Spring 2015

A meta-analytic tutorial and a narrative review on motivation interventions in education

Rory A. Lazowski
James Madison University

A Meta-Analytic Tutorial and a Narrative Review on Motivation Interventions in Education

Rory A. Lazowski

A dissertation submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

Department of Graduate Psychology

May 2015

Dedication

This work is dedicated to my grandmother, Anne Fekete, who instilled in her children and grandchildren the value of education. I will never forget her words to me: "My dear Rory, get as much education as you possibly can. Nobody can take that away from you." Even though she passed away prior to me starting the Ph.D. program, I've felt her presence and strength every step of the way, especially when it was most needed. I love you pieces, Gramma.

Acknowledgments

I would first like to thank Dena Pastor, my advisor and mentor. I feel extremely fortunate to work with you and you've been the best advisor a student could ever ask for. You are a lifesaver. If there was an analysis for the completion of this dissertation, your contributions would be significant and the associated effect size would be huge. I'd also predict the remaining variance would be inconsequential. Thank you. I truly feel I owe you a debt that cannot be repaid.

I would also like to thank my other committee members, Debbi Bandalos, Robin Anderson, and Chris Hulleman for all of their contributions and recommendations toward my dissertation. This work represents an enormous source of pride and I genuinely appreciate your help in crafting a product that I am profoundly honored to have written. Debbi - thank you also for taking me under your measurement wing and offering to include me in various projects and activities to round out my skill-set. Your thoughtfulness does not go unnoticed. Robin - thank you for providing me with opportunities to stretch my professional development throughout the years, not only academically but also in my employment at JMU. And Chris - I wouldn't be at this point without you, both in terms of meta-analysis or motivation research. There is no way to put into words my gratitude for all you have done for me academically, professionally, and most of all, personally. You've had a tremendous impact on my life, my friend. Bigger than a 28-inch pizza.

Next, I'd like to thank Kenn Barron. Thank you for providing the Expectancy-Value Workshop that started it all 7 years ago. Thank you also for always being there to listen and provide academic and personal guidance over good food, good coffee, or good beer. I am also indebted to my current supervisor, Kurt Schick, and my previous supervisor, Valerie Schoolcraft, for supporting my scholastic efforts while I was working in their departments. Your encouragement and support to pursue and accomplish this dream would not be possible without you. I can only hope that my work and service to your programs was sufficient to demonstrate my gratitude. Matt Trybus, you are a true friend and I love you like a brother. You might find several Cohen's *d*'s throughout this paper, which should put a smile on your face. Thank you, Matt, for always being there, for your compassion, and for your authenticity. The world is a much better place with you in it. Thank you also for noticing my hard work and dedication via youtube...I would not have been at JMU if it wasn't for you. The next Sierra Nevada is on me.

Lisa Trimble, you are my best friend. You have been my rock and have provided me with strength when I needed it most throughout this entire process. You continually amaze me and you are a guiding light and a source of tremendous inspiration. You bring out the best in me. I'll always circle back.

To my dog, Georgie – thank you for always wagging your tail and being so excited to see me when I got home. No matter the day I've had, everything was okay when I saw you. My lifelong goal is to be the man that you think I am, and while I may never achieve it, Daddy will always try to accomplish that goal. I promise.

E, C, and G – you know who you are. I couldn't have managed without you. Thank you also to Bruce Springsteen and the E-Street Band for providing the soundtrack of my life these past several years. "Tonight I'll be on that hill 'cause I can't stop, I'll be on that hill with everything I got, Lives on the line where dreams are found and lost, I'll be there on time and I'll pay the cost, For wanting things that can only be found, In the darkness on the edge of town." I found that dream, Boss Man.

Finally, none of this would ever be possible without my family. My brothers Duke and Jeff have inspired me, each in their own way. I am incredibly proud to call you both my brother. This work represents that Lazowski work-ethic, which seems to be present in all of us. That we are all in education in some way or another is also a tremendous nod to my parents, Phil and Liz. Mom and Dad – thank you so much for being the best parents a son could ever ask for and for always pushing me to do my best academically. Without you, this would not be possible. Thank you also for the Sunday night conversations, whether I had good news or bad. You were always there for me and your support throughout the years cannot be measured. Thank you for the unconditional love and always know that I love you the same. For me personally, this work is a testament to the upbringing you've provided, and is thus, a shared accomplishment.

Preface

This dissertation is comprised of two separate papers, both of which draw from a recent meta-analysis conducted by Lazowski and Hulleman (2015). The first paper is a more technical, methodological treatment of meta-analysis that is presented as a tutorial using illustrations based on data from the Lazowski and Hulleman (2015) meta-analysis throughout. Because all meta-analyses focus on an effect size measure, the choice of effect size measure is examined and the concept of a weighted effect size is introduced and illustrated. Next, different types of models used to analyze the effect sizes are presented, namely fixed effects and random effects models. Various issues are examined, including technical aspects of the models, how the researcher determines which model to use, and implications for incorrect use of the models. I then extend these approaches to a multilevel approach to meta-analysis and draw comparisons from the regression models discussed earlier to the multilevel approach. This paper concludes with a treatment surrounding issues of publication bias, different techniques to examine the presence of publication bias, and the inclusion of published and unpublished studies in meta-analysis.

The second paper then shifts from the more technical, methodological focus in the tutorial paper to a more substantive focus about the importance of intervention work in educational research, primarily in the area of achievement motivation. Given the growing body of research over the past 50 years demonstrating the impact of motivation on various educational outcomes, most of this research has focused primarily on correlational or laboratory studies, with far fewer field experiments. This growing body of motivation research has also resulted in a proliferation of different theories to help explain motivation. Although these theories have helped develop substantial knowledge of the factors facilitating or thwarting motivation, the proliferation of theories and constructs has also contributed to some uncertainty about what factors are most salient to student motivation. In addition, there exists some overlap in the constructs and terminology used for seemingly distinct motivation theories, possibly resulting in jingle/jangle fallacies. Subsequently, it can be difficult to interpret motivation theory and research for practical use in the classroom. To address this difficulty and to bring some cohesion to the various similar (or dissimilar) constructs among the theories, I use the expectancy-value framework as an umbrella to categorize the various theoretical approaches and the interventions produced thus far. The primary sources, or drivers, of the interventions are also identified for the overarching constructs of expectancy, value, and cost. To illustrate the cohesion in constructs and these primary sources of motivation interventions, a narrative review of the interventions included in the Lazowski and Hulleman (2015) meta-analysis is presented. The narrative review serves as a qualitative complement to the quantitative analyses presented in our 2015 meta-analysis.

Table of Contents

Paper 1

Table of Contents

Paper 2

List of Tables

Paper 1

Paper 2

List of Figures

Paper 1

Abstract (Paper 1)

Dating back to only the early 1970's, the use of meta-analysis has recently grown steadily in the fields of psychology and education, after initially being used in the physical sciences. Meta-analysis is often lauded as an effective analytic tool to inform practice and policy, disentangle conflicting results among single studies, and identify areas that require additional information for a certain topic. However, because routine use of meta-analysis is relatively recent, there remain methodological issues that require clarity. In addition, as more advanced analytical and statistical techniques emerge, there is a need to examine how these techniques can be applied to meta-analysis and how these techniques differ from more traditional approaches to meta-analysis. Using data from a recent meta-analysis conducted by Lazowski and Hulleman (2015), this work is intended to be a tutorial to examine some of the methodological issues associated with meta-analysis. More specifically, the tutorial first examines the concept of effect size use in meta-analysis, the choice of analytic technique (fixed versus random effects models using traditional approaches), and comparisons of traditional approaches to a more recent approach to meta-analysis, multilevel modeling. The tutorial highlights differences in results that can be obtained depending on whether a fixed effects or random effects model is adopted. The tutorial also largely demonstrates similarities in the results obtained between traditional approaches to meta-analysis and the multilevel approach, although some differences are discussed in areas of notation, output, initial models used, and the advantage of additional flexibility associated with the multilevel analyses. Next, the issue of publication bias is discussed and the methods to detect publication bias (funnel plot, Orwin's fail safe *n*, and the trim and fill method) are presented and subsequently

illustrated using the Lazowski and Hulleman (2015) data. Finally, the present investigation concludes with an examination of best practices related to the inclusion of both published and unpublished (grey) literature in meta-analyses.

Abstract (Paper 2)

Intervention studies are a particularly important and valuable facet of educational research. This paper first discusses how intervention work can be used to help inform theory, research, and policy/practice in a multitude of ways. However, despite these benefits, intervention research in the field of education has been on the decline over the past two decades (Hsieh et al., 2005; Robinson et al., 2007). The field of academic motivation research is no different. Notwithstanding the considerable volume of theoretical, qualitative, observational, and correlational studies, there have been fewer experimental tests of motivation theory in the field of education (Wentzel & Wigfield, 2007). In order to systematically evaluate what has been done to date, Lazowski and Hulleman (2015) conducted a meta-analysis examining motivation interventions that were conducted in authentic educational field settings (e.g., classrooms, workshops) and found that the motivation interventions in this meta-analytic review were promising, averaging approximately a half a standard deviation effect size (d = 0.49; 95% CI = [0.42, 0.56]). However, although formal meta-analytic techniques can provide a quantitative analysis that can be useful in summarizing the interventions, one limitation is that there is often not enough space to also provide a comprehensive narrative review of the studies included. Thus, a narrative review can offer qualitative insight that can complement the quantitative analyses found via meta-analysis. Toward this end, in this paper I offer a more thorough narrative review of the studies included in our meta-analysis. Given the conceptual overlap among the theories and constructs therein, the expectancy-value framework is proposed as a means to organize the various intervention studies. In accordance with this organization, theories are categorized based on whether

the studies primarily target student (a) expectancies, (b) values, or (c) cost. In addition, within the general categories of expectancies, values, and cost I identify specific sources or pathways of expectancies, values, and cost that can be targeted by interventions. These sources or pathways refer to the underlying psychological processes that both serve as antecedents and that are potentially amenable to intervention by educational practitioners, including teachers, parents, and administrators (Hulleman et al., in press).

Meta-Analysis Tutorial Paper – Paper 1

Introduction

The Importance of Systematic Reviews in Educational Research

Gene Glass defined meta-analysis as the “statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings” (1976, p. 3). Meta-analysis is a technique included in the larger field of literature reviews/research syntheses and has proliferated since the 1970’s as a valuable means to summarize a collection of studies on a particular topic and identify and analyze the similarities and differences among the study characteristics (Baldwin and Shadish, 2011). In fact, use of meta-analyses in decision-making for policy and practice in the fields of education and psychology has increased exponentially over the course of the past 25 years. Williams (2012) found that the rate of published meta-analyses has increased steadily every year since 1990. In 2010, for instance, more than 800 published meta-analyses were found in the *PsycInfo* database and over 200 were found in the ERIC database (Educational Resources Information Center, 2013).

Meta-analyses are critically important to education research, especially in the role of informing practice and policy as well as identifying areas that need further investigation. Because meta-analyses do not rely on a single study, but rather an organized synthesis of several studies, they can be regarded as a tool to build stronger arguments for reliability and validity. Although it is difficult to discern just how many educational decisions about policies and practices are based on only one or a handful of studies, there has been criticism that all too often educational research “fails the policy-making and broader educational community by the non-cumulative nature of its findings”

(Davies, 2000, p. 365). In addition, the relevance, practicality, and quality of educational research have come under some scrutiny by critics who note that there exists a gap between practitioners and those undertaking research (2000). More recently, former IES Director John Easton (2013) advocated for the importance of making research more relevant and usable, the importance of rigorous methodological techniques to collect and analyze data, and the importance of the feasibility of findings to inform educators who strive to improve student success. Meta-analyses can serve to meet these challenges and can be a valuable tool for researchers to synthesize information, summarize a topic in a field, and then share these results with practitioners in a digestible, accessible way so as to bridge the researcher/practitioner divide. The advantages here are two-fold. One advantage is the use of sound research to inform practice based on an accumulation of evidence, not simply a single study. The second advantage is the potential to close the gap between the researcher and practitioner by providing resolution of what works and what does not work.

The remainder of this paper is intended to be a tutorial for conducting meta-analyses. First, common characteristics, categories, and features of meta-analyses are presented. Because all meta-analyses focus on an effect size measure, various types of effect sizes are then provided given different types of dependent variables; that is, whether the study design examines the relationship between a combination of continuous and dichotomous variables (Cohen's *d*), only dichotomous variables (Odds Ratio; *OR*), or only continuous variables (product-moment correlation; *r*). Next, different types of regression models used to analyze the effect sizes are presented, namely fixed effects and random effects models. Various issues are examined, including technical aspects of the models, how the researcher determines which model to use, and implications for incorrect use of the models. I then extend these approaches to a multilevel approach to meta-analysis and draw comparisons from the regression models discussed earlier to the multilevel approach. The tutorial concludes with an examination of publication bias, the inclusion of published and unpublished studies, and different techniques to examine the presence of publication bias.

A meta-analysis conducted by Lazowski and Hulleman (2015) is used to illustrate these topics. Our meta-analysis systematically reviewed educational interventions that were guided by academic motivation theories. We identified theoretically grounded motivation interventions that had been experimentally tested in educational contexts and examined the extent to which the interventions impacted various student outcomes. In summary, the meta-analysis included 66 published and unpublished papers of 84 field studies grounded in motivation theory, accounting for 37,239 participants. Data from this meta-analysis will be used as a recurring example throughout the tutorial.

For pedagogical purposes, one study (Yeager, unpublished; Study 2) was omitted for the illustrative examples in this tutorial. This study had a substantially larger sample size than any other study in the meta-analysis and an explanation pertaining to its omission is described in one of the sections that follow (***Illustration: RE model, no moderators***, p. 31).

Table 1 presents all of the studies represented in the Lazowski and Hulleman (2015) meta-analysis, along with their associated effect size, sample size, sampling variance, and study characteristics (e.g., grade of participants and type of experimental design).

Effect Sizes

When conducting meta-analyses, effect sizes are derived from the summary data found in each study in the analysis. The importance of effect sizes is well-discussed in the literature and the interested reader is encouraged to consult Kirk (1996) for an introduction to effect sizes and corresponding practical significance. The *American Psychological Association* (APA) recommends that effect sizes be included when reporting results as they provide the reader with a measure of the magnitude of the observed effect (2010). *P*-values, most commonly reported in primary studies, reflect the likelihood of observing a result, or something even more extreme in the direction of the alternative hypothesis, if the null hypothesis were true. Although *p*-values are partly a function of the size of an effect, they are also a function of sample size. Thus, a study with a large sample may yield a significant *p*-value but a small effect size. Correspondingly, a study with a small sample may yield a non-significant *p*-value but the effect size may be large. For these reasons, *p*-values are not used in meta-analyses and effect sizes are preferred. Effect sizes are typically used in their standardized form so that they are more comparable across studies, even when different measures or outcomes are used from study to study. In instances where effect sizes are not reported in a study, effect sizes are calculated by the researcher given the descriptive or inferential statistics provided (e.g., *t*-test, *F*-test).

The type of effect size ultimately used by the meta-analyst depends on the nature of the results, the types of statistical information reported, and the hypotheses and research questions surrounding the meta-analysis (Lipsey & Wilson, 2001). It is important to note that results for each of the primary studies included in a meta-analysis needs to be encoded into the same effect size statistic. In addition, the type of effect size must be appropriate given the relationships between or among variables in the studies, and to the statistical forms reported in the results (2001).

Borenstein (2009) suggests four major factors that should influence the choice of the effect size statistic to be used in the meta-analysis. First, the effect sizes from the primary studies should be comparable and approximately measure the same thing. For instance, the effect size should not be dependent on aspects of the research design that may differ across studies (e.g., use of covariates). That is, the meaning of the effect size should be the same regardless of the research design. The second factor is that the effect size should be interpretable and meaningful to the substantive researchers whose studies are represented in the meta-analysis. Third, if need be, the meta-analyst should be able to compute effect sizes from the information provided in the primary studies and not depend on raw data for re-analysis. Finally, the effect sizes should have sound technical properties such as known sampling distributions in order to compute variances and confidence intervals.

Like many statistics, sample size will impact the precision of the effect sizes included in the meta-analysis. In general, studies with smaller samples will have correspondingly larger estimates of sampling error for effect sizes compared to studies with larger samples, which will have smaller estimates of sampling error. Therefore, the values of every effect size in the analysis will have different degrees of reliability, and if not accounted for, effect sizes with large amounts of sampling error will contribute just as much as effect sizes with small amounts of sampling error in the final analyses. This is clearly problematic. To effectively address this problem, statistical models in meta-analysis weight each effect size by a term that reflects its precision. The optimal weighting term is a function of the standard error of the effect size (Hedges, 1982; Hedges & Olkin, 1985). Specifically, the weight is the inverse of the squared standard error (i.e., sampling variance) and is commonly termed the inverse variance weight. Therefore, both the effect size and the associated inverse variance weight are incorporated into the statistical analyses. Because it can be difficult to determine the standard error (and thus calculate the inverse variance weights), meta-analyses are generally conducted using effect size measures with known standard error formulas. The most common effect sizes used in meta-analysis include the standardized mean difference, the odds-ratio, and the correlation coefficient (Beretvas, 2010). The choice depends on the nature of variables in the study; that is, whether the study design examines the relationship between a combination of continuous and dichotomous variables, only dichotomous variables, or only continuous variables. The effect size measure most commonly used for the combination of continuous and dichotomous variables is discussed next. A detailed treatment of other effect size measures for only dichotomous variables and only continuous variables is provided in the Appendix.

