 | 2 | 1 | 2 | 2 | 3 | 0 |

| | | | | | | | | | | | | | | | | | | | |
|---------------|-------|------|--------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | dissatisfied | | | | | | | | | | | | | | | |
| Smart 2014 | 0.16 | 0.13 | 0.0182 | Perception of teacher as helpful/friendly | 4 | 3 | 2 | 1 | 0 | 1 | 1 | 4 | 2 | 2 | 1 | 2 | 2 | 3 | 0 |
| Smart 2014 | 0.18 | 0.13 | 0.0182 | Perception of teacher as helpful/friendly | 4 | 3 | 3 | 1 | 0 | 1 | 1 | 4 | 2 | 2 | 1 | 2 | 2 | 3 | 0 |
| Smart 2014 | 0.30 | 0.14 | 0.0185 | Perception of teacher leadership | 3 | 3 | 2 | 1 | 0 | 1 | 1 | 4 | 2 | 2 | 1 | 2 | 2 | 3 | 0 |
| Smart 2014 | 0.47 | 0.14 | 0.0191 | Perception of teacher leadership | 3 | 3 | 3 | 1 | 0 | 1 | 1 | 4 | 2 | 2 | 1 | 2 | 2 | 3 | 0 |
| Smart 2014 | -0.28 | 0.14 | 0.0184 | Perception of teacher strictness | 4 | 1 | 2 | 1 | 0 | 1 | 1 | 4 | 2 | 2 | 1 | 2 | 2 | 3 | 0 |
| Smart 2014 | -0.10 | 0.13 | 0.0181 | Perception of teacher strictness | 4 | 1 | 3 | 1 | 0 | 1 | 1 | 4 | 2 | 2 | 1 | 2 | 2 | 3 | 0 |
| Smart 2014 | -0.30 | 0.14 | 0.0185 | Perception of student freedom | 3 | 1 | 2 | 1 | 0 | 1 | 1 | 4 | 2 | 2 | 1 | 2 | 2 | 3 | 0 |
| Smart 2014 | -0.41 | 0.14 | 0.0188 | Perception of student freedom | 3 | 1 | 3 | 1 | 0 | 1 | 1 | 4 | 2 | 2 | 1 | 2 | 2 | 3 | 0 |
| Smart 2014 | -0.16 | 0.13 | 0.0182 | Perception of teacher as understanding | 4 | 3 | 2 | 1 | 0 | 1 | 1 | 4 | 2 | 2 | 1 | 2 | 2 | 3 | 0 |
| Smart 2014 | -0.08 | 0.13 | 0.0181 | Perception of teacher as understanding | 4 | 3 | 3 | 1 | 0 | 1 | 1 | 4 | 2 | 2 | 1 | 2 | 2 | 3 | 0 |
| Spearman 2013 | 0.07 | 0.28 | 0.0793 | Teacher relatedness | 4 | 3 | 2 | 1 | 0 | 1 | 1 | 4 | 3 | 2 | 3 | 0 | 0 | 0 | 1 |
| Sungur 2009 | 1.35 | 0.08 | 0.0070 | Classroom Goal Structure | 3 | 2 | 3 | 1 | 1 | 1 | 1 | 4 | 3 | 2 | 1 | 4 | 0 | 2 | 1 |
| Swarat 2012 | 0.55 | 0.05 | 0.0022 | Hands-on Work | 1 | 1 | 4 | 2 | 0 | 1 | 1 | 4 | 1 | 2 | 1 | 3 | 2 | 0 | 0 |
| Swarat 2012 | 0.76 | 0.05 | 0.0024 | Technology | 1 | 1 | 4 | 2 | 0 | 1 | 1 | 4 | 1 | 2 | 1 | 3 | 2 | 0 | 0 |
| Tapola 2013 | 0.45 | 0.14 | 0.0206 | Abstract vs. concrete | 1 | 2 | 2 | 2 | 0 | 1 | 1 | 1 | 3 | 4 | 1 | 1 | 0 | 0 | 1 |
| Turkmen 2009 | 0.73 | 0.21 | 0.0434 | Tech-based | 2 | 1 | 2 | 3 | 0 | 1 | 1 | 2 | 3 | 1 | 1 | 4 | 0 | 0 | 1 |

| | | | | | | | | | | | | | | | | | | | |
|-------------------|-------|------|--------|--|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | inquiry approach | | | | | | | | | | | | | | | |
| Vedder-Weiss 2011 | 0.28 | 0.18 | 0.0310 | Democratic vs. Traditional Schools | 3 | 1 | 3 | 3 | 1 | 1 | 1 | 4 | 4 | 4 | 5 | 6 | 4 | 4 | 1 |
| Vedder-Weiss 2011 | 0.09 | 0.13 | 0.0160 | Democratic vs. Traditional Schools | 3 | 1 | 5 | 3 | 1 | 1 | 1 | 4 | 4 | 4 | 5 | 6 | 4 | 4 | 1 |
| Vedder-Weiss 2012 | -0.03 | 0.20 | 0.0390 | Democratic vs. Traditional Schools | 3 | 1 | 3 | 3 | 1 | 1 | 1 | 4 | 4 | 4 | 5 | 6 | 4 | 4 | 1 |
| Vedder-Weiss 2012 | -0.57 | 0.06 | 0.0040 | Democratic vs. Traditional Schools | 3 | 1 | 5 | 3 | 1 | 1 | 1 | 4 | 4 | 4 | 5 | 6 | 4 | 4 | 1 |
| Vedder-Weiss 2013 | 0.41 | 0.05 | 0.0026 | Teacher Goal Orientation | 3 | 2 | 3 | 1 | 1 | 1 | 1 | 4 | 4 | 4 | 5 | 6 | 4 | 4 | 1 |
| Vedder-Weiss 2013 | 0.30 | 0.05 | 0.0025 | Teacher Goal Orientation | 3 | 2 | 5 | 1 | 1 | 1 | 1 | 4 | 4 | 2 | 5 | 6 | 4 | 4 | 1 |
| Wolf 2008 | 0.03 | 0.16 | 0.0241 | Inquiry | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 4 | 5 | 2 | 0 | 2 | 0 | 0 | 0 |
| Wolf 2008 | 0.15 | 0.09 | 0.0083 | Inquiry | 1 | 1 | 1 | 3 | 1 | 1 | 1 | 4 | 5 | 2 | 0 | 2 | 0 | 0 | 0 |
| Wolf 2008 | 0.23 | 0.16 | 0.0242 | Inquiry | 1 | 1 | 3 | 3 | 1 | 1 | 1 | 4 | 5 | 2 | 0 | 2 | 0 | 0 | 0 |
| Wu 2007 | 0.90 | 0.28 | 0.0799 | SC vs. TC digital learning environment | 3 | 1 | 2 | 3 | 0 | 1 | 1 | 4 | 3 | 3 | 1 | 3 | 3 | 4 | 1 |
| Yoon 2009 | 0.41 | 0.16 | 0.0253 | Science as Inquiry | 1 | 1 | 3 | 1 | 0 | 1 | 1 | 4 | 5 | 2 | 1 | 2 | 0 | 0 | 1 |
| Zepeda 2015 | 0.85 | 0.30 | 0.0919 | Problem-solving practice | 1 | 2 | 2 | 4 | 0 | 1 | 1 | 4 | 3 | 2 | 1 | 2 | 3 | 0 | 0 |
| Zepeda 2015 | 0.57 | 0.25 | 0.0650 | Problem-solving practice | 1 | 2 | 3 | 4 | 0 | 1 | 1 | 4 | 3 | 2 | 1 | 2 | 3 | 0 | 0 |
| Zhang 2015 | 0.61 | 0.09 | 0.0090 | Rubric + FB vs. total points | 1 | 1 | 6 | 3 | 0 | 1 | 1 | 2 | 3 | 2 | 1 | 2 | 2 | 0 | 0 |
| Zhang 2015 | 0.36 | 0.09 | 0.0080 | Rubric vs. total points | 1 | 1 | 6 | 3 | 0 | 1 | 1 | 2 | 3 | 2 | 1 | 2 | 2 | 0 | 0 |

Appendix C

Overview of Included Studies

Table C1

Overview of Included Studies

| Study | Sample size | Country | Predictor category | Engagement type* | Method |
|--------------------------|-------------|--------------|--------------------|------------------|--------------------|
| Akcay & Ergin, 2008 | 609 | US (IA) | Instructional | A | Quasi-experimental |
| Aktamis & Ergin, 2008 | 40 | Turkey | Instructional | A | Quasi-experimental |
| Al Khauri, 2007 | 1636 | Oman | Class | C | Correlational |
| Alsup 2015 | 158 | US (MI) | Instructional | A | Quasi-experimental |
| Bathgate et al. 2013 | 252 | US (PA) | Instructional | A | Single group |
| Banaweh et al. 2012 | 357 | Jordan | Instructional | A+C | Quasi-experimental |
| Benjamin 2014 | 69 | West Indies | Instructional | B+A | Quasi-experimental |
| Bilgin 2006 | 55 | Turkey | Social | A | Quasi-experimental |
| Blanchard et al. 2015 | 1808 | US (TX) | Instructional | A | Quasi-experimental |
| Bowling et al. 2013 | 856 | US | Technology | A | Correlational |
| Bozdoğan & Yalçın, 2009 | 31 | Turkey | Class | A | Single group |
| Brown et al. 2013 | 170 | US | Technology | A | Single group |
| Cetin-Dindar 2015 | 243 | Turkey | Class | A | Correlational |
| Chen & Howard. 2010 | 273 | US | Instructional | A | Single group |
| Chen & Yang, 2006 | 76 | Taiwan | Instructional | C | Quasi-experimental |
| Chen 2014 | 170 | Taiwan | Instructional | C | Quasi-experimental |
| Chen et al. 2015 | 50 | US | Social | A, C | Experimental |
| Cheng et al. 2014 | 132 | US | Technology | A,C | Quasi-experimental |
| Cirik 2014 | 1375 | Istanbul | Social | C | Correlational |
| Degenhart 2007 | 1225 | US (TX) | Instructional | A | Single group |
| Dettweiler et al. 2015 | 84 | Germany | Instructional | C | Quasi-experimental |
| Doll et al. 2010 | 1019 | US (NE) | Social | A | Correlational |
| Furtak & Kunter 2012 | 48 | Germany | Class | B,A | Experimental |
| Genç 2015 | 30 | Turkey | Instructional | A | Single group |
| Glenn & Wingenbach, 2015 | 84 | Guatemala | Instructional | A | Single group |
| Grolnick et al. 2007 | 90 | US | Class | C, A+B+C | Experimental |
| Haugen 2013 | 45 | US (Midwest) | Instructional | A | Quasi-experimental |
| Hidiroğlu 2014 | 900 | Turkey | Class | A,B,C | Correlational |
| Hong 2010 | 96 | Taiwan | Instructional | A | Quasi-experimental |
| Hope 2012 | 20 | US (Midwest) | Instructional | B+A | Single group |
| Horak 2013 | 443 | US (Mid- | Instructional | A | Quasi-experimental |

| | | | | | |
|--------------------------------|--------|--------------------------------|---------------------------|-------|--------------------|
| | | Atlantic) | | | |
| Işık & Gücüm 2013 | 70 | Turkey | Instructional | A, C | Quasi-experimental |
| Kahraman & Sungur 2013 | 977 | Turkey | Class | A,C | Correlational |
| Kaloti-Hallak et al. 2015 | 59 | Israel & Palestine | Instructional | A | Single Group |
| Kanter & Konstantopoulos 2010 | 197 | US (Midwest) | Instructional | A | Correlational |
| Kingir et al. 2013 | 802 | Turkey | Class | A,C | Correlational |
| Kiran 2010 | 1932 | Turkey | Instructional | C | Correlational |
| Köse et al. 2010 | 68 | Turkey | Social | B+A | Quasi-experimental |
| Kulo 2011 | 101 | US (PA) | Technology | A | Single Group |
| Kurbanoğlu & Nefes 2015 | 70 | Turkey | Instructional | A | Quasi-experimental |
| Larson 2014 | 219 | US (Midwest) | Instructional | A+C | Quasi-experimental |
| Linnenbrink-Garcia et al. 2013 | 99 | US | Class, Social | A | Correlational |
| Little 2015 | 34 | US (TX) | Technology | C | Experimental |
| Liu 2014 | 58 | Spain | Instructional | A | Single Group |
| Long & Fraser 2015 | 367 | US (TX&CA) | Instructional | A,B,C | Quasi-experimental |
| Luckay 2010 | 1955 | South Africa | Class, Social | A | Correlational |
| McConney et al. 2014 | 10,437 | Australia, New Zealand, Canada | Instructional | A | Quasi-experimental |
| Mo 2008 | 8544 | US | Instructional | B | Correlational |
| Moote et al. 2013 | 73 | Scotland | Instructional | A,C | Correlational |
| Nelson & DeBacker 2008 | 253 | US (South) | Social | C | Correlational |
| Ng et al. 2015 | 732 | Singapore | Class, Social | A,C | Correlational |
| Nugent et al. 2010 | 122 | US (NE) | Technology | A,C | Single Group |
| Ochsendorf et al. 2006 | 1891 | US (MD) | Instructional | B,C | Correlational |
| Odom et al. 2011 | 294 | US (MO) | Instructional, Technology | B | Correlational |
| O'Leary 2011 | 45 | Ireland | Instructional | B | Single Group |
| Osborne et al. 2013 | 745 | UK | Instructional | A,C | Correlational |
| Özkal 2007 | 1152 | Turkey | Class | C | Correlational |
| Pamuk 2014 | 3281 | Turkey | Class | A,C | Correlational |
| Park et al. 2009 | 234 | Korea | Technology | A | Single Group |
| Peng 2009 | 255 | US (OH) | Technology | A | Quasi-experimental |
| Saad 2014 | 112 | US (ND) | Instructional | A | Single Group |
| Serin et al. 2015 | 43 | Turkey | Technology | A | Experimental |
| Skinner et al. 2012 | 310 | US (PNW) | Class | B+A | Correlational |
| Smart 2014 | 223 | US | Class, Social | A,C | Correlational |