Cohen's *d*. Research designs that incorporate group contrasts (comparison of one group to another) are widely used in meta-analysis (Lipsey & Wilson, 2001). Most frequently, experimental and quasi-experimental design studies are utilized, with a comparison of an experimental or treatment group with a control group (a dichotomous variable) on one or more dependent variables (a continuous variable). Different studies commonly use different instruments to measure a dependent variable or construct of interest and thus may not be numerically comparable across studies. This can include situations where the same construct is operationalized in a different manner or when different constructs are measured across studies. In both of these instances, the effect size statistic used to aggregate the findings in the meta-analysis must be standardized so that the values on the original measures are comparable. One effect size that can be used to compare the magnitude of the difference between two groups (e.g., experimental vs. control) across different measures is the standardized mean difference, or Cohen's *d* (Cohen, 1988). Cohen's *d* is appropriate when the dependent variables are continuous in nature. This effect size statistic is calculated using the following formula:

$$d = \frac{\overline{X}_{G1} - \overline{X}_{G2}}{S_{pooled}} \tag{1}$$

where $\overline{X}_{G1}$ and $\overline{X}_{G2}$ are the means for groups 1 and 2, respectively, and S_{pooled} reflects the pooled standard deviation (Hedges & Olkin, 1985). The sampling variance of Cohen's *d* is calculated as:

$$v_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)} \tag{2}$$

where n_1 and n_2 represent the sample sizes for each group; $\frac{n_1 + n_2}{n_1 n_2}$ represents the uncertainty in the estimate for the mean difference, the numerator in Equation 1; and $\frac{d^2}{2(n_1 + n_2)}$ represents the uncertainty in the estimate for the S_{pooled}, the denominator in Equation 1 (Borenstein, 2010).

The standard error of Cohen's *d* is calculated as the square root of v_d:

$$SE_d = \sqrt{v_d} \tag{3}$$

Thus, the inverse variance weight for Cohen's d (w_d), noted earlier as the weight assigned to the effect size to account for sampling error, is:

$$w_d = \frac{1}{v_d} = \frac{2n_1 n_2 (n_1 + n_2)}{2(n_1 + n_2)^2 + n_1 n_2 (d)^2} \tag{4}$$

Cohen's d can be biased upward when sample sizes are small (< 20) (Hedges, 1981). To account for this bias, a correction is applied and the estimate is referred to as Hedges' g (1981). Hedges' g and the associated variance and inverse variance weight are as follows:

$$g = \left[1 - \frac{3}{4N - 9}\right] d, \tag{5}$$

$$v_g = \frac{n_1 + n_2}{n_1 n_2} + \frac{g^2}{2(n_1 + n_2)}, \tag{6}$$

$$w_g = \frac{1}{v_g} = \frac{2n_1 n_2 (n_1 + n_2)}{2(n_1 + n_2)^2 + n_1 n_2 (g)^2} \tag{7}$$

where N represents the total sample size.

Illustration

The Froiland (2011) study from the Lazowski and Hulleman (2015) meta-analysis will be use to illustrate these calculations. This study included four separate dependent variables, with values of Cohen's d averaged across these measures to capture the average effect size. For clarity in this example, only one of the dependent variables will be presented – the Parent Questionnaire of Child Motivation to Learn (PQCML).

First, both the treatment and control groups consisted of 15 students each. At posttest, the mean of the treatment group on the PQCML was 123.3, with a standard deviation of 24.3. Correspondingly, the mean of the control group on the PQCML at posttest was 116.3, with a standard deviation of 20.7. Given these data, Cohen's *d* (Equation 1) comparing the treatment to control group on the PQCML can be calculated as follows:

$$d = \frac{\overline{X}_{G1} - \overline{X}_{G2}}{S_{pooled}} = \frac{123.3 - 116.3}{22.5} = 0.31$$

Next, the sampling variance (Equation 2) can be calculated as follows:

$$v_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)} = \frac{15 + 15}{(15)(15)} + \frac{0.31^2}{2(15 + 15)} = 0.135$$

The standard error (Equation 3) associated with this sampling variance is then:

$$SE_d = \sqrt{v_d} = \sqrt{0.135} = 0.367$$

Finally, the inverse variance weight for this effect size is computed as:

$$w_d = \frac{1}{v_d} = \frac{2n_1 n_2 (n_1 + n_2)}{2(n_1 + n_2)^2 + n_1 n_2 (d)^2} = 7.404$$

As noted above, Cohen's *d* can be upwardly biased when sample sizes are less than 20, and in these instances, Hedges' *g* is generally recommended. However, as Table 1 demonstrates, none of the studies included in the Lazowski and Hulleman (2011) meta-analysis had less than 20 participants. Only one study had exactly 20 participants (Reeve et al., 2004). Therefore, Cohen's *d* was chosen over Hedges' *g*.

In addition to the choice of effect size measure, there are several other methodological concerns that the meta-analyst must consider and address. Some of these include, but are not limited to: the choice of fixed vs. random effects models; combining effect sizes from different study designs (e.g., independent sample and dependent sample studies); combining effect sizes within studies (e.g. averaging effect sizes within one study vs. the use of procedures to account for the intercorrelation among the measures); inter-rater reliability between or among study coders; power; choice of analytic technique to conduct the meta-analysis; and the inclusion of published studies only vs. both published and unpublished studies (e.g., conference presentations not published in peer-reviewed journals, dissertations, master's theses).

Although a detailed treatment of all of these issues is beyond the scope of the current work, the following sections will more thoroughly address two of these issues: 1) choice of analytic technique to conduct the meta-analysis and 2) inclusion of published-only vs. published and unpublished studies.

Choice of Analytical Technique

Statistical models are used in meta-analysis to estimate the overall weighted effect size and variability of effect sizes across studies and to explore variables that may explain variability in the effect sizes comprising a meta-analysis. The four statistical models that are used in the vast majority of meta-analyses for these purposes are shown in Table 2, which differentiates models by whether they are a fixed effects or random effects model and also by whether or not moderators are included. The models shown in Table 2 align with those presented by Hedges and Olkin (1985) and are the same models presented using different notation in Lipsey and Wilson's popular primer on meta-analysis. In 1985 Raudenbush and Bryk described the connection between these meta-analytic statistical models and multilevel models. The connection is simple: the meta-analytic statistical models provided in Table 2 are a type of multilevel model. Although many methodologists have emphasized this connection (e.g., Hox, 2010, Marsh et al., 2009), researchers using the traditional approach to meta-analysis may not realize that they are in fact using the same models that are used in a multilevel approach. For instance, researchers using Lipsey and Wilson's primer as a guide to meta-analysis are likely unaware that they are using the same models as researchers who are using a multilevel approach.

One of the purposes of this tutorial is to emphasize the fact that the same statistical models underlie the traditional approaches to meta-analysis, such as the approach provided in the Lipsey and Wilson primer, and multilevel approaches. To emphasize this fact, the statistical models in Table 2 were fit to the Lazowski and Hulleman (2015) data using different procedures within IBM SPSS Statistics for

Windows, Version 21 (IBM Corp., 2012) and SAS 9.3 software (SAS Institute, Cary NC, 2011). The SPSS macros developed Lipsey and Wilson were first used for the traditional approach and PROC MIXED (Sheu & Suzuki, 2001) in SAS was then used for the multilevel approach. The estimates (which will be described in more detail below) obtained using the two approaches are similar and are provided in Tables 3 and 4.

Although the traditional and multilevel approaches use the same statistical models and yield essentially the same parameter estimates, there are differences in the analytical orientation of researchers adopting the different approaches. Specifically, there are differences in terminology, aspects of the results that are emphasized, which initial model is typically fit to the data, estimation procedures, and the number of moderators included in the model simultaneously. These differences are explained further in the sections below.

Traditional Approach

A traditional approach to meta-analysis aligned with the Lipsey and Wilson primer and using their associated SPSS macros for analysis is provided below. First, a fixed effects model is presented that estimates the overall weighted effect size and assesses variability in effect sizes (Cell A of Table 2). Next, a fixed effects model that incorporates moderators to explain significant variability in effect sizes is presented (Cell B of Table 2). Here, the term moderator refers to different study characteristics that are considered independent variables or predictors that help explain excess variability in effect sizes across studies in the meta-analysis. The term moderator will thus be used throughout this tutorial to reflect the independent variable(s) or predictor(s) in the models. Predictors are called moderators in this context because they moderate the relationship between the experimental conditions (treatment vs. control) and the dependent variable.

Two different approaches are illustrated using this model – one where each moderator is analyzed separately in an ANOVA framework followed by one that explores the effects of various moderators simultaneously in a regression framework. Following the presentation of the fixed effects models, the use of random effects models is then introduced to estimate the overall effect size and between-study variance in effect sizes using the intercept-only model (Table 2, Cell C). The use of moderators to explain variability in effect sizes with the random effects model is then introduced (Table 2, Cell D). This section again illustrates the more traditional approach to meta-analysis by demonstrating how the results from the various fixed and random effects models can be obtained and the differences that can be expected when a fixed versus random effects model is employed.

The Fixed Effects Model

No moderators. Once effect sizes are obtained from each of the primary studies that are to be included in the meta-analysis, the first step is to combine them in such a way to arrive at a single value that we use as the estimate of the population effect size. The fixed effects model (Cell A of Table 2) can be used to accomplish this task and makes the assumption that the effect sizes are simply direct replications of one another (i.e., the effect size is the same in all studies) and that the only differences among the effect sizes are due to sampling error (Hedges & Olkin, 1985).

In the fixed effects model each observed effect size represents an estimate of the population parameter, γ_0, and any variation in the observed effect sizes from γ_0 is only attributable to sampling variance. The population effect size, γ_0, is estimated using a weighted average across the k observed effect sizes in the primary studies, with the weights assigned to each study (w_j) being equal to the inverse variance weight. For Cohen's d, this is the inverse variance weight that was discussed and presented in Equation 4 and thus this overall effect size represents a weighted average of the effect sizes, as shown in Equation 8:

$$\hat{\gamma}_0 = \frac{\sum_{j=1}^{k} w_j d_j}{\sum_{j=1}^{k} w_j} \quad . \tag{8}$$

The standard error of the estimated population effect size is a function of the weights associated with each effect size,

$$SE_{\hat{\gamma}_0} = \sqrt{\frac{1}{\sum_{j=1}^{k} w_j}} \, . \tag{9}$$

The estimate of the population effect size along with its standard error are used in the calculation of significance tests and confidence intervals for the estimate.

Oftentimes in meta-analysis, the various studies that are included are not exact replications and may differ from one another in a variety of ways. For instance, studies may differ on the operational definition of the outcome variable, the population from which the sample is derived, and type, length, or dose of treatment delivered (2003).

Despite these differences, it does not necessarily follow that the effects differ across studies. An important step in the meta-analysis is to test for homogeneity of effect sizes included in the study, which can be conducted using the Q-statistic (Cochran, 1954):

$$Q = \sum_{j=1}^{k} w_j (d_j - \hat{\gamma}_0)^2 \sim \chi^2(k-1) \tag{10}$$

where k reflects the number of effect sizes. If the null hypothesis of homogeneity is not rejected, the effect sizes differ from the population mean by sampling error only. In this case, the researcher would use a fixed effects model with no moderators (refer to Table 2).

Illustration: FE – no moderators

Using Equation 8, the weighted average ($\hat{\gamma}_0$) of the 83 effect sizes using a fixed effects model in the Lazowski and Hulleman (2015) meta-analyses is 0.403 (95% *CI* [0.370, 0.436]), with a corresponding standard error (Equation 9) of 0.017 (see Table 3 under FE: Traditional). The overall homogeneity statistic, Q, was statistically significant, $\chi^2(82) = 297.239$, $p < .001$. Thus, one would reject the hypothesis of homogeneity and conclude that the variance in the population of effect sizes was greater than would be expected from sampling error alone. These computations were performed via the MeanES macro[1] provided by Lipsey and Wilson (2001).

With moderators. When the null hypothesis of homogeneity is rejected, then the variability of the effect sizes is larger than would be expected by sampling error alone, and thus each effect size does not estimate a common population mean (Lipsey & Wilson, 2001). In this case, one option is to continue with a fixed effects model, but include moderator variables based on study characteristics discussed above. One way to go about doing so (that is more aligned with a traditional approach to meta-analysis) is to divide the studies into homogenous groups (again, based on study characteristics) and perform separate moderator analyses via a meta-analytic analog to ANOVA (Hedges, 1982b; Hedges & Olkin, 1985). For example, studies could be categorized according to the type of experimental design (e.g., either randomized or quasi-experimental design) and also categorized by grade level (e.g., elementary school, middle school, high school, post-secondary). Then, two analyses would be run using the analog to ANOVA – one using the type of experimental design as the moderator and the other for the age group of the sample as the moderator. There is no particular statistical reason why one moderator is examined at a time. Rather, this is traditionally how the approach has been conducted.

Illustration: FE with moderators, separate ANOVAs

To illustrate, the results of two separate ANOVAs using the example data were obtained using the MetaF macro provided by

---

[1] Note that the same results could be obtained using the MetaReg macro, the macro appropriate for weighted least squares regression analyses in meta-analysis, specifying no predictors and asking for a fixed effects model. The only difference in the output would be the omission of the *Q* statistic in the MetaReg macro results.

Lipsey and Wilson (2001). When the MetaF macro is used, attention is paid to two quantities: Q_B and Q_W. Q_B represents the between-group variance in effect size, and Q_W represents the within-group variance in effect size. In the moderator analyses, if Q_B is significant, this indicates that there are significant differences in effect sizes across groups and that a significant amount of variability is explained by the moderator. If Q_W is significant, this indicates that there is additional variance in effect sizes not explained by the moderator. However, if Q_W is not significant, the moderator sufficiently captures the excess variability in effect sizes (Lipsey & Wilson, 2001).

With respect to experimental design, there were 61 studies coded as randomized experiments and 22 studies coded as quasi-experimental designs. Results of the moderator analyses on experimental design indicated that this variable explained a statistically significant amount of variability in effect sizes (Q_B = 39.159, $p < .001$). This suggests that the weighted mean effect sizes between experimental designs differed by more than sampling error. The weighted mean effect size for randomized experiments was 0.347 (95% *CI* [0.310, 0.384]) and the weighted mean effect size for quasi-experimental designs was 0.599 (95% CI [0.370, 0.436]). The experimental design of the study explained 13% of the variance in effect sizes (Q_B /(Q_B + Q_W)). The pooled within group variance was also significant (Q_W = 258.080, $p < .001$), suggesting the variability within experimental designs was significant and that the categorical variable represented in Q_B (experimental design) was not sufficient alone to account for the excess variability in the effect size distribution.

We also coded studies according to the grade level of participants. In all, there were 8 studies conducted with elementary students, 22 with middle school students, 14 with high school students, and 39 with students enrolled at a post-secondary institution. Results of the moderator analyses using grade level indicated that this variable explained a statistically significant amount of variability in effect sizes (Q_B = 26.412, $p < .001$). This suggests that the weighted mean effect sizes among grade level differed by more than sampling error. The weighted mean effect size for middle school students was largest (0.543; 95% *CI* [0.469, 0.617], followed by post-secondary students (0.411; 95% *CI* [0.363, 0.459]) elementary students (0.372; 95% *CI* [0.267, 0.477]), and high school students (0.2801; 95% CI [0.035, 0.212]). The grade level of the participants in the study explained 9% of the variance in effect sizes ($Q_B /(Q_B + Q_W)$). The pooled within group variance was also significant (Q_W = 270.827, $p < .001$), suggesting the variability within the different age groups was significant and that the categorical variable represented in Q_B (age group) was not sufficient alone to account for the excess variability in the effect size distribution.

Illustration: FE with moderators, separate regressions

Although the MetaF macro was used in these two separate ANOVAs, the same analyses could be executed within a multiple regression framework by including code variables as predictors to represent the categorical variables. The analog to multiple regression (Hedges, 1982b, 1983b; Hedges & Olkin, 1985) in the context of meta-analysis is often called meta-regression (Konstantopoulos & Hedges, 2009). The MetaReg macro provided by Lipsey and Wilson (2001) can be used for meta-regression and uses the same model (Cell B of Table 2) as the MetaF macro; the only difference in the macros is the nature of the output provided. Specifically, the MetaReg macro output contains two quantities: Q_R and Q_E. Q_R represents the regression sum of squares and tests whether the regression model is significant; that is, whether the regression model explains a significant amount of variability in effect sizes (Lipsey & Wilson, 2001). This quantity is the same as Q_B discussed in the ANOVA framework. In addition, the regression model will also yield Q_E which represents the sum of squares residual; that is, the unexplained variability in effect sizes that is not accounted for in the model (2001). It is therefore the same as Q_W in the ANOVA framework. The correspondence between Q_R and Q_E in the MetaReg macro output and Q_B and Q_W in the MetaF macro output highlights the well-known fact that ANOVA are regression are equivalent. When experimental design and grade level are entered as predictors (using a series of code variables) in separate meta-regression models, the results obtained used the Meta-Reg macro are the same as the ANOVA results obtained using the MetaF macro.