| | | | | | |
|-------------------------------|------|-----------------------|---------------|--------|--------------------|
| | | (Southeast) | | | |
| Spearman & Watt 2013 | 52 | Australia | Social | A | Correlational |
| Sungur & Güngören, 2009 | 900 | Turkey | Class | C | Correlational |
| Swarat et al., 2012 | 533 | US (Midwest) | Instructional | B+A | Single Group |
| Tapola 2013 | 52 | Finland | Instructional | A | Single Group |
| Toprac 2008 | 44 | US (TX) | Technology | A | Single Group |
| Türkmen 2009 | 97 | Turkey | Technology | A | Quasi-experimental |
| Vedder-Weiss & Fortus 2011 | 1181 | Israel | Class | C, B+C | Quasi-experimental |
| Vedder-Weiss & Fortus 2012 | 658 | Israel | Class | C, B+C | Quasi-experimental |
| Vedder-Weiss & Fortus 2013 | 1614 | Israel | Class | C, B+C | Correlational |
| Wolf & Fraser 2008 | 165 | US (NY) | Instructional | A,B,C | Quasi-experimental |
| Wu & Huang 2007 | 54 | Taiwan | Instructional | C | Quasi-experimental |
| Yoon 2009 | 166 | Korea | Instructional | C | Correlational |
| Zepeda 2015 | 46 | US | Instructional | A,C | Experimental |
| Zhang & Mislak, 2015 | 136 | US (Midwest) | Instructional | A+C | Quasi-experimental |
| Zheng 2012 | 75 | US (Southeast) | Technology | A+B+C | Experimental |

*Engagement Type: A= affective, B= behavioral, C= cognitive

Appendix D

Descriptive Statistics for Included Studies

Table D1

Descriptive Statistics for Included Studies

| Source Characteristics | | | | |
|--|--------------------------------|---------|--------------------|---------|
| Variable | Number of studies (<i>k</i>) | | Percent of studies | |
| Published | | | | |
| No | 21 | | 26.6% | |
| Yes | 58 | | 73.4% | |
| Peer-reviewed | | | | |
| No | 27 | | 34.2% | |
| Yes | 52 | | 67.6% | |
| Study Characteristics | | | | |
| Variable | Point estimates | | Studies | |
| | <i>n</i> | Percent | <i>k</i> | Percent |
| Predictor Classification: Type | | | | |
| Instructional Method | 57 | 36.1% | 40 | 48.2% |
| Technology | 15 | 9.5% | 13 | 15.7% |
| Class Characteristics | 60 | 37.9% | 20 | 24.1% |
| Social Characteristics | 26 | 16.5% | 10 | 12% |
| Predictor Classification: SDT | | | | |
| Autonomy | 94 | 59.4% | 22 | 23.9% |
| Competence | 29 | 18.4% | 21 | 22.8% |
| Relatedness | 35 | 22.2% | 49 | 53.3% |
| Engagement Type | | | | |
| Behavioral | 10 | 6.3% | 7 | 6.7% |
| Affective | 84 | 53.2% | 56 | 53.3% |
| Cognitive | 49 | 31% | 31 | 29.5% |
| Behavioral + Affective | 6 | 3.8% | 3 | 2.9% |
| Behavioral + Cognitive | 3 | 1.9% | 3 | 2.9% |
| Affective + Cognitive | 4 | 2.5% | 3 | 2.9% |
| All three | 2 | 1.3% | 2 | 1.9% |
| Grade Level | | | | |
| 5 th (10-11 years old) | 3 | 1.9% | 3 | 3.8% |
| 6 th -8 th (11-14 years old) | 110 | 70% | 52 | 65.8% |
| 9 th (14-15 years old) | 13 | 8.2% | 7 | 8.9% |
| Mix | 32 | 20.3% | 17 | 21.5% |
| School Structure | | | | |
| Not specified | 55 | 34.8% | 26 | 32.9% |
| Elementary | 2 | 1.3% | 2 | 2.5% |
| Middle School | 44 | 27.8% | 21 | 26.6% |
| Junior High | 7 | 4.4% | 6 | 7.6% |
| K-8 | 37 | 23.4% | 16 | 20.3% |
| High School | 3 | 1.9% | 2 | 2.5% |
| Other/Mix | 10 | 6.3% | 6 | 7.6% |

| | | | | |
|--|-----|-------|----|-------|
| School Type | | | | |
| Not specified | 43 | 27.2% | 21 | 26.6% |
| Public | 97 | 61.4% | 46 | 58.2% |
| Private | 6 | 3.8% | 4 | 5.1% |
| Mix | 12 | 7.6% | 8 | 10.1% |
| School Setting | | | | |
| Not specified | 83 | 52.5% | 45 | 57% |
| Rural | 5 | 0.6% | 5 | 6.3% |
| Suburban | 28 | 17.7% | 8 | 10.1% |
| Urban | 24 | 15.2% | 13 | 16.5% |
| Mix | 18 | 11.4% | 8 | 10.1% |
| Socio-economic Status | | | | |
| Not specified | 101 | 63.9% | 44 | 55.7% |
| Low (> 60% FRL) | 9 | 5.7% | 6 | 7.6% |
| Average (35-59% FRL) | 7 | 4.4% | 3 | 3.8% |
| High (<35% FRL) | 19 | 12% | 13 | 16.5% |
| Mix | 22 | 13.9% | 13 | 16.5% |
| Study Methodology | | | | |
| Correlation/Regression/SEM | 77 | 48.7% | 23 | 29.1% |
| Single Group (Repeated Measures) | 21 | 13.3% | 18 | 22.8% |
| Quasi-experimental | 45 | 28.5% | 31 | 39.2% |
| Experimental | 15 | 9.5% | 7 | 8.9% |
| Reliability | | | | |
| Not reported | 6 | 3.8% | 5 | 6.3% |
| References external instrument | 4 | 2.5% | 4 | 5% |
| References external instrument reliability | 21 | 13.3% | 11 | 13.9% |
| Internal reliability <.70 | 28 | 17.7% | 13 | 16.5% |
| Internal reliability > .70 | 99 | 62.7% | 46 | 58.2% |
| Validity | | | | |
| Face/content validity (assessed by investigator) | 8 | 5.1% | 4 | 4.9 |
| Face/content validity (assessed by study) | 17 | 10.8% | 3 | 3.7 |
| References external instrument | 72 | 45.6% | 45 | 54.9 |
| References external instrument validity | 20 | 12.7% | 14 | 17.1 |
| Internal validity measures (e.g., EFA, CFA) | 41 | 25.9% | 16 | 19.5 |
| Geographic Location | | | | |
| United States | 72 | 45.6% | 35 | 44.3 |
| Outside of US | | | | |
| Australia | 1 | 1.2% | 1 | 2.3 |
| Finland | 1 | 1.2% | 1 | 2.3 |
| Germany | 7 | 8.1% | 2 | 4.5 |
| Guatemala | 1 | 1.2% | 1 | 2.3 |
| Ireland | 1 | 1.2% | 1 | 2.3 |
| Israel | 6 | 7% | 3 | 6.8 |
| Jordan | 1 | 1.2% | 1 | 2.3 |
| Korea | 2 | 2.3% | 2 | 4.5 |
| Oman | 1 | 1.2% | 1 | 2.3 |
| Scotland | 2 | 2.3% | 1 | 2.3 |
| Singapore | 6 | 7% | 1 | 2.3 |

| | | | | |
|----------------|----|-------|----|------|
| South Africa | 6 | 7% | 1 | 2.3 |
| Spain | 1 | 1.2% | 1 | 2.3 |
| Taiwan | 4 | 4.7% | 4 | 9 |
| Turkey | 39 | 45.3% | 18 | 40.9 |
| United Kingdom | 2 | 2.3% | 1 | 2.3 |
| West Indies | 2 | 2.3% | 1 | 2.3 |

Note. Descriptive statistics are based on 79 studies and 158 point estimates. For some variables, the number of studies sums to greater than 79 because a particular study might contribute two different outcomes (e.g., both affective and cognitive).