Examining moderators via regression (i.e., the approach taken in Cell B in Table 2) is preferred over performing separate analyses, as illustrated with the above ANOVAs. One main reason is that the moderators can be examined together, taking potential intercorrelations between or among the moderators into account which is not the case when separate analyses are performed (Viswesvaran & Sanchez, 1998). That is, the effects of one moderator can be examined after controlling for the effects of other moderators. Furthermore, regression offers flexibility in the types of moderators to be analyzed through the ability to handle both categorical and/or continuous moderator variables (Marsh, Bornmann, Mutz, Daniel, & O'Mara, 2000; Van den Noortgate & Onghena, 2003) as well as interactions between moderators.

Illustration: FE with moderators, single regression

To illustrate, a single meta-regression model was estimated using the MetaReg macro with both experimental design and grade level entered as predictors. Because dummy coding was used for each categorical variable a total of four code variables were entered as predictors into the model (one code variable for experimental design and three for grade). The quasi-experimental group served as the reference group for the experimental variable and the elementary grade level served as the reference group for the grade variable.

The parameter estimates and standard errors of this model are provided in Table 4 under FE: Traditional. Results indicated that Q_R was significant (Q_R = 59.911, p < 0.001), suggesting the regression model explains a significant amount of variability across effect sizes. Q_E was also significant (Q_E = 237.327, p < 0.001), suggesting that the unexplained variability was greater than would be expected from sampling error alone. Despite that fact that significantly variability in effect sizes remains once controlling for these two predictors, the experimental design and grade level associated with the studies together explain 20% of the variance in effect sizes (Q_R /(Q_R + Q_E)).

In addition to ascertaining the variance explained by the set of predictors, entering in predictors simultaneously allows the effects of one predictor to be examined once controlling for the effects of the other predictor. For instance, a comparison of the coefficient[2] associated with experimental design in the model including only this predictor (β_1 = -0.253, p<0.001) to the model including both

[2] Because dummy-coding was used with quasi-experimental designs as the reference group, β_1 represents the difference between the effect sizes associated with randomized designs and those associated with quasi-experimental designs.

this predictor and grade ((β_1 = -0.249, p<0.001), shows that experimental design is still a significant predictor once controlling for grade.

Another advantage in utilizing a regression approach is that interactions between variables can be explored. For example, additional code variables could be entered into the model to represent the interaction between experimental design and grade level. Given the low number of quasi-experimental studies in the elementary, middle and high school grade levels, this analysis was not pursued.

The Random Effects Model

With moderators in the fixed effects model discussed above, it is assumed that the known study characteristics included in the model are able to account for all the variability in the true effect sizes and/or the remaining variance is negligible (Hedges, 1983a). However, in practice the studies comprising the meta-analysis are rarely exact replications with regard to these study characteristics, nor do they often account for all the heterogeneity in effect sizes (Hedges, 1983a); thus the assumption of homogeneity is tenuous at best, making the fixed effects models unrealistic (Erez, Bloom, & Wells, 1996; Hedges & Vevea, 1998; Hunter & Schmidt, 2000; National Research Council, 1992; Schmidt, Oh, & Hayes, 2009). For a fixed effects model, one must have strong evidence that the studies included in the meta-analysis were virtually identical (Aronson, Ellsworth, Carlsmith, & Gonzalez, 1990; Schmidt, Oh, & Hayes, 2009). Schmidt, Oh, and Hayes (2009) provide an example, noting the following:

> If the studies drew their samples from the same population (e.g. college sophomores), tested exactly the same hypotheses with exactly the same study design, treatment strength (if an experimental study), measures, instructions, time limits, etc, then one might assume *a priori* that the same population parameter was estimated in all the primary studies (i.e. σ_{δ}^{2} or $\sigma_{\rho}^{2} = 0$) and this could be the basis for choosing the FE model. (p. 124)

Because heterogeneous results are common and expected (Engels, Schmidt, Terrin, Olkin, & Lau, 2000), random-effects models are preferred over fixed-effects models (Erez, Bloom, & Wells, 1996; Hedges & Vevea, 1998; Hunter & Schmidt, 2000; National Research Council, 1992; Schmidt, Oh, & Hayes, 2009).

Differences between the random effects and fixed effects models also have implications regarding the inferences that can be drawn from the results. A random effects model allows the researcher to generalize results beyond those found in the study, whereas this generalization is inappropriate for a fixed effects model. For a fixed effects model, inferences can only be made about the studies included in the meta-analysis (Hedges & Vevea, 1998). Because the assumption that the true effect size is the same in all studies (any variation is solely due to sampling variance) is often untenable and the limited generalizability in fixed effects models, random effects models are generally recommended (Baldwin & Shadish, 2011; Borenstein, Hedges, Higgins, & Rothstein, 2009).

As implied above, random effects models differ from fixed effects models in a number of ways. The random effects model allows the true effect size to vary depending on the study, whereas the fixed effects model assumes that all studies comprising the meta-analysis share one common effect size (Hedges & Vevea, 1998) or are strictly a function of moderators. To clarify, a direct comparison of the fixed effects models to the random effects models are provided in Table 2. Most notably, in a random effects model, sources of variation include both the within study estimation error and between study variance, as captured by v_j and τ, respectively; conversely, the fixed effects model only includes the source of variation associated with within study estimation error (v_j) (Borenstein, Hedges, Higgins, & Rothstein, 2009). There are a variety of different estimators that can be used for τ (for a review see Viechtbauer, 2005). A commonly used estimator is DerSimonian and Laird's (1986) estimator for τ, which is a noniterative method of moments (MM) approach based on a quadratic form of Q:

$$\tau = \frac{Q-(k-1)}{c} \tag{11}$$

where Q represents the homogeneity test statistic, k represents the number of studies, and c is calculated by $\sum_{j=1}^{k} w_j - (\sum_{j=1}^{k} w_j^2)/(\sum_{j=1}^{k} w_j)$.

The Q-statistic used in Equation 11 is the same as was used with the fixed effects model. Thus, the same Q-statistic computed using Equation 10 is used to test the null hypothesis of homogeneity of effect sizes, which indicates the between study variance (represented using τ) equals zero. Although the same Q-statistic is used in the fixed and random approaches, the actions that follow based on its results vary across researchers. For instance, researchers wedded to a fixed effects model may interpret the Q-statistic as an indication that moderator variables need to be included in additional fixed effects models. Other researchers might use the *Q*-statistic to decide which kind of model (fixed or random effects) to use in estimating the overall effect size and in moderator analyses. For instance, a researcher may begin with a fixed effects model, but switch to random effects models if a significant *Q*-statistic is obtained. Proceeding in this fashion is called a conditionally random effects approach because adoption of a random effects model is conditional upon the outcome of the test for homogeneity, or *Q* (Hedges & Vevea, 1998). Similarly, another researcher might begin with a random effects model and only switch to a fixed effects model if the *Q*-statistic is non-significant.

No moderators. When using a random effects model without any predictors, in other words, when using a random effects model to estimate the overall effect size across studies, studies are weighted by their estimated precision to produce an overall weighted mean effect size using Equation 8. Thus, the same equation to obtain the overall effect size is used in the fixed and random effects models.

The difference is in the weights assigned to each study (w_j). The weights in a random effects model are now equal to $w_j = 1/(v_j + \tau)$. If τ is non-zero, the weights will be smaller in a random effects model compared to a fixed effects model. The standard error of the estimated population effect size is still computed using Equation 9, but with the weights now equal to those associated with a random effects model. Because the weights are smaller when τ is non-zero, the standard error of the overall effect size will be larger in a random effects model compared to a fixed effects model. As before, the estimate of the population effect size along with its standard error is used in the calculation of significance tests and confidence intervals for the estimate.

Illustration: RE model, no moderators

Using the formula in Equation 8, the weighted average of the 83 effect sizes using a random effects model in the Lazowski and Hulleman (2015) meta-analyses is .497 (95% *CI* [0.428, 0.566]), with a corresponding standard error (Equation 9) of 0.035(see Table 3 under FE: Traditional - MM). The overall homogeneity statistic, *Q*, which is the same as that computed for the fixed effects model, was statistically significant, χ^2 (82) = 297.239, $p < .001$) indicating that the estimate of the between-study variance in effect sizes ($\hat{\tau}$= 0.063) is significantly different than zero. The square root of $\hat{\tau}$ is the standard deviation of population effect sizes. Its value is 0.251 indicating that population effect sizes vary from the overall effect size of 0.497 by about 0.251 units.

When a fixed effects approach was adopted, the overall effect size using the fixed effects model was presented and the significant *Q*-statistic was used to justify incorporating moderators into further fixed effects models. In this example, when a random effects model was used to estimate the overall effect size, the significant *Q*-statistic was used to justify retention of the overall estimate based on the random effects model and to pursue further analyses incorporating moderators into the random effects model.

Before moderators are included, it is important to point out two noteworthy differences between the results for the fixed effects and random effects approaches to computing the overall effect size. First, the standard error is larger in the random effects approach. This is expected and is a result of incorporating the non-zero between-study variance (τ) into the computations. The second difference pertains to the differences in the estimates of the overall effect sizes. In the fixed effects model γ_0 was estimated as 0.403 and in the random effects model as 0.497. Differences between the estimates from the two models are expected when τ is non-zero. Recall that in the fixed effects model, variability in effect sizes is only assumed to be due to sampling variance. However, in the random effects model, variability in effect sizes is assumed to be due to both sampling variance and between study variance. In each of the models, the inverse variance weight is a function of sample size, with more weight placed on studies that have larger sample sizes. As a result, these studies have a larger impact on the overall weighted effect size. This weight also has a larger impact in fixed effects models where the only variability in effect sizes are attributed to sampling variance. On the other hand, the between study error variance in the random effects models attenuates the weight placed on larger studies as some of the variance is also attributed to between study variation.

For instance, the Paunesku et al. (unpublished) study (Table 1) in the Lazowski and Hulleman (2015) meta-analysis had a substantially larger sample size compared to most other studies. The effect size associated with this study was also smaller (d = 0.14) than the overall weighted effect sizes in both the fixed effects and random effects models (0.403 and 0.497, respectively). As is demonstrated here, this study had a larger impact on the overall weighted effect size associated with the fixed effects model by comparison to the random effects model. In the fixed effects model, the inverse variance weight associated with this study was calculated as 144.409. In comparison, for the random effects model, the inverse variance weight associated with this study was reduced to 14.301. Again, this is due to the additional source of variability (between study variability) that is incorporated in the inverse variance weight calculated for each study in the random effects model.

As mentioned earlier, Yeager (unpublished) Study 2 was omitted from the analyses in this tutorial due to the much larger sample size in this study compared to other studies in the meta-analysis. The sample size for this study was 21,559 students. When this study is included, the magnitude of the difference in the weighted average effect sizes between the fixed and random effects models was even more pronounced. Including this study yielded a weighted average effect size of 0.258 for the fixed effects model compared to 0.489 for the random effects model – a difference of 0.231!

With moderators. To explore the effects of moderators in a random effects model, either ANOVA approaches (using only categorical moderators) or regression (using either categorical or continuous moderators) can be used. Given the aforementioned weaknesses of relying on separate ANOVA models, only regression is considered here. The random effects regression equation used to examine the moderating effects of the study characteristics and is presented in Cell D from Table 2. As can be seen, this equation is similar to the one presented for the fixed effects model in Cell B with the addition of u_j in the random effects model. Weighted least squares is used to estimate regression coefficients, with the weights now being equal to $w^*_j = 1/(v_j + \tau^*)$, with τ^* representing the between-study variance once controlling for the predictors in the model (computational details for computing τ^* can be found in Raudenbush, 2009).

Illustration: RE model, with moderators

To illustrate, a random effects regression model was estimated using the MetaReg macro with code variables in the model to explore the effects of experimental design and grade level on the effect sizes, simultaneously. The estimation method for τ was specified as noniterative method of moments (MM; Raudenbush, 2009). Parameter estimates and standard errors are presented in Table 4 under RE: Traditional-MM. Results indicated that Q_R was significant ($Q_R = 10.16$, $p < 0.001$), suggesting the regression model explains a significant amount of variability across effect sizes. Q_E was also significant ($Q_E = 74.94$, $p < 0.001$), suggesting that the unexplained variability was greater than would be expected from sampling error alone. Despite that fact that significantly variability in effect sizes remains once controlling for these two predictors, the experimental design and grade level associated with the studies together explain 12% of the variance in effect sizes ($Q_R / (Q_R + Q_E)$).

It is informative to contrast these results with those from the fixed effects model provided earlier. First, note that the parameter estimates are somewhat different than those in the fixed effect model. Differences are due to the weights used in their estimation. Specifically the weights used in the random effects model now incorporate conditional between study variance (τ^*) in addition to sampling error. Differences in the parameter estimates between fixed and random effects will be larger as τ increases. Second, note that the standard errors are larger in the random effects model in comparison to the fixed effects model. Again, this is due to the addition of τ^* in the random effects model. Third, note that compared to the traditional random effects model with no moderators, in the traditional random effects model with moderators the between study variance has been reduced. This reflects a reduction in the between study variance due to the addition of moderators in the model.

Misspecifying the model. The choice of which model to use is an important decision the meta-analyst must make. Applying fixed effects models when random effects models are more appropriate (and vice versa) can result in substantial biases and distortions in conclusions. For instance, Schmidt, Oh, and Hayes (2009) re-analyzed previously published meta-analyses using random effects models that were originally analyzed using fixed effects models. Their results demonstrated that fixed effects standard errors were much smaller and confidence intervals around mean effect sizes were substantially narrower compared to the random effects re-analyses (2009). Therefore, if the studies included in the meta-analyses were truly random but analyzed using fixed effects models, standard errors of parameter estimates would be too small and correspondingly, Type I error rates would be inflated. It is interesting to note that the authors indicated that none of the meta-analyses in their study mentioned the plausibility that the studies included in the meta-analyses were exact replications of one another, a primary argument for using fixed effects models. Therefore, the authors argue that the precision of findings reported in meta-analyses could potentially be overestimated, leading to important consequences for research and practice that have been based on faulty grounds (2009).

Applying this same logic to the use of random effects models when fixed effects models are more appropriate, an opposite trend emerges. More specifically, in this instance the standard errors will be too large and confidence intervals will be too wide, resulting in lower power and an increased likelihood of Type II error rates. However, this is less of a concern given that for most meta-analyses, a random effects model is more appropriate.

Illustration: Conflicting results using fixed and random effects

Conflicting results and interpretations of findings between fixed effects and random effects models is demonstrated with the Lazowski and Hulleman (2015) data. As can be seen from Table 4, the parameter estimate for middle school in the traditional fixed effects model (γ_{02} = 0.151) was statistically significant (p = 0.022), indicating that the effect size for studies in middle schools is significantly higher than the effect size for studies in elementary schools (controlling for the experimental design of the study). This information would indicate that the moderator, grade level, was significant (i.e., one of the groups was significantly different than at least one other group controlling for the experimental design of the study). However, the corresponding parameter estimate in the random effects model (γ_{02} = 0.165) was not statistically significant (p = 0.206), nor were any other estimates for grade level. Thus, in the random effects model, one would conclude that grade level was not a significant moderator. It is important to note that we justified and used a random effects model in our meta-analysis. Should we have used a fixed effects model instead, we may have arrived at a different conclusion for the grade level moderator effect.

Meta-analysis Using Multilevel Modeling

Extending upon the fixed effects and random effects regression models already discussed, another analytic approach to conducting a meta-analysis is through the use of

Hierarchical Linear Models (HLM), otherwise known as multilevel models. Meta-analyses can be considered a special case of multilevel modeling (Hox, 2010) with multilevel modeling providing a useful approach to distinguishing the various sources of variability discussed already (e.g., within study sampling error variance and random effects or between studies variance) (Raudenbush & Bryk, 2002).

Using a multilevel approach, the data are considered hierarchical, with subjects (Level 1) nested within studies (Level 2) (2002). At Level 1, the estimated effect size for each study varies randomly due to sampling error around the population or true effect size (Raudenbush, 2009) for that study. At Level 2, the true effect sizes vary among studies due to different study characteristics plus a random effect that represents unknown or unobserved sources of variability in true effect sizes (2009). Thus, the multilevel meta-analysis model is usually represented as the random-effects model shown in Cell D (Level 1) from Table 2 indicating that the observed effect sizes include both the true effect size and error.

In the multilevel model, the moderator effects are treated as fixed and the u_js are treated as random, and therefore, this model is sometimes called a mixed-effects model. There is a correspondence between the fixed-effects or random-effects models discussed earlier (Van den Noortgate & Onghena, 2003) and multilevel meta-analysis model (Cell D from Table 2). This model simplifies to the fixed-effects regression model with moderators (Cell B, Combined Equation in Table 2) when the between study variance (τ) is zero; the model also simplifies to the fixed-effects model with no moderators (Cell A in Table 2) when the between study variance is zero (the Level 2 variance is zero) and no moderators are included. Finally, the model simplifies to the random-effects model with no moderators (Cell C in Table 2) when no moderators are included (2003).

One difference between traditional and multilevel approaches is which model is used as the initial model in the analysis. In the multilevel approach, the meta-analyst always starts with the random effects model but may simplify to the fixed effects model when the between study variance is zero. However, in the traditional approach, as aforementioned, meta-analysts use a variety of starting points, including starting and staying with a fixed effects model or starting with a fixed effects model and switching to a random effects model based on the *Q* statistic results.

Perhaps the largest difference between the multilevel approaches and the traditional approaches lies in the estimation procedures that are typically used for the between-study variance. In the traditional approaches, several estimation procedures can be used, but the most common include noniterative method of moments (MM), full maximum likelihood (ML), and restricted maximum likelihood (REML; Lipsey & Wilson, 2001). In the multilevel approach, maximum likelihood methods are most commonly used [either ML or REML, though Hox and de Leeuw (2003) note that REML procedures are preferred over ML in situations with small samples]. Therefore, the multilevel approaches most commonly use iterative estimation techniques whereas traditional approaches most commonly rely on closed form estimation techniques. Note that differences in estimation techniques across the two approaches only pertain to the random effects models, where the between-study variance is estimated. The use of different estimation procedures also invokes different assumptions about the distribution of u_js. When maximum likelihood techniques are used in either approach, the assumption of normality is made. When method of moments techniques are employed, which are limited to the traditional approaches to meta-analysis, an assumption about the probability distribution of the u_js is not made (Raudenbush, 2009).

Illustration: Comparison of results using traditional vs. multilevel modeling in meta-analysis

The illustrative examples for the traditional techniques provided thus far utilized noniterative method of moments (MM) techniques for estimating the between-study variance in random effects models. To illustrate the similarities and differences between traditional and multilevel approaches when the estimation method is held constant, the traditional random effects model results[3] using ML estimation are also provided in Tables 3 and 4 along with the multilevel results using ML.

The parameter estimates and standard errors across the traditional and multilevel fixed effects models in Tables 3 and 4 are almost identical, with the exception of small differences in the *p*-values, which are a result a result of using the normal distribution for significance testing in the former and *t*-distributions in the latter. With respect to the random effect models in Tables 3 and 4, there are small differences between traditional

[3] The MetaReg macro was used to acquire the ML results for both the random effects intercept only model and the random effects with moderators model. Although the MeanES macro was used to acquire the random effects intercept only results when noniterative method of moments was used, the use of a different estimation method (such as ML) is not an option with this macro. For this reason, we used the MetaReg macro to acquire the results for this model, since this macro can estimate the between-study variance in random effects models using noniterative MM, ML or REML. To use this macro to acquire the results of the intercept-only model, no predictors are specified when calling the macro.

> and multilevel parameter estimates and standard errors when different estimation procedures are used, but not when the same estimation procedure is used. It is noteworthy that the same conclusions would be made about effect sizes in this example using either traditional or multilevel approaches.

Other authors (e.g., Hox, 2010) who have utilized both traditional and multilevel approaches with the same data have noted the same similarities in the results of the two approaches. Simulation studies show a similar trend; specifically that the results obtained via the traditional random effects approach do not substantially differ from results of the multilevel approach. Van den Noortgate and Onghena (2003) conducted a simulation study comparing these approaches with varying mean group sizes ($\bar{n}$ spanning 3 to 100), number of studies (*k* spanning 3 to 100), varying sample sizes across studies (slightly unbalanced, largely unbalanced), overall effect size (0, 0.5, 1), variance in true effect sizes (i.e., between study variance; 0, 0.05, 0.1), and varying distributions of the true effect sizes (normal, symmetric with heavy tails, skewed with heavy tails). The true model was a random effects model without moderators. Each data set was analyzed using four traditional methods, including: fixed effects models, random effects models with τ calculated according to two different method of moments estimators (e.g., DerSimonion & Laird, 1986; Hedges & Olkin, 1985), conditional random effects models (i.e., using a random effects model if Q test significant, fixed effects models otherwise). Each data set was also analyzed using a multilevel random effects model and restricted maximum likelihood estimation. The performance of these five different approaches were compared with respect to estimation of the overall effect size (γ_0) and between study variance in effect sizes (τ). The significance tests of these parameters were also compared across the five approaches.

The authors concluded that the performance of the multilevel approach was comparable to traditional approaches. In considering estimators, likelihood estimators[4] (ML, REML) might be preferred because they are more efficient in large samples than method of moments (MM) estimators (Raudenbush, 2009). The normality assumption invoked by the likelihood estimators regarding u_js might be considered a drawback, but the simulation results of Van den Noortgate and Onghena (2003) indicated the robustness of methods, including the multilevel likelihood methods, with non-normal distributions of true effect sizes.

A notable finding in the Van den Noortgate and Onghena (2003) study pertained to the performance of the various significance tests of the between-study variance. Traditional methods often use some form of the *Q*-statistic (Equation 10). What differs across these forms is whether the weights and overall effect size are based on the fixed effects model or the random effects model and if the latter, which procedure was used to estimate τ. For instance, τ could be calculated using noniterative method of moments approaches, which include DerSimonian and Laird's (1986) estimator (Equation 11) and Hedges and Olkin's (1985) estimator. Likelihood-based estimators (ML, REML) of τ are also available and Van den Noortgate and Onghena (2003) call *Q*-statistics using

---

[4] If likelihood methods are chosen, researchers might prefer REML since the ML estimates of τ biased when k is small. If REML is used it is important to keep in mind that likelihood ratio tests (LRTs) of nested models differing in fixed effects are not appropriate.

likelihood-based weights and an overall effect size the multilevel *Q*-statistic[5] because likelihood-based estimators are more commonly used in multilevel approaches. Other options available for testing τ in multilevel meta-analysis (or when likelihood-based estimators are used in traditional approaches) include the Wald test, where the ratio of τ to its standard error is compared to a normal distribution, and likelihood-ratio tests (LRTs), which compare the deviances (-2LLs) of the random effects model and a fixed effects model where τ is constrained to zero. To illustrate these tests and their results are shown for the random effect model with no moderators in Table 5. In this example, all tests indicate the same conclusion – that there are significant between-study variances in effect sizes. The more thorough investigation of the performances of these tests by Van den Noortgate and Onghena (2003) indicate poor performance of the Wald and LRT significance tests of τ. For this reason, it was recommended that the multilevel *Q*-statistic (i.e., the chi-square statistic proposed by Raudenbush and Bryk (1985)) be used to test τ when a likelihood-based estimator is adopted.

The Van den Noortgate and Onghena (2003) study also highlighted the poor performance of all approaches in situations where there are few studies in the meta-analysis, particularly when those studies have small sample sizes, a finding also noted by Marsh et al. (2009) in a meta-analysis examining gender effects for peer reviews of grant proposals. Typically, small numbers of primary studies in meta-analyses are problematic for all approaches. However, this is likely more problematic for multilevel approaches (or when likelihood-based estimators are used in traditional approaches) as parameter

[5] This *Q*-statistic does not appear in the output of proc mixed, but can be computed by hand. It does appear in the output produced by the software program HLM (Bryk, Raudenbush, & Congdon, 1996) as the chi-square significance test of τ.

estimates can be less stable and more prone to nonconvergence (2009). This was evident in the Marsh et al. (2009) meta-analysis when they attempted to examine moderator effects with some categorical variables only represented by a small (one, two, or three) number of studies.

In summary, the results obtained using the traditional and multilevel approaches with the Lazowski and Hulleman (2014) data would lead the researcher to the same conclusions. For example, comparisons of the parameter estimates and standard errors between the traditional and multilevel approaches for the fixed effects/no moderators, fixed effects/moderators, random effects/no moderators, and random effects/moderators models all show similar results. In fact, the same parameter estimates and standard errors are obtained in the two approaches when the same estimator is used. These results underscore the fact that the same models underlie the traditional and multilevel orientations (those in Table 1) and that essentially the same results are obtained for these two orientations, particularly when the same estimator is employed. The only concern for meta-analysts adopting a multilevel approach (or using likelihood-based estimators in the traditional approach) is the poor performance of some significance tests used to assess between-study variability in effect sizes.

Because our comparison highlighted the similarities between the two approaches, it is useful at this point to be reminded of the differences. First, there is a difference in estimation procedures. Traditional approaches more typically use MM, whereas multilevel approaches use ML or REML. Second, there is a difference in the models used. Limiting the analysis to fixed effects models is more pervasive in the traditional approach as is the investigation of moderators in separate analyses. Third, there is a difference in the software and thus, in the output. In the software used in our illustrations, a notable difference in their associated output was the presence of the kind of information you would find in a source table (e.g., Q_B, Q_W, Q_R, Q_E) in the Lispey and Wilson macro's output and the absence of this information in the output of PROC MIXED. Although this information is not provided in the output from PROC MIXED, it can easily be computed with additional code using the equations for those quantities.

Given the similarity in the results of the two approaches, it appears surprising that the traditional approach is sometimes viewed as being quite different from the multilevel approach. At the very least, traditional approaches are utilized at a much larger rate in the applied meta-analytic literature compared to multilevel approaches (Marsh et al., 2009). One reason may lay in the fact that meta-analysis, and the traditional approaches therein, emerged prior to the advent of multilevel modeling. As such, the traditional approach may simply be the recommended way of conducting meta-analyses based on historical acceptance. In addition, the adoption of more traditional approaches compared to multilevel approaches may also be a function of familiarity. More specifically, most researchers have likely taken coursework or conducted research using ANOVA or regression frameworks (i.e., required as part of a graduate degree) but far less likely to have taken coursework or have practical experience using multilevel modeling. Some researchers, then, may not be able to make the connection that the same models underlie both orientations due to this lack of exposure.

Another point related to this issue is that the equations for the statistical models underlying the analyses in the traditional approach are not always provided, though this appears customary in multilevel studies. Even if equations are provided from the traditional approach are provided, the notation might look quite different than the notation used in multilevel modeling. In both instances, it may be particularly difficult to make the connection.

Provided the similarities between these orientations, using a multilevel approach may offer some advantages that address some limitations in the traditional approach. For example, an interesting extension of the multilevel approach is the flexibility in adding more than two levels (Hox & de Leeuw, 2003). This can be particularly beneficial when there are several different outcome measures within each study. In the classical approaches discussed earlier, the options are either to average these into one single effect size per study (Lipsey & Wilson, 2001) or to conduct separate meta-analyses for each different outcome (Hedges, 1982b; Hedges & Olkin, 1985). However, using multilevel models, a researcher can specify a multivariate outcome model (Hox & de Leeuw, 2003). In this case, the multiple outcomes within a single study can be incorporated into the model using an additional, third level without violating assumptions of independence (Marsh et al., 2009).

In sum, Van den Noortgate and Onghena (2003) concluded that while the multilevel approach does not necessarily provide superior results in comparison to traditional approaches, researchers can feel confident when using multilevel approaches and can capitalize on the large amount of flexibility they provide in modeling the data. Hox and de Leeuw (2003) emphasized the flexibility of multilevel approaches in the facility of adding additional levels to the model (e.g., adding a third level to capture multiple outcomes within the same study). Additionally, multilevel meta-analysis makes it easier for a researcher to transition to Bayesian procedures, which are less sensitive to the problems that occur with small samples by including prior distributions for model parameters. This prior distribution can also be used to indicate *a priori* beliefs regarding the likelihood of publication bias and provides a method to investigate this type of common bias (2003). More traditional approaches to investigating publication bias are discussed next.

Publication Bias and the Inclusion of Published and Unpublished Data

Publication bias refers to the notion that a larger percentage of statistically significant results are likely to be published (and thus included in meta-analyses) in comparison to those found not significant or in the opposite direction of researchers' hypotheses (Sterne et al., 2000; Torgerson, 2006). Also inherent in this notion is the tendency of published studies to have larger effect sizes in comparison to unpublished studies (Durlak & Lipsey, 1991) and that sample sizes tend to be larger for published studies (Torgerson, 2006). Statistically speaking, underpowered studies with small sample sizes will need to demonstrate larger effects to be found statistically significant. This can be problematic especially in cases when a meta-analysis includes a large number of published studies with small sample sizes but large effect sizes; in this case, the likelihood of publication bias increases (Begg & Berlin, 1988). In addition, the publication bias can arise when the researchers of primary studies do not submit statistically non-significant findings for publication, leading to the "file drawer" problem (Wilson & Lipsey, 2001). Therefore, it may be easier for the meta-analyst to retrieve published studies that have statistically significant, positive results compared to non-significant or negative results. This will bias the meta-analysis in a more positive direction (over-estimating the effects) because of the overrepresentation of these published studies in the review (2001). Publication bias is therefore widely considered to be a threat to validity in meta-analyses (Torgerson, 2006). This threat has been well documented in educational research and the social sciences dating back over 50 years ago (e.g., Sterling, 1959). Discussed next are the methods to identify, assess, and address publication bias.

Publication bias can be detected using both graphical and statistical methods (Torgerson, 2006). The methods described here include the funnel plot (graphical method), the fail-safe *n* test (statistical method), and the trim-and-fill method (statistical method). In addition, the inclusion of unpublished as well as published studies in meta-analyses is strongly recommended and moderator analyses that compare both types of studies are encouraged.

Funnel plot. First, the funnel plot (Light & Pillemer, 1984) is the most commonly used method to detect publication bias (Torgerson, 2006). Funnel plots graphically depict a point estimate for each study on the x-axis (usually the effect size) against a measure of the precision for each study on the y-axis (usually the sample size or standard error) (2006). An example funnel plot based on fake data is shown in Figure 1. As Figure 1 depicts, studies that demonstrate the highest precision will be located at the top of the graph with other studies dispersed in equal measure on both sides below. Because the precision of an effect size estimate decreases as the sample size decreases, more variability is expected at the bottom of the graph where studies with smaller sample sizes are located (Soeken & Sripusanapan, 2003). Therefore, when little to no publication bias is present, the data points on the graph will look like an inverted funnel. However, when publication bias may be present, one side of the funnel will have missing data points typically on left side of the graph which would depict the absence of negative or null results among the studies. Another indication of possible publication bias would be a "hollowing out" (Torgerson, 2006, p. 97) in the center of the funnel plot suggesting that statistically significant results in either a positive or negative direction were published but not those without significant effects.

Limitations of the funnel plot are well-documented. Asymmetry in the funnel plots can be due to several factors other than publication bias. Three possible reasons include true substantive or methodological heterogeneity between the studies, data irregularities such as poor methodological design, and chance (Sterne et al., 2000). Asymmetry due to chance is more likely to occur when the number of studies included in the meta-analysis is small (e.g., < 20), and therefore, an asymmetrical funnel plot may be due to the fact that no studies with an extreme result had yet been produced (Torgerson, 2006). Another limitation of the funnel plot method is the difficulty in interpreting the findings. More specifically, because the funnel plot requires visual inspection, individuals may differ on their interpretation of the results (Soeken & Sripusanapan, 2003) or the results may simply be unclear.

Illustration: Funnel plot

A funnel plot based on the studies in the Lazowski and Hulleman (2015) meta-analysis is presented in Figure 2. First, note that there does not appear to be missing data points on the left side of the plot. Second, the funnel plot shows that studies with smaller sample sizes are distributed around the mean effect size on both sides of the distribution. Both of these features of the funnel plot suggest minimal publication bias.

Orwin's fail-safe *n* test. Another method to examine the presence of publication bias is the fail-safe *n* test. The fail-safe *n* test was first developed by Rosenthal (1979), then adapted by Orwin (1983) for use with the standardized mean difference effect size. Rosenthal first developed the fail-safe *n* for use in combining *z*-values across studies and his formula determined "the number of unpublished studies reporting null results needed to reduce the cumulated effect across studies to the point of non-significance" (Lipsey & Wilson, 2001, p. 166).

Orwin's fail-safe *n* approach determines "the number of studies with an effect size of zero needed to reduce the mean effect size to a specified or criterion level" (Lipsey & Wilson, 2001, p. 166). Therefore, in order to calculate the fail-safe *n*, the researcher must determine a criterion effect size that would be too small to be of theoretical or practical significance. Orwin (1983) recommended a *d* of 0.20 as the criterion effect size, which reflects the magnitude of an effect size conventionally considered to be small (Cohen, 1988). The fail-safe *n* provided by Orwin (1983) is calculated using the following formula:

$$N_{fs} = \frac{k(\gamma_0 - d_c)}{d_c} \tag{12}$$

where *k* represents the number of studies in the meta-analysis, γ_0 represents the weighted average effect size for the studies in the meta-analysis, and d_c represents the criterion value selected that *d* would equal (typically 0.20) when the number of hypothetical studies (N_{fs}) were added to the meta-analysis. Therefore, N_{fs} equals the number of hypothetical studies necessary to change the obtained effect size (γ_0) into a small effect size with little to no theoretical or practical significance.

The fail-safe *n* approach has limitations as well. The method assumes that the hypothetical or unpublished studies represent a random sample of all the studies that were conducted (Iyengar & Greenhouse, 1988), an assumption that is likely tenuous. Additionally, this method does not account for studies that may have a negative effect size. Assuming that unpublished studies are more likely to include negative effect sizes, the number of hypothetical studies as indicated by the fail-safe *n* may be overestimated (Soeken & Sripusanapan, 2003). Therefore, fail-safe *n* should be applied and interpreted with due caution (2003).