Appendix E

Selection and Use of Predictor and Criterion Variables by Study

Table E1

Selection and Use of Predictor and Criterion Variables by Study

| Study | Predictor Variable | Criterion Variable & Type of Engagement |
|------------------------|--|---|
| Akçay & Ergin, 2008 | Science-Technology-Society vs. Textbook | Combined self-designed science attitudes scale across 6 th -9 th grades (affective) |
| Aktamis & Ergin, 2008 | Science process skills instruction | Self-designed attitudes scale (affective) |
| Al Khaursi, 2007 | Perceptions of a learning classroom assessment environment | Mastery goal orientation subscale of self-designed student questionnaire (cognitive) |
| Alsup 2015 | Viewing interviews of STEM professionals | Subject attitudes subscale from STEM Semantics Survey (affective) |
| Banaweh et al. 2012 | Whole brain teaching | Student Motivation Toward Science Learning Questionnaire (SMSLQ) (cognitive+affective) |
| Bathgate et al. 2013 | Analyzing vs. Action Consuming new knowledge vs. Action | Self-designed motivational scales (affective) |
| Benjamin 2014 | Performance assessment | Self-designed Science Attitude Scale (SAS) (affective) Self-designed Science Motivation Questionnaire (SMQ) (behavioral + affective) |
| Bilgin 2006 | Cooperative learning vs. teacher demonstration | ASTS scale (affective) |
| Blanchard et al. 2015 | Innovation Club after school inquiry intervention | Two items from self-designed scale for science and engineering interest (affective) |
| Bowling et al. 2013 | Playing Flash game <i>Disease Defenders</i> | Self-designed science attitudes scale (affective) |
| Bozogan & Yalçın, 2009 | Participating in exhibitions and activities at learning center | Self-designed interest scale (affective) |
| Brown et al. 2013 | Participating in online problem-based learning | Self-designed instrument (affective) |

| | | |
|--------------------------|---|--|
| Cetin-Dindar 2015 | Perceptions of a constructivist learning environment (CLES) | Science Motivation Questionnaire (SMQ) (affective) |
| Chen et al. 2010 | Live simulation | Combined scientific inquiry attitude, scientific attitude adoption, and science lesson enjoyment subscales of TOSRA (affective) |
| Chen & Yang, 2006 | Project-based learning | Combined Self-Directed Science Learning Readiness Scale (SDSLRS) and Students' Motivation toward Science Learning (SMTSL) scales (cognitive) |
| Chen 2014 | Scaffolding in an e-learning environment | Academic Motivation Scale (AMS) (cognitive) |
| Chen et al. 2015 | Collaborative vs. individual digital game play | Intrinsic motivation subscale of MSLQ (cognitive) Task value subscale of MSLQ (affective) |
| Cheng et al. 2014 | Playing web game Humunology vs. learning through online content | Self-designed perceived usefulness scale (affective) Self-designed peer learning and help-seeking (cognitive) |
| Cirik 2014 | Perceived teacher social support from Child and Adolescent Social Support Scale (CASSS) | Combined 6 motivation scales from MSLQ (cognitive) |
| Degenhart 2007 | Interaction with NSF Fellows in Classroom | Combined STEM beliefs and STEM interests surveys for the science group (affective) |
| Dettweiler et al. 2015 | Outdoor vs. Indoor Education | Self-determination index (cognitive) |
| Doll et al. 2010 | ClassMaps Survey (MT-My Teacher Scale) | Student Engagement Survey (SES) (affective) |
| Furtak & Kunter 2012 | Autonomy support (procedural, cognitive, or procedural + cognitive) | Combined interest and value subscales of IMI (affective) Self-designed scale for motivated behavior (behavioral) |
| Genç2015 | Scientific study (research, interaction with professionals, discussion with peers) | SAS (affective) |
| Glenn & Wingenbach, 2015 | Junior Master Gardeners' Program | Survey of Students' Attitudes Toward Science (affective) |
| Grolnick et al. 2007 | Participation in After-School Program promoting student to student interaction, autonomy, and relatedness | Relative Autonomy Index (RAI) (cognitive); science data from Engagement survey (all three engagement types) |
| Haugen 2013 | Project-based learning | Combined attitude to inquiry, adoption of scientific attitudes, and enjoyment of science subscales of TOSRA (affective) |

| | | |
|-------------------------------|---|--|
| Hidirođlu 2014 | Survey of Classroom Goal Structure (autonomy support subscale) | Affective scale from Engagement Questionnaire (EQ) (affective); behavior scale from EQ (behavioral); combined cognitive and agentic engagement (cognitive) |
| Hong 2010 | Collaborative science intervention supporting inquiry and cooperative learning | Attitudes scale derived from Secondary School Student Questionnaire (SSSQ) & Waering Attitudes Toward Science Protocol (WASP) (affective) |
| Hope 2012 | SciJourn program participation in Mary Connor's Classroom (9 th graders) | Youth Engagement with Science and Technology (YEST) (behavioral + affective) |
| Horak 2013 | Problem-based learning | Appeal and Meaning subscales of the Student Perceptions Of Class Questionnaire (SPOCQ) (affective) |
| Iřik & Gęcüm 2013 | Project-based learning | Strategy use and self-regulated learning subscales from MSLQ (cognitive) |
| Kahraman & Sungur 2013 | Teacher's mastery goals from PALS | Intrinsic value subscale from MSLQ (affective) |
| Kaloti-Hallak et al. 2015 | Lego robotics competition | Task value subscale from MSLQ (affective) |
| Kanter & Konstantopoulos 2010 | Supporting students explaining concepts to one another Supporting students analyzing data | Combined metacognition scale from MSLQ and mastery approach goals from AGQ (cognitive) |
| Kingir et al. 2013 | Perceptions of personal relevance, uncertainty, critical voice, shared control, and student negotiation | Intrinsic motivation scale (affective) |
| Kiran 2010 | Sources of Science Self-Efficacy (SSSE) Verbal persuasion sub-scale | Combined students' post-perceptions of value, relevance, and interest in science (affective) |
| Köse et al. 2010 | Cooperative learning | Intrinsic value subscale of MSLQ (affective) |
| Kulo 2011 | GIS-supported inquiry | Combined mastery approach subscale of AGQ and self-regulation subscale of MSLQ (cognitive) |
| | | Mastery approach subscale of Achievement Goal Questionnaire (AGQ) (cognitive) |
| | | Attitude Scale Toward Science (ASTS) (affective) |
| | | Combined data from all achievement proficiency levels of the Energy Unit Science and Technology Survey subscale of Science and Technology Survey (affective) |