Illustration: Orwin's fail safe n

Using the 83 studies from the Lazowski and Hulleman (2015) meta-analysis and weighted effect size of 0.497, Orwin's fail-safe *n* suggests that an additional 123 studies with a mean effect size of zero would be needed to reduce the mean effect size to 0.20, as calculated below:

$$N_{fs} = \frac{83(0.497 - 0.20)}{0.20} = 123.255$$

The trim and fill method. The trim and fill method (Duval & Tweedie, 2000) aims to identify and adjust for funnel plot asymmetry that may be due to publication bias. The method provides an "objective approach to estimate the number of studies missing from the funnel plot (through trimming), but also a means to replace them and obtain an adjusted estimate of the overall ES (through filling)" (Soeken & Sripusanapan, 2003, p.

60). The method also assumes that studies on left hand side of the funnel plot (where effect sizes are smaller than the average effect size) are missing, and thus the method determines the number of studies that would be needed to be trimmed from the right side (where effect sizes are larger than the average) to achieve a symmetric center (Sutton, 2009).

Two estimators are used for the number of missing studies: $R_0 = \gamma^* - 1$, where γ^* represents the length of the rightmost run of ranks for positive values. The rightmost run of ranks for positive values reflects the effect size estimates that deviate the most (in a positive direction) from the average effect size estimate (2009). The second estimator is $L_0 = [4T - k(k\text{-}1)]/[2k\text{-}1]$, where T represents the sum of the positive ranks and k represents the number of studies (Soeken & Sripusanapan, 2003). The estimation process will continue until the value of missing studies stabilizes which is generally after two or 3 iterations (2003). As an example, if the estimate chosen converges to a value of 3, then the “mirror image” of the 3 largest effect sizes are filled in or added to the data and the average effect size is recalculated. Therefore, if the 3 largest effect sizes had values of 1.2, 1.1, and 1.0, then the “mirror image” of these effect sizes would be -1.2, -1.1, and -1.0, respectively.

After the number of missing studies is determined, the funnel plot is then “filled” with the missing studies around the center of the funnel plot (Higgins & Green, 2011). The adjusted average effect size can be compared to the original effect size to capture the impact of missing studies in the meta-analysis (Soeken & Sripusanapan, 2003). The average effect size adjusted following the "trimming" and will generally be smaller than the original average.

Illustration: Trim and fill method

To illustrate the trim and fill method, the Lazowski and Hulleman (2015) data were analyzed using the PubBias macro in SAS provided by Rendina-Gobioff and Kromrey (2006). Although this macro does not provide the adjusted average effect size after trimming, it does provide a test to determine the presence of publication bias for three indicators – the right tail, the left tail, and both tails. In this regard, publication bias is present when $R_0 > 3$ (2006). Results suggested that all three indicators (right tail, left tail, and both tails) indicate no publication bias in the Lazowski and Hulleman (2015) data.

Like the other methods described thus far, the trim and fill method has limitations. First, implicit in the method is an assumption that the funnel plot should be symmetrical. However, it may be difficult to determine whether the adjusted intervention effect would mirror what would have been obtained without publication bias. This is because the true reason for publication bias itself cannot be determined (Higgins & Green, 2011). Correspondingly, the trim and fill method does not account for various mechanisms behind funnel plot asymmetry other than publication bias. The adjusted average effect size estimates from the trim and fill method should therefore be interpreted with due caution (2011). The trim and fill method also been demonstrated to perform poorly in cases with substantial between-study heterogeneity (Peters, Sutton, Jones, Abrams, & Rushton, 2007; Terrin, Schmidt, Lau, & Olkin, 2003). Finally, following the trim and fill estimation procedures, the inferences drawn are based on a dataset that includes imputed effect size estimates. Some argue that imputed estimates may inappropriately contribute information that impacts the uncertainty in the overall effect size estimate (Higgins & Green, 2011).

Inclusion of unpublished (grey) literature. Although the methods described above can help the researcher detect the existence of publication bias, each of these methods suffers from limitations as described earlier. Rather than assess the existence of publication bias after the studies have been collected (post-hoc), one of the most effective ways to minimize publication is through an extensive and exhaustive search of the literature and by including unpublished, or grey, literature. The most commonly accepted definition of grey literature was operationalized at the Third International Conference on Grey Literature, defined as: "that which is produced on all levels of government, academics, business and industry in electronic and print formats not controlled by commercial publishers" (Auger, 1998). Grey literature includes, but is not limited to, unpublished reports, dissertations and theses, conference abstracts/papers, policy documents, reports to funding agencies, unpublished manuscripts (rejected or not submitted), and technical reports (Conn, Valentine, Cooper, & Rantz, 2003). The inclusion of grey literature in meta-analysis may attenuate the potential problem of publication bias and provide a more comprehensive, complete, and objective answer to the research question the meta-analyst seeks to understand (McAuley, Pham, Tugwell, & Moher, 2000). Although grey literature is more difficult to locate and retrieve in comparison to published work, the current consensus is that there is little justification for conducting meta-analyses that purposefully exclude grey literature (Rothstein &

Hopewell, 2009). The exclusion of grey literature impacts the validity and reliability of meta-analyses, especially in situations where unpublished findings differ systematically from published findings (Dickersin, 1997). These systematic differences are discussed next.

There exists evidence that grey literature differs from research published in well-known journals in sundry ways (Conn, Valentine, Cooper, & Rantz, 2003). First, the most important and persistent difference between published and unpublished work is that results from published work are more likely to be statistically significant (2003). The resulting bias against the null hypothesis (2003) has been demonstrated to exist in both the social and biomedical sciences as well as for both experimental and observational studies (Dickersin, 2005). Research has indicated this may be due to a variety of reasons, including: the tendency that studies with statistically significant results are more likely to be published in journals with high impact factors, widely distributed, and indexed in computerized databases (Begg & Berlin, 1989; Egger & Smith, 1998); the tendency for authors to only submit research that replicates previous findings (Cooper, DeNeve, & Charlton, 1997); the tendency that statistically non-significant findings are less likely (or take longer) to be published by comparison to statistically significant findings (Hopewell & Clarke, 2001); and a negative correlation between sample size and effect size in published data (Rothstein & Hopewell, 2009).

Another systematic difference between unpublished and published studies is the tendency of researchers to stop studies when preliminary or pilot studies suggest no treatment effect (Dickersin, Chan, Chalmers, Sacks, & Smith, 1987). In addition, although unpublished studies are more likely to contain smaller samples, these studies may also contain hard-to-recruit participants, novel pilot studies, or innovative interventions (Conn, Valentine, Cooper, & Rantz, 2003). It can be particularly inopportune to exclude such studies in a meta-analysis, especially given the method's ability to handle and summarize results across a range of sample sizes (2003).

One final systematic difference between unpublished and published studies is related to externally funded research. More specifically, externally funded research is more likely to be published compared to work that is not funded (2003). The importance of funding may be particularly salient in educational intervention work, where it may be quite costly to implement even small-scale studies, especially in K-12 settings. Therefore, it is likely that there is valuable grey literature from intervention studies that were not funded from external sources.

Because of these systematic differences, it is not implausible that the inclusion or exclusion of grey literature in meta-analyses will yield different results, and correspondingly, have important consequences on the interpretations of findings. Excluding grey literature attenuates the breadth of coverage of the available evidence and thus may introduce systematic error and pose a threat to validity (Moher, Cook, Eastwood, Olkin, Rennie, & Stroup, 2000). Nonetheless, a majority of meta-analyses in different fields (e.g., medical, education) do not include grey literature (McAuley, Pham, Tugwell, & Moher, 2000; Rothstein & Hopewell, 2009). Because publication bias is most directly and consistently linked with statistical significance of findings, meta-analyses that exclude grey literature risk overestimating the effect sizes associated with interventions (McAuley, Pham, Tugwell, & Moher, 2000).

Indeed, studies examining this link have demonstrated this trend. When effect sizes are broken down by publication status, published studies tend to have larger effect sizes compared to unpublished studies (Rothstein & Hopewell, 2009). Lipsey and Wilson's (1993) seminal study of meta-analyses surrounding psychological, educational, and behavioural intervention research found that estimates of treatment effects (experimental groups compared to control groups) from published studies were approximately one-third larger compared to those from unpublished studies.

More recently, Webb and Sheeran (2006) meta-analysed randomized experiments designed to influence behavioral intentions and found that published studies reported larger effect sizes for treatment vs. control conditions compared to unpublished studies, at a rate of approximately one-third of a standard deviation. They also indicated that studies without statistically significant findings were less likely to be published (33%) compared to those reporting significant findings (89%) (2006). Similarly, McLeod and Weisz (2004) examined 121 dissertations and 134 published studies in the area of youth psychotherapy and found that published studies reported effects more than twice as large as dissertations. Rothstein and Hopewell (2009) express that this is particularly noteworthy, given the dissertations were more methodologically sound and there appeared to be no differences in treatment fidelity between dissertations and published articles in the study.

In the field of education, similar trends emerge; however, the difference between unpublished and published studies is less pronounced. For example, Elbaum (2002) conducted a meta-analysis on the effects of classroom placement on self-concept for students diagnosed with learning disabilities. They found only a small effect size (d =

0.05) for published studies and no effect for unpublished studies (d = 0.00) (2002). Similarly, Swanson (1999) found no differences in publication status (published vs. unpublished studies) in a meta-analysis examining interventions designed to improve reading skills for students diagnosed with learning disabilities. In a follow-up meta-analysis using single-subject designs, Swanson and Sachse-Lee (2000) reported a similar pattern of results. The average treatment effect for published studies (d = 1.42) was only slightly larger than those found in dissertations and technical reports (d = 1.27).

Illustration: Examining moderator effects of published vs. unpublished studies

With respect to publication status, there were 71 published studies and 12 unpublished studies included in the Lazowski and Hulleman (2015) meta-analysis. A random effects regression model was estimated using the MetaReg macro with code variables in the model to explore the effects of publication status. The estimation method for τ was specified as noniterative method of moments (MM; Raudenbush, 2009). Results indicated that Q_R was significant (Q_R = 9.383, p = 0.002), suggesting the regression model explains a significant amount of variability across effect sizes. Q_E was not significant (Q_E = 78.489, p = 0.558), suggesting that the unexplained variability was not any greater than would be expected from sampling error alone. Publication status associated with the studies explained approximately 11% of the variance in effect sizes (Q_R /(Q_R + Q_E)). These results also suggest that published studies

> in the Lazowski and Hulleman (2015) meta-analysis had a significantly larger weighted effect size (0.537) in comparison to unpublished studies (0.265).

The argument against including grey literature is primarily a methodological one. More specifically, authors often justify the exclusion of unpublished studies in a meta-analysis as a quality check or quality control; in other words, they argue that unpublished studies are likely to be of lower quality or have less treatment fidelity compared to published work (Rothstein & Hopewell, 2009; Torgerson, 2006). However, there appears to be a prevailing opinion that unpublished material should be included in meta-analyses. Surveys conducted by Cook and Guyatt (1993) and more recently by Tetzlaff and her colleagues (2006) suggest that a substantial majority of meta-analysts and methodologists believe that research syntheses should include both published and unpublished studies. Both surveys, however, revealed that journal editors possessed less favorable views toward unpublished studies compared to meta-analysts and methodologists, though the Tetzlaff et al. survey demonstrated that this difference is diminishing. The less favorable views among journal editors may be due in part to the fact that the unpublished studies have not undergone peer review (McAuley et al., 2000); however, earlier studies have suggested that unpublished and published studies do not differ with regard to scientific rigor (e.g., Chalmers et al., 1990; Easterbrook et al., 1991). More recent work, however, suggests that studies in the grey literature can be difficult to assess (Hopewell, Clarke, & Mallett, 2005) but that the quality of studies included in a meta-analysis should be assessed no matter if they were retrieved from published or unpublished sources (Rothstein & Hopewell, 2009). Ultimately, however, best practices dictate that researchers be explicit and document in the meta-analysis exactly the sources that have been searched, the search strategies used, and the inclusion/exclusion criteria so that readers can evaluate the validity of the conclusions based on the search results (2009).

Conclusion

The use of meta-analysis in the fields of psychology and education has proliferated in the past 25 years and has been a useful analytic tool to inform practice and policy, identify areas that need further investigation, and provide some resolution to conflicting results among primary studies. Like most analytic tools, however, the quality of the findings and inferences drawn are dependent upon the methodological rigor and quality with which the meta-analysis was conducted. This tutorial, designed for practitioners and researchers interested in conducting meta-analyses, covered a host of methodological issues that should be considered to help inform best practices and to illuminate similarities and differences among the models typically used in meta-analytic work. In this regard, the tutorial was designed to inform the decision-making process about the type of effect size to use, the choice between fixed verses random effects models, traditional verses multilevel modeling approaches, and the importance of assessing publication bias and including grey literature in meta-analysis.

Of course, the issues presented here are not exhaustive of all issues present in meta-analysis; however, these issues are particularly salient for researchers to consider. As shown throughout the tutorial, the results and interpretation of findings may differ depending on whether a fixed or random effects model is chosen. This distinction is an important one. The choice or either a fixed effects or random effects model may have bearing not only on the overall weighted effect size but also on the results of moderator analyses. As advocated here and supported by other researchers (e.g., Aronson, Ellsworth, Carlsmith, & Gonzalez, 1990; Schmidt, Oh, & Hayes, 2009), the use of fixed effects models is rarely justified and often a random effects model is most appropriate.

In contrast, although some may believe that traditional and multilevel approaches differ dramatically, this tutorial and the work of other researchers (e.g., Hox, 2010; Van den Noorgate & Onghena, 2003) demonstrated that the two approaches produce similar results and that the models are essentially the same. Differences between the two approaches largely surround the estimation procedures used to estimate between-study variance – method of moments (MM) are most commonly used in the traditional approach, whereas maximum likelihood (ML) procedures are most common in the multilevel approaches. Perhaps the largest difference is the model used as the starting point for analyses. In the traditional approach, a researcher typically begins with a fixed effects model, then moves to a random effects model if significant heterogeneity exists among the effect sizes. On the other hand, in the multilevel approach, a researcher typically begins with the random effects model from the start.

Finally, the importance of assessing publication bias, and the most common methods for evaluating publication bias, were discussed. Given that each method for assessing publication bias has limitations, it is critically important that researchers conducting meta-analyses search for both published and unpublished literature. The inclusion of unpublished literature may provide a more thorough and exhaustive breadth of studies that were conducted on a given topic. As well, some research suggests (and evidenced through the illustration presented in this tutorial) that the effect sizes associated with unpublished studies are smaller than those from published studies. In this regard, including only published studies lends risk to an overestimation of the true effect size.

Given the potential power of meta-analyses to inform policy, practice, theory, and research, it is paramount that the meta-analyses themselves be conducted with appropriate rigor and sound methodology. This tutorial was intended to contribute to the growing body of literature surrounding best practices in meta-analysis through an instructional, illustrative manner so that researchers and practitioners alike are better equipped to use the meta-analytic tool as part of their existing toolbox.

References

References marked with an asterisk indicate studies included in the Lazowski and Hulleman (2015) meta-analysis.

* Acee, T. W., & Weinstein, C. E. (2010). Effects of a value-reappraisal intervention on statistics students' motivation and performance. *The Journal of Experimental Education*, *78*, 487-512.

Alexander, R. A., Scozzaro, M. J., & Borodkin, L. J. (1989). Statistical and empirical examination of the chi-square test for homogeneity of correlation and meta-analysis. *Psychological Bulletin, 106,* 329-331.

Aronson, E., Ellsworth, P., Carlsmith, J., & Gonzalez, M. (1990). *Methods of research in social psychology* (Second Edition). Reading, MA: Addison-Wesley.

* Aronson, J., Fried, C. B., & Good, C. (2002). Reducing the effects of stereotype threat on african american college students by shaping theories of intelligence. *Journal of Experimental Social Psychology*, *38*, 113-125.

Auger, C. P. (1998). *Information sources in grey literature.* Guides to Information Sources, 4th ed., London: Bowker Saur.

Baldwin, S., & Shadish, W. (2011). A primer on meta-analysis in clinical psychology. *Journal of Experimental Psychopathology, 2*, 294–317.

Begg, C. B., & Berlin, J. A. (1988). Publication bias: a problem in interpreting medical data. *Journal of the Royal Statistical Society, Series A (Statistics in Society), 151*(3), 419-463.

Beretvas, S. N. (2010). Meta-analysis. In G.R. Hancock, R. O. Mueller (Eds.). *The* *reviewer's guide to quantitative methods in the social sciences (pp. 256-263). New York, NY: Routledge.*

*Bernacki, M., Nokes-Malach, T., Richey, J. E., & Belenky, D. M. (2014). Science diaries: A brief writing intervention to improve motivation to learn science, *Educational Psychology: An International Journal of Experimental Educational Psychology,* doi:10.1080/01443410.2014.895293.

* Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development*, *78*, 246-263.

*Boese, G. D., Stewart, T. L., Perry, R. P., & Hamm, J. M. (2013). Assisting failure-prone individuals to navigate achievement transitions using a cognitive motivation treatment (attributional retraining). *Journal of Applied Social Psychology*, *43*, 1946-1955.

*Bordine, J. J. (1998). *Improving reading, self-concept, and motivation in at-risk kindergarten students: An intervention* (Unpublished doctoral dissertation). The University of Texas at Austin, Austin, TX.