| | | |
|-------------------------------|--|--|
| Kurbanoğlu & Nefes 2015 | Context-based questions | Science Attitudes Scale (affective) |
| Larson 2014 | Generative Vocabulary Matrix from Engagement Model of Academic Literacy for Learning (EngageALL) | Engagement subscale from Experience Sampling Form (ESF) (cognitive + affective) Combined intrinsic interest and positive affect subscales from ESF (affective) |
| Linnebrink-Garcia et al. 2013 | Perceived choice, teacher approachability, connections to real life, supportsStiutVerb for involvement, group work | Triggered situational interest sub-scale from the Situational Interest (SI) survey (affective) |
| Little 2015 | Digital game-based learning vs. lab work | Cognitive subscale of RAPS-SM (cognitive) |
| Long & Fraser 2015 | Specific vs. general curricula | Investigation/Involvement subscale of Outcomes-related learning environment scale (ORLES) (behavioral) Task orientation subscale of ORLES (cognitive) Enjoyment of science subscale of ORLES (affective) |
| Luckay 2010 | Working with ideas Respect for differences Personal relevance Collaboration Critical Voice Uncertainty | Combined adaptation of the enjoyment of science subscales of TOSRA for different SES levels (affective) |
| McConney et al. 2014 | Inquiry vs. Non-inquiry learning from PISA | Combined general interest, enjoyment, personal value, and general value from PISA 2006 across geographic locations (affective) |
| Mo 2008 | Opportunity to Learn (OTL1)from PISA 2006 for emphasizing instruction in scientific investigation Opportunity to Learn (OTL2) emphasizing connections between science and society | Combined composites from PISA 2006 for participating in scientific investigation activities and for participating in connecting science to society (behavioral) |

| | | |
|------------------------|--|---|
| Moote et al. 2013 | CREST student-run science project | Combined self-regulation and cognitive strategy use subscales of MSLQ with self-regulated learning sub-scale of the Five Component Scale for Self-Regulation (FCSSR) (cognitive) Combined intrinsic value subscale of MSLQ with personal relevance (IMPR) subscale of the Science Motivation Questionnaire (SMQ) (affective) |
| Nelson & DeBacker 2008 | Class involvement from the Classroom Environment Scale; Class belongingness from the Psychological Sense of School Membership (PSSM) | Approaches to Learning (ATL) Questionnaire (cognitive) |
| Ng et al. 2015 | Combined perceptions of teacher autonomy support from Learning Climate Questionnaire (LCQ) and autonomy subscale of Psychological Needs Questionnaire Competence Relatedness | Task value subscale of MSLQ (affective) Learning strategies subscale of MSLQ (cognitive) |
| Nugent et al. 2010 | Participation in short-term robotics intervention | Science task value subscale from MSLQ-adapted scale (affective) Problem approach learning strategies subscale from MSLQ-adapted scale (cognitive) |
| Ochsendorf et al. 2006 | Inquiry/activity-based method | Basic Learning Engagement Scale (behavioral) Combined Advanced Learning Engagement Scale and Mastery Goal Orientation Scale (cognitive) |
| Odom et al. 2011 | Student-centered learning environments Teacher-centered learning environments Computer Usage | Attitudes toward science sub-scale of Science Achievement Influences Survey Version 2 (SAIS v.2) (affective) |
| O'Leary 2011 | Universally vs. Commercially-designed worksheets | Effort and Persistence in Learning (EPL) subscale of the Student Approaches to Learning Survey (behavioral) |

| | | |
|-------------------------|---|--|
| Osborne et al. 2013 | Scientific argumentation | Combined 7 th and 9 th grade data for interest in learning science and science enjoyment (affective) Combined 7 th and 9 th grade data for task orientation (cognitive) |
| Özkal 2007 | Personal relevance, critical voice, student negotiation, and shared control subscales from Classroom Learning Environment Survey (CLES) | Meaningful learning orientation subscale from the Learning Approaches Questionnaire (LAQ) (cognitive) |
| Palmer 2009 | Experiments; Demonstrations | Tukey's HSD test results from Experiment vs. Note-Taking and Demonstration vs. Note-Taking (affective) |
| Pamuk 2014 | Personal Relevance, Critical Voice, and Student Negotiation from Classroom Learning Environment Survey (CLES) | Task value subscale from MSLQ (affective) Combined mastery approach subscale from AGQ and self-regulated learning subscale from MSLQ (cognitive) |
| Park et al. 2009 | Interactive Computer Technology (ICT) Involvement | Total attitudes toward science score across clusters (affective) |
| Peng 2009 | STEAM games and interaction with STEAM fellow | Combined Value of Science and Motivation in Science subscales (affective) |
| Liu 2014 | Participating in hands-on inquiry activities | Combined two items about interest and fun (affective) |
| Saad 2014 | NEAR-space ballooning project | One item about astronomy interest (affective) |
| Serin et al. 2015 | Computer-assisted instruction | Attitudes Scale Toward Science (ASTS) (affective) |
| Skinner et al. 2012 | Autonomy and Competence in Garden-Based Education | Classroom Engagement Survey (behavioral + cognitive) |
| Smart 2014 | Perceptions of the teacher (leadership, helping/friendly, understanding, student freedom, strict, admonishing, dissatisfied) | Partial correlations with Mastery Orientation subscale of PALS (cognitive) Partial correlations with Task Value for Learning Science subscale of PALS (cognitive) |
| Spearman & Watt 2013 | Teacher relatedness perceptions from Student-Reported Teacher Style Scale (STRS) | Intrinsic Interest Value from Student Motivation Questionnaire (SMQ) (affective) |
| Sungur & Güngören, 2009 | Survey of Classroom Goal Structure | Combined goal orientation and strategy use scales from ATL Questionnaire (cognitive) |

| | | |
|----------------------------|--|--|
| Swarat et al., 2012 | Technology and Hands-on activities vs. purely cognitive activities | Self-designed 2-item scale (behavioral + affective) |
| Tapola 2013 | Abstract vs. Concrete activities | Situational Interest scale (affective) |
| Toprac 2008 | PBL digital game (Alien Rescue III) | Combined attainment value, intrinsic value, and utility value subscales of the Dimensions of Continuing Motivation in Science (DCMS) scale (affective) |
| Türkmen 2009 | Technology-based inquiry approach | ASTS scale (affective) |
| Vedder-Weiss & Fortus 2011 | Democratic vs. Traditional Schools | Combined grades 5-8 data from personal mastery goal sub-scale of Patterns of Adaptive Learning (PALS) (cognitive) Adapted engagement scale (cognitive + behavioral) |
| Vedder-Weiss & Fortus 2012 | Democratic vs. Traditional Schools | Combined grades 5-8 data from personal mastery goal sub-scale of Patterns of Adaptive Learning (PALS) (cognitive) Adapted engagement scale (cognitive + behavioral) |
| Vedder-Weiss & Fortus 2013 | Patterns of Adaptive Learning Scale (PALS) for perceived teacher goal orientations | Combined grades 5-8 data from personal mastery goal sub-scale of Patterns of Adaptive Learning (PALS) (cognitive) Adapted engagement scale (cognitive + behavioral) |
| Wolf & Fraser 2008 | Inquiry vs. non-inquiry learning | Combined involvement, investigation, and cooperation subscales of What is Happening in this Class? (WIHIC) Questionnaire (behavioral) Task orientation sub-scale of WIHIC (cognitive) Enjoyment of Science Lessons subscale of Test of Science-Related Attitudes (TOSRA) (affective) |
| Wu & Huang 2007 | Student-centered vs. Teacher-centered Digital Learning Environments | Student attitudes scale (affective) |
| Yoon 2009 | Assessing Student Understanding in Science Inventory (Science-as-inquiry) subscale | MSLQ (intrinsic value scale) for goal orientation (cognitive) |

| | | |
|----------------------|--|--|
| Zepeda 2015 | Problem-solving practice | Task value subscale from MSLQ (affective) Combined mastery approach subscale from AGQ and metacognitive awareness inventory (MAI) (cognitive) |
| Zhang & Mislak, 2015 | Assessment practices: rubric vs. total points and rubric + feedback vs. total points | SMTSL (cognitive + affective) |
| Zheng 2012 | Digital game | Combined four scales from Game Flow Questionnaire (all three engagement types) |

Appendix F

Studies Included in the Meta-Analysis: References

- Akcay, H., Yager, R. E., Iskander, S. M., & Turgut, H. (2010). Change in student beliefs about attitudes toward science in grades 6-9. *Asia-Pacific Forum on Science Learning and Teaching, 11*(1).
- Aktamis, H., & Ergin, O. (2008). The effect of scientific process skills education on students' scientific creativity, science attitudes and academic achievements. *Asia-Pacific Forum on Science Learning and Teaching, 9*(1), 1-21.
- Al Kharusi, H. A. (2007). *Effects of teachers' assessment practices on ninth grade students' perceptions of classroom assessment environment and achievement goal orientations in Muscat science classrooms in the Sultanate of Oman* (Doctoral dissertation). Retrieved from Dissertation Abstracts database. (Order No. 3274102).
- Alsop, P. R. (2015). *The effect of video interviews with STEM professionals on STEM-subject attitude and STEM-career interest of middle school students in conservative Protestant Christian schools* (Doctoral dissertation). Retrieved from Dissertation Abstracts database. (Order No. 3685438)
- Bathgate, M. E., Schunn, C. D., & Correnti, R. (2014). Children's motivation toward science across contexts, manner of interaction, and topic. *Science Education, 98*(2), 189-215. <http://dx.doi.org/10.1002/sc.21095>
- Bawaneh, A. K. A., Zain, A. N. M., Saleh, S., & Abdullah, A. G. K. (2012). Using Herrmann Whole Brain Teaching Method to enhance students' motivation towards science learning. *Journal of Turkish Science Education, 9*(3), 3-22.

- Benjamin, A. (2014). *The impact of performance assessment on students' interest and academic performance in science* (Thesis).
- Bilgin, I. (2006). The effects of hands-on activities incorporating a cooperative learning approach on eighth grade students' science process skills and attitudes toward science. *Journal of Baltic Science Education*, (9), 27-37.
- Blanchard, S., Judy, J., Muller, C., Crawford, R. H., Petrosino, A. J., White, C. K., ... Wood, K. L. (2015). Beyond blackboards: Engaging underserved middle school students in engineering. *Journal of Pre-College Engineering Education Research*, 5(1), 1-14. <http://dx.doi.org/10.7771/2157-9288.1084>
- Bowling, K. G., Klisch, Y., Wang, S., & Beier, M. (2013). Examining an online microbiology game as an effective tool for teaching the scientific process. *Journal of Microbiology & Biology Education*, 14(1), 58-65.
<http://dx.doi.org/10.1128/jmbe.v14i1.505>
- Bozdogan, A. E., & Yalçın, N. (2009). Determining the influence of a science exhibition center training program on elementary pupils' interest and achievement in science. *Eurasia Journal of Mathematics, Science and Technology Education*, 5(1), 27-34.
- Brown, S. W., Lawless, K. A., & Boyer, M. A. (2013). Promoting positive academic dispositions using a web-based PBL environment: The GlobalEd 2 project. *Interdisciplinary Journal of Problem-Based Learning*, 7(1), 67-90.
<http://dx.doi.org/10.7771/1541-5015.1389>

- Cetin-Dindar, A. (2015). Student motivation in constructivist learning environment. *Eurasia Journal of Mathematics, Science & Technology Education*, 12(2), 233-247. <http://dx.doi.org/10.12973/eurasia.2016.1399a>
- Chen, C. H. (2014). An adaptive scaffolding e-learning system for middle school students' physics learning. *Australasian Journal of Educational Technology*, 30(3), 342-355. <http://dx.doi.org/10.14742/ajet.v30i3.430>
- Chen, C. H., & Howard, B. (2010). Effect of live simulation on middle school students' attitudes and learning toward science. *Journal of Educational Technology & Society*, 13(1), 133-139.
- Chen, C. H., Wang, K. C., & Lin, Y. H. (2015). The comparison of solitary and collaborative modes of game-based learning on students' science learning and motivation. *Journal of Educational Technology & Society*, 18(2), 237-248.
- Chen, Y. K., & Yang, K. Y. (2006, November). Using problem-based learning teaching model to promote the self-directed science learning readiness and science learning motivation of eighth-grade students. *APER A conference*, 28-30.
- Cheng, M. T., Su, T., Huang, W. Y., & Chen, J. H. (2014). An educational game for learning human immunology: What do students learn and how do they perceive? *British Journal of Educational Technology*, 45(5), 820-833. <http://dx.doi.org/10.1111/bjet.12098>
- Cirik, I. (2015). Relationships between social support, motivation, and science achievement: Structural equation modeling. *Anthropologist*, 20(1-2), 232-242.
- Degenhart, H. S. (2007). *Relationship of inquiry-based learning elements on changes in middle school students' science, technology, engineering, and mathematics*