Borenstein, M. (2009). Effect sizes for continuous data. *The handbook of research synthesis and meta-analysis, 2, 221-235.*

Bryk, A.S., Raudenbush, S.W., & Congdon, R. (1996). HLM 4 for Windows [Computer software]. Chicago, IL: Scientific Software International, Inc.

Chalmers, I., Adams, M., Dickersin, K., Hetherington, J., Tarnow-Mordi, W., Meinert, C. (1990). A cohort study of summary reports of controlled trials. *Journal of the American Medical Association, 263,* 1401-1405.

Cheung, M. W.-L. (2008). A model for integrating fixed-, random-, and mixed-effects meta-analyses into structural equation modeling. *Psychological Methods, 13*, 182-202.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10*, 101-129.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

* Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science*, *313*, 1307-1310.

* Cohen, G. L., Garcia, J., Purdie-Vaughns, V., Apfel, N., & Brzustoski, P. (2009). Recursive processes in self-affirmation: Intervening to close the minority achievement gap. *Science, 324*, 400-403.

Conley, A. M. (2012). Patterns of motivation beliefs: Combining achievement goal and expectancy-value perspectives. *Journal of Educational Psychology, 104(1)*, 32–47.

Conn, V. S., Valentine, J. C., Cooper, H. M., & Rantz, M. J. (2003). Grey literature in meta-analysis. *Nursing Research, 52*(4), 256-261.

Connell, J. P., Klem, A. M., Lacher, T., Leiderman, S., & Moore, W., with Deci, E. L. (2009). *First Things First: Theory, research, and practice*. Toms River, NJ: Institute for Research and Reform in Education. Available online: http://www.irre.org/publications

Cook, D. J., & Guyatt, G. H. (1993). Should unpublished data be included in meta-analyses? *Journal of the American Medical Association, 269*(21), 2749-2753.

* Cook, J. E., Purdie-Vaughns, V., Garcia, J., & Cohen, G. L. (2012). Chronic threat and contingent belonging: Protective benefits of values affirmation on identity development. *Journal of Personality and Social Psychology*, *102*, 479-496.

Cooper, H. M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society*, *1*(1), 104-126.

Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods, 2,*447-452.

Cooper, H., & Hedges, L. V. (2009). Research synthesis as a scientific process. *The handbook of research synthesis and meta-analysis, 2, 3-17.*

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis*. New York, NY: Russell Sage Foundation.

*Craven, R. G., Marsh, H. W., & Debus, R. L. (1991). Effects of internally focused feedback and attributional feedback on enhancement of academic self-concept. *Journal of Educational Psychology*, *83*, 377-392.

*Cueva, C. (2006). *An achievement motivation program for young minority students utilizing cognitive-behavioral techniques* (Unpublished doctoral dissertation). Texas A & M, Corpus Christi, TX.

Davies, P. (2000). The relevance of systematic reviews to educational policy and practice. *Oxford Review of Education, 26,* 365-378.

* Day, J. D., Borkowski, J. G., Punzo, D., & Howsepian, B. (1994). Enhancing possible selves in mexican american students. *Motivation and Emotion*, *18*, 79-103.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled*

Clinical Trials, 7, 177-188.

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.),*Publication bias in meta-analysis: Prevention, assessment and adjustments* (p. 11-33). London: John Wiley & Sons.

Dickersin, K. (1997). How important is publication bias? A synthesis of available data, *AIDS Education & Prevention, 9(Supp Al),* 15-21.

Dickersin, K., Chan, S., Chalmers, T. C., Sacks, H. S., & Smith, H. Jr. (1987). Publication bias and clinical trials. *Controlled Clinical Trials, 8,* 343-353.

*Duckworth, A.L., Kirby, T., Gollwitzer, A., & Oettingen, G. (in press). From fantasy to action: Mental contrasting with implementation intentions (MCII) improves academic performance in children.

*Durik, A. M., Schmidt, J., Shumow, L., Rodenbeck, B. (April, 2014). *Self-generated utility among a diverse sample of adolescent students: An analysis of grade level and gender.* Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Durlak, J. A. & Lipsey, M. W. (1991). A practitioner's guide to meta-analysis. *American Journal of Community Psychology, 19*(3), 291-332.

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.

Easterbrook, P. J., Berlin, J. A., Gopalan, R., & Matthews, D. R. (1991). Publication bias in clinical research. *Lancet, 337,* 867-872.

Easton, J.Q. (2013, June). *Using measurement as leverage between developmental research and educational practice*. Center for the Advanced Study of Teaching and Learning: Charlottesville, Virginia.

Engles, E. A., Schmidt, C. H., Terrin, N., Olkin, I., & Lau, J. (2000). Heterogeneity and statistical significance in meta-analysis: An empirical study of 125 meta-analyses. *Statistics in Medicine, 19*(13), 1707-1728.

Erez, A., Bloom, M. C., Wells, M. T. (1996). Using random rather than fixed effects models in meta-analysis: Implications for situational specificity and validity generalization. *Personnel Psychology, 49*, 275-306.

Fleiss, J. L., & Berlin, J. A. (2009). Effect sizes for dichotomous data. *The handbook of research synthesis and meta-analysis, 2*, *237-253.*

*Fordyce, M. W. (1983). A program to increase happiness: Further studies. *Journal of Counseling Psychology*, *30*, 483-498.

*Fordyce, M. W. (1977). Development of a program to increase personal happiness. *Journal of Counseling Psychology*, *24*, 511-521.

* Froiland, J. M. (2011, April). Parental autonomy support and student learning goals: A preliminary examination of an intrinsic motivation intervention. In *Child & Youth Care Forum* (Vol. 40, No. 2, pp. 135-149). Springer US.

*Gehlbach, H., Brinkworth, M. E., King, A., Hsu, L., McIntyre, J., & Rogers, T. (2014). *Creating birds of similar feathers: Leveraging similarity to improve teacher-student relationships and academic achievement.* Manuscript submitted for publication.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3-8.

*Good, C., Aronson, J., & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology*, *24*, 645-662.

* Greenstein, T. (1976). Behavior change through value self-confrontation: A field experiment. *Journal of Personality and Social Psychology*, *34*, 254-262.

* Guthrie, J. T., Wigfield, A., Humenick, N. M., Perencevich, K. C., Taboada, A., & Barbosa, P. (2006). Influences of stimulating tasks on reading motivation and comprehension. *The Journal of Educational Research*, *99*, 232-246.

* Hall, N. C., Perry, R. P., Goetz, T., Ruthig, J. C., Stupnisky, R. H., & Newall, N. E. (2007). Attributional retraining and elaborative learning: Improving academic development through writing-based interventions. *Learning and Individual Differences*, *17*, 280-290.

* Hall, N. C., Hladkyj, S., Perry, R. P., & Ruthig, J. C. (2004). The role of attributional retraining and elaborative learning in college students' academic development. *The Journal of Social Psychology*, *144*, 591-612.

* Harackiewicz, J. M., Rozek, C. S., Hulleman, C. S., & Hyde, J. S. (2012). Helping parents to motivate adolescents in mathematics and science: An experimental test of a utility value intervention. *Psychological Science*, *23*, 899-906.

*Hausmann, L. R., Ye, F., Schofield, J. W., & Woods, R. L. (2009). Sense of belonging and persistence in White and African American first-year students. *Research in Higher Education*, *50*, 649-669.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 7*, 119-137.

Hedges, L. V. (1982a). Estimating effect size from a series of independent experiments. *Psychological Bulletin, 92,* 490-499.

Hedges, L. V. (1982b). Fitting continuous models to effect size data. *Journal of Educational Statistics, 7(4),* 245-270.

Hedges, L. V. (1983a). A random-effects model for effect sizes. *Psychological Bulletin, 93,* 388-395.

Hedges, L. V. (1983b). Combining independent estimators in research synthesis. *British Journal of Mathematical and Statistical Psychology, 36*, 123-131.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* London, England: Academic Press Inc.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*, 486-504.

Higgins, J. P. T., & Green, S. (2011). Cochrane handbook for systematic reviews of interventions. Retrieved from www.cochrane-handbook.org

Higgins, J.P.T & Thompson, S.G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*, 1539-1558.

*Hidi, S., Berndorff, D., & Ainley, M. (2002). Children's argument writing, interest and self-efficacy: An intervention study. *Learning and Instruction*, *12*, 429-446.

*Hong, H., & Lin-Siegler, X. (2011). How learning about scientists' struggles influences students' interest and learning in physics. *Journal of Educational Psychology*. Advance online publication. doi:10.1037/a0026224.

Hopewell, S., & Clarke, M. (2001). Methodologists and their methods. Do methodologists write up their conference presentations or is it just 15 minutes of fame? *International Journal of Technology Assessment in Health Care, 17*(4), 601-603.

Hopewell, S. Clarke, M., & Mallett, S. (2005). Grey literature and systematic reviews. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments* (p. 48-72). Chichester, UK: Wiley.

Hopewell, S., McDonald, S., Clarke, M., & Egger, M. (2007). Grey literature in meta-analyses of randomized trials of health care interventions. *Cochrane Database Syst Rev*, *2*(2).

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

Hox, J. J., & de Leeuw, E. D. (2003). Multilevel models for meta-analysis. In S. P. Reise & N. Duan (Eds.), *Multilevel modelling: Methodological advances, issues, and applications* (pp. 90–111). Mahwah, NJ: Lawrence Erlbaum.

* Hoyert, M. S., & O'Dell, C. D. (2006). A brief intervention to aid struggling students: A case of too much motivation? *Journal of Scholarship of Teaching and Learning*, *6*(1), 1-13.

* Hulleman, C. S., Godes, O., Hendricks, B. L., & Harackiewicz, J. M. (2010). Enhancing interest and performance with a utility value intervention. *Journal of Educational Psychology*, *102*, 880-895.

* Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science*, *326*, 1410-1412.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8*(4), 275-292.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. New York, NY: Sage Publications, Inc.

IBM Corp. (2012). *IBM SPSS Statistics for Windows, Version 21.0.* Armonk, NY: IBM.

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem, *Statistical Science, 3*(1), 109-117.

*Jamieson, J. P., Mendes, W. B., Blackstock, E., & Schmader, T. (2010). Turning the knots in your stomach into bows: Reappraising arousal improves performance on the GRE. *Journal of Experimental Social Psychology*, *46*, 208-212.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17*, 137-152.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746-759.

* Kitsantas, A., Robert, A. R., & Doster, J. (2004). Developing self-regulated learners: Goal setting, self-evaluation, and organizational signals during acquisition of procedural skills. *The Journal of Experimental Education*, *72*, 269-287.

Konstantopoulos, S. & Hedges, L. V. (2009). Analyzing effect sizes: Fixed effects models. In H. Cooper, L. V. Hedges, and J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis, 2nd Edition* (p. 279-293). New York: Russell Sage Foundation.

Lazowski, R. A., Hulleman, C. S. (2015, April). *Motivation interventions in* *education: A meta-analytic review. Paper presented at the annual conference of the American Educational Research Association, Chicago, IL.*

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. Confirmation of from meta-analysis. *American Psychologist, 48*(12)*,* 1181-1209.

Lipsey, M.W., & Wilson, D. B. (2001). *Practical meta-analysis*. Sage: Thousand Oaks, CA.

Marsh, H. W., Bornmann, L., Mutz, R., Daniel, H., & O'Mara, A. (2009). Gender effects in the peer reviews of grant proposals: A comprehensive meta-analysis comparing traditional and multilevel approaches. *Review of Educational Research, 79*(3), 1290-1326.

* Martin, A. J. (2008). Enhancing student motivation and engagement: The effects of a multidimensional intervention. *Contemporary Educational Psychology*, *33*, 239-269.

McAuley, L., Pham, B., Tugwell, P., & Moher, D. (2000). Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? *Lancet, 356*(9237), 1228-1231.

*McGinley, J. J., & Jones, B. D. (2014). A brief instructional intervention to increase students' motivation on the first day of class. *Teaching of Psychology, 41*, 158-162.

McLeod, B. D., Weisz, J. R. (2004). Using dissertations to examine potential bias in child and adolescent clinical trials. *Journal of Consulting and Clinical Psychology, 72*(2), 235-251.

Messick, S. (1990). *Validity of test interpretation and use.* Research Rep. No. 90-11. Princeton, NJ: Educational Testing Service.

* Miyake, A., Kost-Smith, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L., & Ito, T. A. (2010). Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science*, *330*, 1234-1237.

Moher, D., Cook, D. J., Eastwood, S., Olkin, I., & Stroop, D. E. (2000). Improving quality of reports of meta-analyses of randomized controlled trials: The QUOROM statement. QUOROM Group. *British Journal of Surgery, 87,* 1448-1454.

* Morisano, D., Hirsh, J. B., Peterson, J. B., Pihl, R. O., & Shore, B. M. (2010). Setting, elaborating, and reflecting on personal goals improves academic performance. *Journal of Applied Psychology*, *95*, 255-264.

*Muis, K. R., Ranellucci, J., Franco, G. M., & Crippen, K. J. (2013). The interactive effects of personal achievement goals and performance feedback in an undergraduate science class. *The Journal of Experimental Education*, *81*, 556-578.

National Research Council (1992). *Combining information: Statistical issues and opportunities for research. Washington, DC: National Academy Press.*

Orwin, R.G. (1983). A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics, 8*, 157-159.

*Oyserman, D., Bybee, D., & Terry, K. (2006). Possible selves and academic outcomes: How and when possible selves impel action. *Journal of Personality and Social Psychology, 91*, 188-204.

* Oyserman, D., Terry, K., & Bybee, D. (2002). A possible selves intervention to enhance school involvement. *Journal of Adolescence*, *25*, 313-326.

* Patall, E. A., Cooper, H., & Wynn, S. R. (2010). The effectiveness and relative importance of choice in the classroom. *Journal of Educational Psychology*,*102*, 896-915.

*Paunesku, D., Walton, G., Romero, C., Smith, E., Yeager, D., Dweck, C. (2014). *Mindset interventions are a scalable treatment for academic underachievement.* Manuscript submitted for publication.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2007). Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine*, *26*(25), 4544-4562.

*Pugh, K. J. (2014). *Supporting deep engagement: Development and evaluation of the teaching for transformative experiences in science (TTES) model.* Manuscript under review.

*Quintanilla, Y. T. (2007). *Achievement motivation strategies: An integrative achievement motivation program with first year seminar students* (Unpublished doctoral dissertation), Texas A & M University, Corpus Christi, TX.

*Radil, A. I. (2012). *An autonomy support motivation intervention with pre-services teachers: Do the strategies that they intend to use change?* (Unpublished master's thesis), University of Alberta, Ottawa, ON, Canada.

*Ramirez, G., & Beilock, S. L. (2011). Writing about testing worries boosts exam performance in the classroom. *Science*, *331*, 211-213.

*Ranellucci, J., Hall, N. C., K. R., & Lajoie, S. P. (2014). *Mastery, maladaptive learning behaviors, and academic achievement: An intervention approach.* Manuscript under review.

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. *The Handbook of Research Synthesis and Meta-Analysis, 2, 221-235.*

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

* Reeve, J., Jang, H., Carrell, D., Jeon, S., & Barch, J. (2004). Enhancing students' engagement by increasing teachers' autonomy support. *Motivation and Emotion*, *28*, 147-169.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*, 638-641.

Rosenthal, R. (1994). Statistically describing and combing studies. In H. Cooper & L.V Hedges (Eds.), *The handbook of research synthesis* (p. 231-244). New York: Russell Sage Foundation.

Rothstein, H. R., & Hopewell, S. (2009). Grey literature. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (p. 103-122). New York: Russell Sage Foundation.

* Ruthig, J. C., Perry, R. P., Hall, N. C., & Hladkyj, S. (2004). Optimism and attributional retraining: Longitudinal effects on academic achievement, test anxiety, and voluntary course withdrawal in college students. *Journal of Applied Social Psychology*, *34*, 709-730.

(2011). *SAS 9.3 formats and informats: reference.* Cary, N.C.: SAS Institute.

* Schaffner, E., & Schiefele, U. (2007). The effect of experimental manipulation of student motivation on the situational representation of text. *Learning and Instruction*, *17*, 755-772.

Schmidt F. L., Oh, I. & 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,* 97-128.

*Sherman, D. K., Hartson, K. A., Binning, K. R., Purdie-Vaughns, V., Garcia, J., Taborsky-Barba, S., & Cohen, G. L. (2013). Deflecting the trajectory and changing the narrative: How self-affirmation affects academic performance and motivation under identity threat. *Journal of Personality and Social Psychology.* Advance online publication.

*Siegle, D., & McCoach, D. B. (2007). Increasing student mathematics self-efficacy through teacher training. *Journal of Advanced Academics, 18*, 278-312.

Soeken, K. L., & Sripusanapan, A. (2003). Assessing publication bias in meta-analysis. *Nursing Research, 52*, 57-60.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa. *Journal of American Statistical Association, 54*, 30-34.

Sterne, A. C., Gavaghan, D. & Egger, M. (2000). Publication and related bias in meta-analysis. *Journal of Clinical Epidemiology, 53*(11), 1119-1129.

*Struthers, C. W., & Perry, R. P. (1996). Attributional style, attributional retraining, and inoculation against motivational deficits. *Social Psychology of Education*, *1*, 171-187.

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, and J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis, 2nd Edition* (p. 435-452). New York: Russell Sage Foundation.

*Sverdlik, A., & Hall, N. C. (2013, August). *"Be realistic": A mixed method analysis of a motivational intervention.* Poster presented at the meeting of the American Psychological Association, Honolulu, Hawaii. *withdrawn

Swanson, H. L., & Sachse-Lee, C. (2000). A meta-analysis of single-subject-design Intervention research for students with LD. *Journal of Learning Disabilities, 33*(2), 114-136.

Swanson, H. L. (1999). Reading research for students with LD: A meta-analysis of Intervention outcomes. *Journal of Learning Disabilities, 32*(6), 504-532.

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.

Tetzlaff, J., Moher, D., Pham, B., & Altman, D. (October, 2006). *Survey of views on including grey literature in systematic reviews. Paper presented at the XIV Cochrane Colloquium, Dublin, Ireland.*

Torgerson, C. J. (2006). Publication bias: The achilles' heel of systematic reviews? *British Journal of Educational Studies, 54*, 89-102.

Van Den Noortgate, W., & Onghena, P. (2003). Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational and Psychological Measurement*, *63*(5), 765-790.

*Vansteenkiste, M., Simons, J., Lens, W., Soenens, B., & Matos, L. (2005). Examining the motivational impact of intrinsic versus extrinsic goal framing and autonomy-supportive versus internally controlling communication style on early adolescents' academic achievement. *Child Development*, *76*, 483-501.

*Vansteenkiste, M., Simons, J., Lens, W., Sheldon, K. M., & Deci, E. L. (2004). Motivating learning, performance, and persistence: The synergistic effects of intrinsic goal contents and autonomy-supportive contexts. *Journal of Personality and Social Psychology*, *87*, 246-260.

*Vansteenkiste, M., Simons, J., Lens, W., Soenens, B., Matos, L., & Lacante, M. (2004). Less is sometimes more: Goal content matters. *Journal of Educational Psychology*, *96*, 755-764.

*Vansteenkiste, M., Timmermans, T., Lens, W., Soenens, B., & Van den Broeck, A. (2008). Does extrinsic goal framing enhance extrinsic goal-oriented individuals' learning and performance? An experimental test of the match perspective versus self-determination theory. *Journal of Educational Psychology*, *100*, 387-397.

Viswesvaran, C., & Sanchez, J. I. (1998). Moderator search in meta-analysis: A review and cautionary note on existing approaches. *Educational and Psychological Measurement, 58*, 77-87.

* Walton, G. M., & Cohen, G. L. (2007). A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology*, *92*, 82-96.

* Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science*, *331*, 1447-1451.

*Wilson, T. D., & Linville, P. W. (1982). Improving the academic performance of college freshmen: Attribution therapy revisited. *Journal of Personality and Social Psychology*, *42*, 367-376.

* Wilson, T. D., & Linville, P. W. (1985). Improving the performance of college freshmen with attributional techniques. *Journal of Personality and Social Psychology*, *49*, 287-293.

Wilson, D. B., & Lipsey, M. W. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis, *Psychological Methods, 6*(4), 413-429.

*Yeager, D. S., Johnson, R., Spitzer, B. J., Trzesniewski, K. H., Powers, J., Dweck, C. S. (2014). The far-reaching effects of believing people can change: Implicit theories of personality shape stress, health, and achievement during adolescence. *Journal of Personality and Social Psychology.* Advanced online publication. doi:10.1037/a0036335.

*Yeager, D. S., Miu, A. S., Powers, J., & Dweck, C. S. (2013). Implicit theories of personality and attributions of hostile intent: A meta-analysis, an experiment, and a longitudinal intervention. *Child Development,* 1-17.

*Yeager, D. S., Purdie-Vaughns, V., Garcia, J., Apfel, N., Brzustoski, P., Master, A., & Cohen, G. L. (2013). Breaking the cycle of mistrust: Wise interventions to provide critical feedback across the racial divide. *Journal of Experimental Psychology: General*.

*Yeager, D. S., Walton, G. M., Brady, S. T., Akcinar, E. N., Paunesku, D., Keane, L., . . . Dweck, C. S. (2014). *Brief psychological interventions can reduce post-secondary achievement gaps at scale.* Manuscript submitted for publication.

Table 1

Summary Table of Motivation Intervention Studies

Study	Theory	Avg. d	n_e, n_c [a]	Sampling Variance	Grade[b]	Exp. Design
Fordyce (1983)						
Study 1	Ach. Emotions	0.36	64, 39	0.053	PS	Quasi
Study 2	Ach. Emotions	0.34	50, 21	0.064	PS	Quasi
Study 3	Ach. Emotions	0.68	44, 13	0.070	PS	Quasi
Fordyce (1977)						
Study 1	Ach. Emotions	0.52	68, 27	0.042	PS	Quasi
Study 2	Ach. Emotions	0.70	39, 29	0.068	PS	Randomized
Study 3	Ach. Emotions	1.01	42, 26	0.104	PS	Quasi
Ramirez & Beilock (2011)	Ach. Emotions	0.57	27, 26	0.079	HS	Randomized
Hoyert & O'Dell (2006)	Ach. Goal Theory	0.93	69, 68	0.032	PS	Randomized
Muis et al. (2013)	Ach. Goal Theory	0.12	198, 52	0.024	PS	Randomized
Ranellucci et al. (unpublished)	Ach. Goal Theory	-0.04	135, 42	0.031	PS	Randomized
Boese et al. (2013)	Attribution	0.77	84, 42	0.038	PS	Randomized
Good et al. (2003)						
Attribution vs. Control	Attribution	1.11	34, 35	0.067	MS	Randomized
Hall et al. (2007)	Attribution	0.28	374, 375	0.005	PS	Quasi
Hall et al. (2004)	Attribution	0.43	101, 102	0.020	PS	Quasi
Ruthig et al. (2004)	Attribution	0.59	118, 118	0.018	PS	Quasi
Struthers & Perry (1996)	Attribution	0.41	108, 150	0.016	PS	Randomized
Wilson & Linville (1985)						
Replication 1	Attribution	0.25	20, 20	0.100	PS	Randomized
Replication 2	Attribution	0.13	20, 20	0.100	PS	Randomized
Wilson & Linville (1982)	Attribution	0.73	20, 20	0.107	PS	Randomized
Yeager et al. (2013)						
Study 1	Attribution	0.76	22, 22	0.097	M	Randomized
Study 2	Attribution	0.78	22, 22	0.098	M	Randomized
Study 3	Attribution	0.44	38, 38	0.054	HS	Randomized

Study	Theory	Avg. d	n_e, n_c[a]	Sampling Variance	Grade[b]	Exp. Design
Durik et al. (unpublished)	Expectancy-Value	0.10	158, 157	0.013	MS	Randomized
Harackiewicz et al. (2012)	Expectancy-Value	0.32	94, 94	0.022	HS	Randomized
Hulleman et al. (2010), Study 2	Expectancy-Value	0.38	160, 158	0.013	PS	Randomized
Hulleman & Harackiewicz (2009)	Expectancy-Value	0.27	136, 126	0.015	HS	Randomized
Silva et al. (2011)	Goal Setting	0.71	20, 21	0.104	HS	Randomized
Sverdlik & Hall (unpublished)	Goal Setting	0.67	25, 27	0.081	PS	Randomized
Aronson et al. (2002)	Impl. Theories of Int.	0.53	37, 37	0.056	PS	Randomized
Blackwell et al. (2007)	Impl. Theories of Int.	0.69	49, 50	0.047	MS	Randomized
Good et al. (2003)						
Implicit Theories vs. Control	Impl. Theories of Int.	0.76	34, 35	0.062	MS	Randomized
Yeager et al. (2014)						
Study 2	Impl. Theories of Int.	0.43	39, 39	0.052	HS	Randomized
Study 3	Impl. Theories of Int.	0.36	75, 75	0.027	HS	Randomized
Yeager et al. (2013)						
Study 3	Impl. Theories of Int.	0.65	39, 39	0.054	HS	Randomized
Guthrie et al. (2006)	Interest	0.71	49, 49	0.043	ES	Quasi
Hidi et al. (2002)	Interest	0.67	90, 90	0.023	MS	Quasi
Acee & Weinstein	Multiple Perspectives	0.63	41, 41	0.051	PS	Quasi
Bernacki, et al. (2014)	Multiple Perspectives	0.54	26, 27	0.078	MS	Randomized
Bordine (unpublished)	Multiple Perspectives	0.49	16, 15	0.133	ES	Randomized
Craven et al. (1991)	Multiple Perspectives	0.08	81, 79	0.025	ES	Randomized
Cueva (unpublished)	Multiple Perspectives	0.18	33, 31	0.063	ES	Quasi
Duckworth et al. (in press)	Multiple Perspectives	0.51	38, 39	0.054	ES	Randomized
Good et al. (2003)						
Combined vs. Control	Multiple Perspectives	0.79	34, 35	0.063	MS	Randomized
Hong & Lin-Siegler (2011)	Multiple Perspectives	0.41	88, 93	0.023	HS	Randomized
Jamieson et al. (2010)	Multiple Perspectives	0.87	14, 14	0.156	PS	Randomized
Kitsantas et al. (2004)	Multiple Perspectives	1.14	48, 48	0.048	HS	Randomized
Martin (2008)	Multiple Perspectives	0.48	26, 27	0.078	HS	Quasi
McGinley & Jones (2014)	Multiple Perspectives	0.37	58, 53	0.037	PS	Randomized

Study	Theory	Avg. d	n_e, n_c [a]	Sampling Variance	Grade [b]	Exp. Design
Morisano et al. (2010)	Multiple Perspectives	0.75	43, 42	0.050	PS	Randomized
Paunesku et al. (unpublished)	Multiple Perspectives	0.14	1196, 398	0.003	HS	Randomized
Siegle & McCoach (2007)	Multiple Perspectives	0.33	430, 442	0.005	ES	Randomized
Yeager et al. (unpublished)						
Study 1	Multiple Perspectives	0.24	310, 274	0.007	PS	Randomized
Study 2	Multiple Perspectives	0.13	5281, 16278	0.000	PS	Randomized
Study 3	Multiple Perspectives	0.23	1186, 406	0.003	PS	Randomized
Quintenilla (unpublished)	Need for Ach.	0.36	47, 46	0.044	PS	Randomized
Day et al. (1994)	Possible Selves	0.91	42, 41	0.053	ES	Randomized
Oyserman et al. (2006)	Possible Selves	0.35	141, 123	0.015	HS	Randomized
Oyserman et al. (2002)	Possible Selves	0.37	62, 146	0.020	MS	Quasi
Cohen et al. (2009)	Self-Affirmation	0.52	192, 193	0.011	MS	Randomized
Cohen et al. (2006)	Self-Affirmation	0.37	104, 104	0.018	MS	Randomized
Cook et al. (2012), Study 2	Self-Affirmation	0.36	61, 60	0.034	MS	Randomized
Miyake et al (2010)						
Male Comparison	Self-Affirmation	-0.16	178, 105	0.015	PS	Randomized
Female Comparison	Self-Affirmation	0.21	69, 47	0.036	PS	Randomized
Sherman et al. (2013)						
Study 1	Self-Affirmation	0.34	41, 40	0.050	MS	Randomized
Study 2	Self-Affirmation	0.64	26, 29	0.077	MS	Randomized
Walton & Cohen (2011)	Self-Affirmation	0.52	49, 43	0.045	PS	Randomized
Greenstein (1976)	Self-Confrontation	0.54	87, 84	0.024	PS	Randomized
Froiland (2011)	Self-Determination	0.71	15, 15	0.142	ES	Quasi
Patall et al. (2010)	Self-Determination	0.12	193, 194	0.010	HS	Randomized
Radil (unpublished)	Self-Determination	0.43	25, 37	0.069	PS	Quasi
Reeve et al. (2004)	Self-Determination	1.94 [c]	10, 10	0.241	HS	Randomized
Schaffner & Schiefele (2007)	Self-Determination	0.46	188, 187	0.011	HS	Randomized
Vansteenkiste et al. (2008) Vansteenkiste et al. (2005)	Self-Determination	0.70	68, 70	0.031	MS	Randomized

Study	Theory	Avg. d	n_e, n_c[a]	Sampling Variance	Grade[b]	Exp. Design
Study 1	Self-Determination	0.83	65, 65	0.033	MS	Randomized
Study 3	Self-Determination	0.74	57, 56	0.038	MS	Randomized
Vansteenkiste et al. (2004)						
Study 1	Self-Determination	1.57[c]	100, 100	0.024	PS	Quasi
Study 2	Self-Determination	1.49[c]	189, 189	0.013	PS	Quasi
Vansteenkiste et al. (2004)	Self-Determination	0.42	123, 122	0.017	PS	Randomized
Gehlbach et al. (unpublished)	Social Belongingness	0.15	194, 60	0.023	HS	Randomized
Hausmann et al. (2009)						
Exp. vs. Control (White)	Social Belongingness	0.26	70, 67	0.029	PS	Randomized
Exp. vs. Control (Afr. Amer)	Social Belongingness	-0.04	41, 42	0.048	PS	Randomized
Walton & Cohen (2007)						
Study 1	Social Belongingness	0.91	18, 18	0.122	PS	Randomized
Study 2	Social Belongingness	1.57[c]	18, 18	0.134	PS	Randomized
Pugh (unpublished)	Transformative Exp.	0.67	76, 82	0.027	MS	Randomized
Total		0.49[d]	14200, 23039			

Note: Ach. Emotions = Achievement Emotions; Impl. Theories of Int. = Implicit Theories of Intelligence.
[a]The sample size for the experimental condition (n_e) is reported first, followed by the sample size for the control condition (n_c).
[b]Grade included Elementary School (ES), Middle School (MS), High School (HS), and Post-Secondary (PS).
[c]Extreme outliers were Windsorized and adjusted to 3 standard deviations from the effect size mean.
[d]Mean Effect Size calculated via macro (meanes.sps) provided by Lipsey and Wilson (2001).

Table 2

Fixed-effects and Random-effects Models With and Without Moderators

		Fixed Effects Cell A[a]	Random Effects Cell C[c]
No Moderators	Level 1	$d_j = \delta_j + e_j \quad e_j \sim N(0, v_j)$	$d_j = \delta_j + e_j \quad e_j \sim N(0, v_j)$
	Level 2	$\delta_j = \gamma_0$	$\delta_j = \gamma_0 + u_j \quad u_j \sim N(0, \tau)$
	Combined	$d_j = \gamma_0 + e_j$	$d_j = \gamma_0 + u_j + e_j$
	Inverse variance weight	$1 / v_d$	$1/(v_j + \tau)$

		Fixed Effects Cell B[b]	Random Effects Cell D
Moderators	Level 1	$d_j = \delta_j + e_j \quad e_j \sim N(0, v_j)$	$d_j = \delta_j + e_j \quad e_j \sim N(0, v_j)$
	Level 2	$\delta_j = \gamma_0 + \gamma_1 X_{1j} + \gamma_2 X_{2j} + \ldots \gamma_p X_{pj}$	$\delta_j = \gamma_0 + \gamma_1 X_{1j} + \gamma_2 X_{2j} + \ldots \gamma_p X_{pj} + u_{0j} \quad u_j \sim N(0, \tau)$
	Combined	$d_j = \gamma_0 + \gamma_1 X_{1j} + \gamma_2 X_{2j} + \ldots \gamma_p X_{pj} + e_j$	$d_j = \gamma_0 + \gamma_1 X_{1j} + \gamma_2 X_{2j} + \ldots \gamma_p X_{pj} + u_{0j} + e_{ij}$
	Inverse variance weight	$1 / v_d$	$1/(v_j + \tau)$

[a] d_j reflects the observed effect size in study *j*, δ_j reflects the true effect size for study *j*, γ_0 reflects the population effect size, and e_j reflects the residual due to sampling error (Hedges & Olkin, 1985). Errors of estimation e_j are assumed to be statistically independent, each with a mean of zero and a known variance v_j. The variance of e_j is specific to each study *j* and calculated using the sampling variance formulas provided in the previous section (e.g., the sampling variance for v_j for d_j is provided in Equation 2). [b] *X* reflects the moderators (study characteristics) and *p* reflects the number of moderators. [c] In a random effects model, sources of variation include both the within study estimation error and between studies variance, as captured by v_j and τ, respectively (Borenstein, Hedges, Higgins, & Rothstein, 2009).