- (STEM) beliefs and interests (Doctoral dissertation). Retrieved from Dissertation Abstracts database. (Order No. 3270326)
- Dettweiler, U., Ünlü, A., Lauterbach, G., Becker, C., & Gschrey, B. (2015). Investigating the motivational behavior of pupils during outdoor science teaching within self-determination theory. *Frontiers in Psychology, 6*, 1-16.
<http://dx.doi.org/10.3389/fpsyg.2015.00125>
- Doll, B., Spies, R. A., Champion, A., Guerrero, C., Dooley, K., & Turner, A. (2010). The ClassMaps Survey: A measure of middle school science students' perceptions of classroom characteristics. *Journal of Psychoeducational Assessment, 28*(4), 338-348. <http://dx.doi.org/10.1177/0734282910366839>
- Furtak, E. M., & Kunter, M. (2012). Effects of autonomy-supportive teaching on student learning and motivation. *The Journal of Experimental Education, 80*(3), 284-316.
<http://dx.doi.org/10.1080/00220973.2011.573019>
- Genç, M. (2015). The effect of scientific studies on students' scientific literacy and attitude. *Ondokuz Mayıs Üniversitesi Eğitim Fakültesi Dergisi, 34*(1), 141-152.
<http://dx.doi.org/10.7822/omuefd.34.1.8>
- Glenn, A., & Wingenbach, G. (2015). Effects of the Junior Master Gardener's (JMG) curriculum on Guatemalan students' knowledge gain and attitude toward science. *Journal of International Agricultural and Extension Education, 21*(2), 64-73.
<http://dx.doi.org/10.5191/jiaee.2015.22205>
- Grolnick, W. S., Farkas, M. S., Sohmer, R., Michaels, S., & Valsiner, J. (2007). Facilitating motivation in young adolescents: Effects of an after-school program.

Journal of Applied Developmental Psychology, 28(4), 332-344.

<http://dx.doi.org/10.1016/j.appdev.2007.04.004>

- Haugen, M. I. (2013). *Comparing project-based learning to direct instruction on students' attitude to learn science* (Master's thesis). Retrieved from Dissertation Abstracts database. (Order No. 1543663)
- Hidirođlu, F. M. (2014). *The role of perceived classroom goal structures, self-efficacy, and the student engagement in seventh grade students' science achievement* (Unpublished master's thesis). Middle East Technical University, Turkey.
- Hope, J. M. G. (2012). *Exploring the nature of high school student engagement with science and technology as an outcome of participation in science journalism* (Doctoral dissertation). Retrieved from Dissertation Abstracts database. (Order No. 3507699)
- Hong, Z. R. (2010). Effects of a collaborative science intervention on high achieving students' learning anxiety and attitudes toward science. *International Journal of Science Education*, 32(15), 1971-1988.
- <http://dx.doi.org/10.1080/09500690903229304>
- Horak, A. (2013). *The effect of using problem-based learning in middle school gifted science classes on student achievement and students' perceptions of classroom quality* (Doctoral dissertation). Retrieved from Dissertation Abstracts database. (Order No. 3591082)
- Iřik, Ö., & Gcm, B. (2013). The effect of project based learning approach on elementary school students' motivation toward science and technology course. *Hacettepe niversitesi Eđitim Fakltesi Dergisi*, 28(28-3).

- Kahraman, N., & Sungur, S. (2013). Antecedents and consequences of middle school students' achievement goals in science. *The Asia-Pacific Education Researcher*, 22(1), 45-60. <http://dx.doi.org/10.1007/s40299-012-0024-2>
- Kaloti-Hallak, F., Armoni, M., & Ben-Ari, M. M. (2015). Students' attitudes and motivation during robotics activities. *Proceedings of the Workshop in Primary and Secondary Computing Education*, 102-110. <http://dx.doi.org/10.1145/2818314.2818317>
- Kanter, D. E., & Konstantopoulos, S. (2010). The impact of a project-based science curriculum on minority student achievement, attitudes, and careers: The effects of teacher content and pedagogical content knowledge and inquiry-based practices. *Science Education*, 94(5), 855-887. <http://dx.doi.org/10.1002/sce.20391>
- Kingir, S., Tas, Y., Gok, G., & Vural, S. S. (2013). Relationships among constructivist learning environment perceptions, motivational beliefs, self-regulation and science achievement. *Research in Science & Technological Education*, 31(3), 205-226. <http://dx.doi.org/10.1080/02635143.2013.825594>
- Kiran, D. (2010). *A study on sources and consequences of elementary students' self-efficacy beliefs in science and technology course* (Unpublished doctoral dissertation). Middle East Technical University, Turkey.
- Köse, S., Sahin, A., Ergün, A., & Gezer, K. (2010). The effects of cooperative learning experience on eighth grade students' achievement and attitude toward science. *Education*, 131(1), 169-180.
- Kulo, V. (2011). *Design, development, and formative evaluation of a geographic information system-supported science Web-based inquiry module* (Doctoral

- dissertation). Retrieved from Dissertation Abstracts database. (Order No. 3456166)
- Kurbanoglu, N. İ., & Nefes, F. K. (2015). Effect of context-based questions on secondary school students' test anxiety and science attitude. *Journal of Baltic Science Education, 14*(2), 216-226.
- Larson, S. C. (2014). Exploring the roles of the generative vocabulary matrix and academic literacy engagement of ninth grade biology students. *Literacy Research and Instruction, 53*(4), 287-325. <http://dx.doi.org/10.1080/19388071.2014.880974>
- Linnenbrink-Garcia, L., Patall, E. A., & Messersmith, E. E. (2013). Antecedents and consequences of situational interest. *British Journal of Educational Psychology, 83*(4), 591-614. <http://dx.doi.org/10.1111/j.2044-8279.2012.02080.x>
- Little, T. W. (2015). *Effects of digital game-based learning on student engagement and academic achievement* (Doctoral dissertation). Retrieved from Dissertation Abstracts database. (Order No. 3721273)
- Liu, L. R. (2014). ¿Cómo aumentar el interés por la Ciencia?: Una propuesta didáctica para alumnos de 12-15 años. *Boletín de la Real Sociedad Española de Historia Natural. Sección aula, museos y colecciones, 1*, 139-157.
- Long, C., & Fraser, B. J. (2015). Comparison of alternative sequencing of middle-school science curriculum: Classroom learning environment and student attitudes. *Curriculum and Teaching, 30*(1), 23-36. <http://dx.doi.org/10.7459/ct/30.1.03>

- Luckay, M. B. (2010). *Implementation of social constructivist learning environments in Grade 9 natural science in the Western Cape Province, South Africa* (Unpublished doctoral dissertation). University of Cape Town, Africa.
- McConney, A., Oliver, M. C., Woods-McConney, A., Schibeci, R., & Maor, D. (2014). Inquiry, engagement, and literacy in science: A retrospective, cross-national analysis using PISA 2006. *Science Education*, 98(6), 963-980.
<http://dx.doi.org/10.1002/sce.21135>
- Mo, Y. (2008). *Opportunity to learn, engagement, and science achievement: Evidence from TIMSS 2003 Data* (Doctoral dissertation). Retrieved from Dissertation Abstracts database. (Order No. DP19739)
- Moote, J. K., Williams, J. M., & Sproule, J. (2013). When students take control: Investigating the impact of the crest inquiry-based learning program on self-regulated processes and related motivations in young science students. *Journal of Cognitive Education and Psychology*, 12(2), 178-196.
<http://dx.doi.org/10.1891/1945-8959.12.2.178>
- Nelson, R. M., & DeBacker, T. K. (2008). Achievement motivation in adolescents: The role of peer climate and best friends. *The Journal of Experimental Education*, 76(2), 170-189. <http://dx.doi.org/10.3200/JEXE.76.2.170-190>
- Ng, B. L., Liu, W. C., & Wang, J. C. (2015). Student motivation and learning in mathematics and science: A cluster analysis. *International Journal of Science and Mathematics Education*, 1-18. <http://dx.doi.org/10.1007/s10763-015-9654-1>
- Nugent, G., Barker, B., Grandgenett, N., & Adamchuk, V. I. (2010). Impact of robotics and geospatial technology interventions on youth STEM learning and

attitudes. *Journal of Research on Technology in Education*, 42(4), 391-408.

<http://dx.doi.org/10.1080/15391523.2010.10782557>

Ochsendorf, R., Pyke, C., Lynch, S., & Watson, W. (2006, April). *The impact of a middle school motion and forces curriculum unit on student outcomes: Results from consecutive quasi-experimental studies*. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, San Francisco, CA.

Odom, A. L., Marszalek, J. M., Stoddard, E. R., & Wrobel, J. M. (2011). Computers and traditional teaching practices: Factors influencing middle level students' science achievement and attitudes about science. *International Journal of Science Education*, 33(17), 2351-2374. <http://dx.doi.org/10.1080/09500693.2010.543437>

O'Leary, S. (2011). The inclusive classroom: Effect of a readability intervention on student engagement and on-task behaviour within two mixed-ability science classrooms. *Science Education International*, 22(2), 145-151.

Osborne, J., Simon, S., Christodoulou, A., Howell-Richardson, C., & Richardson, K. (2013). Learning to argue: A study of four schools and their attempt to develop the use of argumentation as a common instructional practice and its impact on students. *Journal of Research in Science Teaching*, 50(3), 315-347.

<http://dx.doi.org/10.1002/tea.21073>

Özkal, K. (2007). *Scientific epistemological beliefs, perceptions of constructivist learning environment and attitude towards science as determinants of students approaches to learning* (Unpublished doctoral dissertation). Middle East Technical University, Turkey.

- Palmer, D. H. (2009). Student interest generated during an inquiry skills lesson. *Journal of Research in Science Teaching*, 46(2), 147-165.
<http://dx.doi.org/10.1002/tea.20263>
- Pamuk, S. (2014). *Multilevel analysis of students' science achievement in relation to constructivist learning environment perceptions, epistemological beliefs, self-regulation and science teachers' characteristics* (Unpublished doctoral dissertation). Middle East Technical University, Turkey.
- Peng, L. W. (2009). *Digital science games' impact on sixth and eighth graders' perceptions of science* (Doctoral dissertation). Retrieved from Dissertation Abstracts database. (Order No. 3371595)
- Saad, M. E. (2014). *Progressing science, technology, engineering, and math (STEM) education in North Dakota with near-space ballooning* (Master's thesis). Retrieved from Dissertation Abstracts database. (UMI No. 1559998)
- Serin, O., Serin, N. B., & Özbaş, F. (2015). The effect of computer-assisted science instruction attitude towards science and the computer. *International Journal of New Trends in Arts, Sports & Science Education*, 4(3), 1-11.
- Skinner, E. A., Chi, U., & The Learning-Gardens Educational Assessment Group. (2012). Intrinsic motivation and engagement as “active ingredients” in garden-based education: Examining models and measures derived from self-determination theory. *The Journal of Environmental Education*, 43(1), 16-36.
<http://dx.doi.org/10.1080/00958964.2011.596856>

- Smart, J. B. (2014). A mixed methods study of the relationship between student perceptions of teacher-student interactions and motivation in middle level science. *RMLE Online*, 38(4), 1.
- Spearman, J., & Watt, H. M. (2013). Perception shapes experience: The influence of actual and perceived classroom environment dimensions on girls' motivations for science. *Learning Environments Research*, 16(2), 217-238.
<http://dx.doi.org/10.1007/s10984-013-9129-7>
- Sungur, S., & Güngören, S. (2009). The role of classroom environment perceptions in self-regulated learning and science achievement. *Elementary Education Online*, 8(3), 883-900.
- Swarat, S., Ortony, A., & Revelle, W. (2012). Activity matters: Understanding student interest in school science. *Journal of Research in Science Teaching*, 49(4), 515-537. <http://dx.doi.org/10.1002/tea.21010>
- Tapola, A., Veermans, M., & Niemivirta, M. (2013). Predictors and outcomes of situational interest during a science learning task. *Instructional Science*, 41(6), 1047-1064. <http://dx.doi.org/10.1007/s11251-013-9273-6>
- Toprac, P. K. (2008). *The effects of a problem-based learning digital game on continuing motivation to learn science* (Doctoral dissertation). Retrieved from Dissertation Abstracts database. (Order No. 3329870)
- Türkmen, H. (2009). An effect of technology based inquiry approach on the learning of "Earth, Sun, & Moon" subject. *Asia-Pacific Forum on Science Learning and Teaching*, 10(1).

- Vedder-Weiss, D., & Fortus, D. (2011). Adolescents' declining motivation to learn science: inevitable or not? *Journal of Research in Science Teaching*, 48(2), 199-216. <http://dx.doi.org/10.1002/tea.20398>
- Vedder-Weiss, D., & Fortus, D. (2012). Adolescents' declining motivation to learn science: A follow-up study. *Journal of Research in Science Teaching*, 49(9), 1057-1095. <http://dx.doi.org/10.1002/tea.21049>
- Vedder-Weiss, D., & Fortus, D. (2013). School, teacher, peers, and parents' goals emphases and adolescents' motivation to learn science in and out of school. *Journal of Research in Science Teaching*, 50(8), 952-988. <http://dx.doi.org/10.1002/tea.21103>
- Wolf, S. J., & Fraser, B. J. (2008). Learning environment, attitudes and achievement among middle-school science students using inquiry-based laboratory activities. *Research in Science Education*, 38(3), 321-341. <http://dx.doi.org/10.1007/s11165-007-9052-y>
- Wu, H. K., & Huang, Y. L. (2007). Ninth-grade student engagement in teacher-centered and student-centered technology-enhanced learning environments. *Science Education*, 91(5), 727-749. <http://dx.doi.org/10.1002/sce.20216>
- Yoon, C. H. (2009). Self-regulated learning and instructional factors in the scientific inquiry of scientifically gifted Korean middle school students. *Gifted Child Quarterly*, 53(3), 203-216. <http://dx.doi.org/10.1177/0016986209334961>
- Zepeda, C. D., Richey, J. E., Ronevich, P., Nokes-Malach, T. J. (2015). Direct instruction of metacognition benefits adolescent science learning, transfer, and motivation:

An in vivo study. *Journal of Educational Psychology*, 107(4), 954-970.

<http://dx.doi.org/10.1037/edu0000022>

Zhang, B., & Mislak, J. (2015). Evaluating three grading methods in middle school science classrooms. *Journal of Baltic Science Education*, 14(2), 207-215.

Zheng, M., & Spires, H. A. (2014). Fifth graders' flow experience in a digital game-based science learning environment. *International Journal of Virtual and Personal Learning Environments*, 5(2), 69-86. <http://dx.doi.org/10.4018/ijvple.2014040106>