Table 3

Results from Fixed and Random Effects Models Based on Lazowski and Hulleman (2015) Data – No Moderators

				Fixed Effects (FE) Models					
	FE: Traditional						FE: Multilevel		
	Est.	*SE*	*p*				Est.	*SE*	*p*
γ_0	0.4032	0.0168	<0.001				0.4032	0.0168	<0.001
				Random Effects (RE) Models					
	RE: Traditional-MM			RE: Traditional-ML			RE: Multilevel-ML		
	Est.	*SE*	*p*	Est.	*SE*	*p*	Est.	*SE*	*p*
γ_0	0.4970	0.0352	<0.001	0.4967	0.0351	<0.001	0.4967	0.0351	<0.001
τ	0.0627	---		0.0619	0.0152		0.0619	0.0144	

Table 4

Results from Fixed and Random Effects Models Based on Lazowski and Hulleman (2015) Data – With Moderators

Fixed Effects (FE) Models

		FE: Traditional			FE: Multilevel		
		Est.	*SE*	*p*	Est.	*SE*	*p*
γ_0		0.5938	0.0660	<0.001	0.5938	0.0660	<0.001
γ_1	ExD	-0.2489	0.0430	<0.001	-0.2489	0.0430	<0.001
γ_2	MS	0.1512	0.0658	0.0215	0.1513	0.0658	0.0242
γ_3	HS	-0.0684	0.0643	0.2874	-0.0684	0.0643	0.2907
γ_4	PS	-0.0237	0.0600	0.6933	-0.0237	0.0600	0.6944

Random Effect (RE) Models

		RE: Traditional-MM			RE: Traditional-ML			RE: Multilevel-ML		
		Est.	*SE*	*p*	Est.	*SE*	*p*	Est.	*SE*	*p*
γ_0		0.5923	0.1272	<0.001	0.5929	0.1239	<0.001	0.5929	0.1239	<0.001
γ_1	ExD	-0.1942	0.0800	0.0153	-0.1967	0.0778	0.0114	-0.1967	0.0778	0.0135
γ_2	MS	0.1652	0.1307	0.2063	0.1649	0.1272	0.1949	0.1649	0.1272	0.1987
γ_3	HS	-0.0125	0.1382	0.9281	-0.0134	0.1343	0.9206	-0.0134	0.1343	0.9209
γ_4	PS	0.0085	0.1233	0.9448	0.0081	0.1200	0.9461	0.0081	0.1200	0.9463
τ		0.0530	---		0.0483	0.0128		0.0483	0.0124	

Note. ExD refers to experimental design, MS refers to Middle School, HS refers to High School, and PS refers to Post-Secondary.

Table 5

Significance Tests Used to Estimate Between-Study Variance (τ)

Test Name	Estimator of τ	τ (from Table 3)	Statistic	*df*	*p*	Notes
Fixed Effects *Q*-statistic	NA	NA	$\chi^2 = 297.24$	82	<0.0001	Provided in output of MeanES macro. Calculated using Equation 10 where weights and overall effect size are those from the fixed effects model with no moderators.
Random Effects *Q*-statistic	MM	0.0627	$\chi^2 = 328.39$	82	<0.0001	Calculated using Equation 10 where weights and overall effect size are those from the random effects model with no moderators and MM estimation of τ.
Multilevel *Q*-statistic	ML	0.0619	$\chi^2 = 328.20$	82	<0.0001	Provided in HLM software output. Calculated using Equation 10 where weights and overall effect size are those from the random effects model with no moderators and ML estimation of τ.
Wald Test	ML	0.0619	$z = 4.31$	NA	<0.0001	Provided in SAS PROC MIXED output (covtest option).
LRT	ML	0.0619	$\chi^2 = 131.30$	1	<0.0001	Calculated as difference between deviances of fixed and random effects models with no moderators.

Note. NA = not applicable; MM = noniterative method of moments estimator of DerSimonian and Laird (1986); ML = maximum likelihood; LRT = likelihood ratio test.

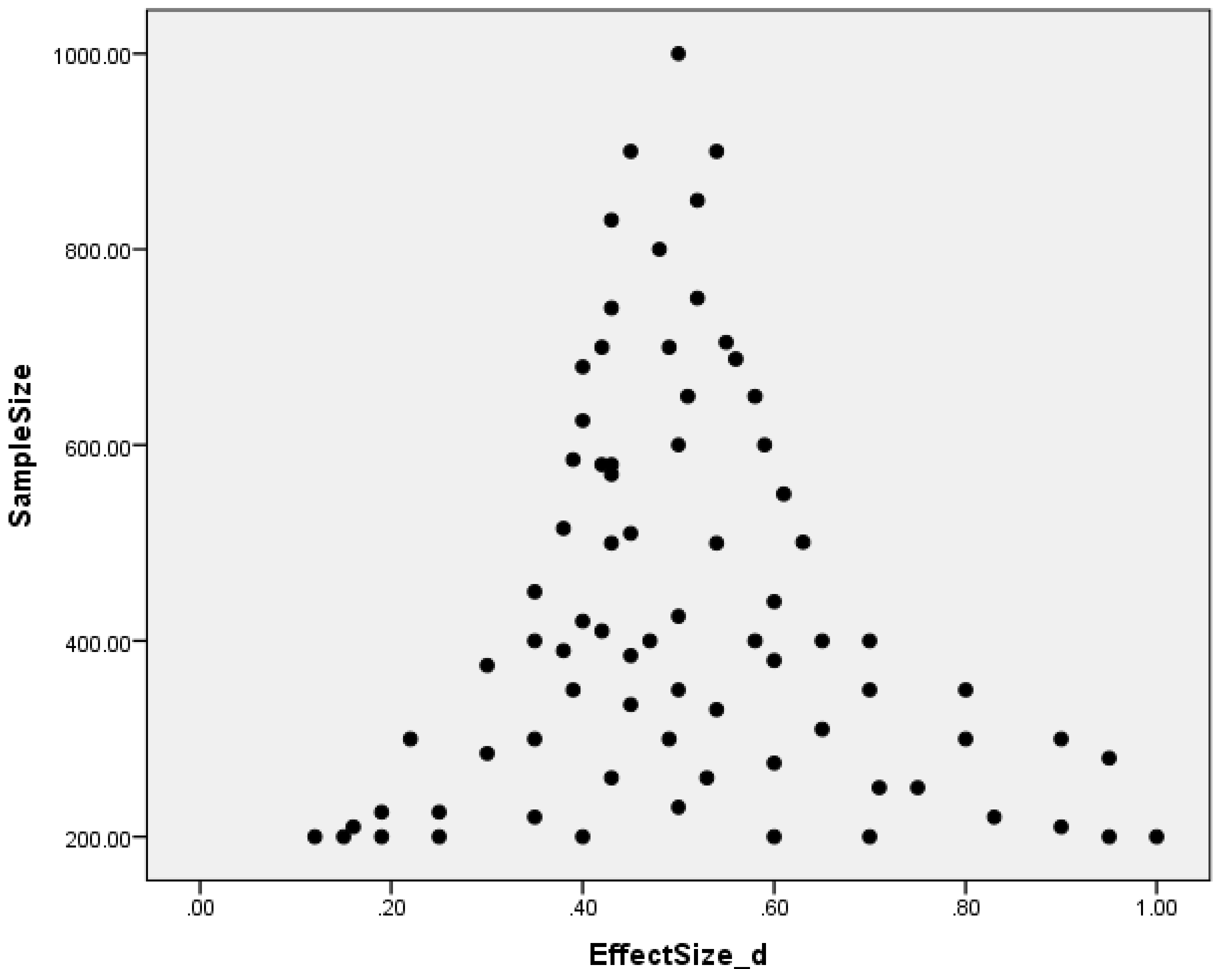

Figure 1. Example funnel plot of effect sizes (x-axis) by sample sizes (y-axis).

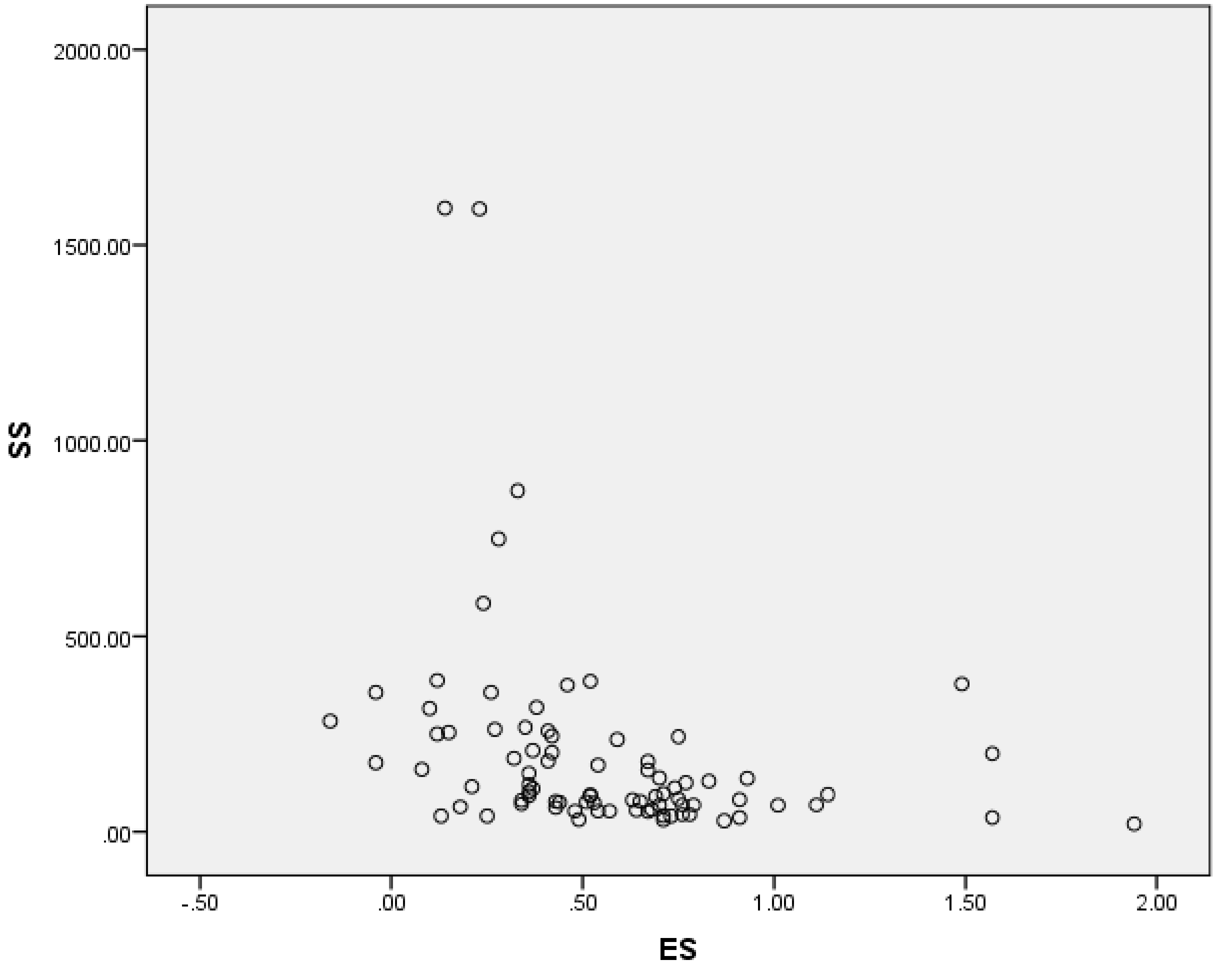

Figure 2. Funnel plot of effect sizes (x-axis) by sample sizes (y-axis). One study with a sample of 21559 was excluded to facilitate the interpretability of the axes. The four studies on the far right of the plot were identified as outliers. These were Windorized to a value 3 standard deviations from the mean of all effect sizes for moderator analyses.

Appendix

Research designs intended to analyze the relationship between two dichotomous variables can also be used in meta-analysis. For instance, studies may compare two groups with respect to the relative odds of some status or event (e.g., successful outcome, diagnosis of illness, dropping out of school) and the data are presented in terms of relative frequencies and proportions, plotted in cross-tabulation tables (Lipsey & Wilson, 2001). In these types of designs, the odds ratio is most commonly used as the measure of effect size. The odds ratio can be used with data collected from cross-sectional, prospective, or retrospective study designs (Fleiss & Berlin, 2009). The viability of the odds ratio across these designs represents an advantage over other potential effect size measures that are more limited in their use. For instance, the phi coefficient is only appropriate for cross-sectional designs; the sample difference and the rate ratio are only appropriate for cross-sectional or prospective designs (2009).

The odds ratio is calculated via cell frequencies or proportions in a 2 X 2 cross-tabulation table using:

$$OR = \frac{ad}{bc} = \frac{p_a p_d}{p_b p_c} = \frac{p_a / p_b}{p_c / p_d} = \frac{p_a(1-p_c)}{p_c(1-p_a)} \tag{13}$$

where *a, b, c*, and *d* represent cell frequencies and p_a, p_b, p_c, and p_d represent the proportion of each group in each status (Lipsey & Wilson, 2001). Because the form of the odds ratio is centered around 1 (indicating no relationship) rather than 0, with values spanning 0 to 1 reflecting a negative relationship, and values greater than 1 reflecting a positive relationship, the analyses are usually performed using the natural log of the odds ratio to ease interpretation (2001). When transformed into the logged odds ratio, the sampling distribution is approximately normal with a mean of 0; as well, a positive value reflects a positive relationship and a negative value reflects a negative relationship (2001). The logged odds ratio, variance, and inverse variance weight are given below:

$$LOR = \log_e(OR), \tag{14}$$

$$v_{LOR} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \tag{15}$$

$$w_{LOR} = \frac{1}{v_{LOR}} = \frac{abcd}{ab(c+d) + cd(a+b)} \tag{16}$$

In order to generate summary statistics such as means and confidence intervals, the logged odds ratio is converted back to an odds ratio using the following calculation:

$$OR = e^{LOR} \tag{17}$$

where *e* is the base of the natural logarithm (2001).

Finally, the product-moment correlation (*r*) is most appropriate when the research design examines the relationship between two continuous variables. Because *r* is already standardized, there is no conversion needed to compare the strength of the relationship between the variables even when they are operationalized in a different manner between or among studies. Put differently, *r* can be conceived as an effect size in and of itself. The product-moment correlation between two variables, *x* and *y*, is calculated by the following formula:

$$r_{xy} = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y} \tag{18}$$

with σ_{xy}^2 representing the covariance between *x* and *y*, and σ_x and σ_y representing the standard deviations of *x* and *y*, respectively (2001). However, when the correlation does not equal zero, the sampling distribution becomes skewed and is not normal. This leads to problems with estimating the standard error and associated confidence intervals (Alexander et al., 1989; Rosenthal, 1994). To address this problem, the correlations are transformed into a *Z*-statistic using the Fisher's Z_r transformation (Hedges & Olkin, 1985):

$$Z_r = .5\log_e\left[\frac{1+r}{1-r}\right] \tag{19}$$

where *r* represents the correlation between the two variables and $\log_e$ represents the natural logarithm. The standard error and inverse variance weight are then computed using the Z_r-transformed correlation with the formulas below:

$$v_{Z_r} = \frac{1}{N-3} \tag{20}$$

$$w_{Z_r} = \frac{1}{v_{Z_r}} = N-3 \tag{21}$$

Narrative Review of Motivation Interventions in Education – Paper 2

Introduction

The Importance of Intervention Studies in Educational Research

Although observational and correlational research can generate and test hypotheses and investigate how constructs operate in various settings, intervention research in education (i.e., empirical investigations that manipulate an independent variable) provides valuable information about what happens when we attempt to enhance educational outcomes through intentional manipulation. From a theoretical perspective, intervention studies help move the field forward by providing insight about the causal relationships between constructs and educational outcomes, or between educational settings and outcomes (Shadish, Cook, & Campbell, 2002; Tunnell, 1977). Because interventions represent an operationalized theory in action, they provide a strong test of the theory as applied in an educational context. From a theoretical perspective, intervention studies can serve as a source of validity evidence linking not only the measurement of constructs to hypothesized outcomes as a result of experimental manipulation (Messick, 1990), but also the degree to which changes in a theoretical construct predict outcomes in hypothesized ways.

Education researchers and practitioners are ultimately interested in how to structure the educational context in order to maximize student learning outcomes. In other words, they aim to develop an intervention, or interventions, that facilitate student learning and academic achievement. This requires testing the extent to which the interventions, based on our theoretical hypotheses, create the kind of change in students and teachers that we had envisioned. If not, we can go back to the drawing board to revise the intervention, our theories, or both. Without this kind of idea testing, our theories will not be pushed to grow, and knowledge about how to best structure educational environments will be limited. Incorporating intervention studies into how we think about our theories is aligned with movements in other fields to more quickly translate research findings into practice, such as improvement science efforts in health care (e.g., Berwick, 2008; Marshall, Pronovost, & Dixon-Woods, 2013) and education (Bryk, Gomez, & Grunow, 2010). When conducted as a part of a complete methodological approach that includes observations, mixed-methods, development/design-based studies, and randomized control trials (see Brown, 1992; Design-based Research Collective, 2003; Harackiewicz & Barron, 2003; Harackiewicz & Hulleman, 2010; Hulleman & Barron, under review), intervention studies offer the opportunity to make great advances in our theoretical and practical knowledge about education.

From a practical perspective, intervention studies facilitate our understanding about which interventions are most effective in improving educational outcomes in a way that observational research cannot. This understanding can guide policy recommendations for educational practice built on appropriate, evidenced-based research. As Raudenbush (2005) notes, "Among policymakers, public and private research funding agencies, and applied education researchers themselves, there is currently an overarching interest in identifying interventions that show strong promise, based on convincing evidence, to improve teaching and learning in U.S. classrooms" (p. 25). Policymakers do not provide direct intervention in the classroom in the same way a teacher or researcher might; rather, their influence is more indirect (2005). In a more indirect manner, intervention research can provide policy makers with critical information to make decisions that impact resource allocation, accountability and instructional practices, and the transformation of school governance (e.g., school choice, charter schools) (2005). No matter a policymaker, researcher, or practitioner, there appears to be growing emphasis on the use of intervention research and experiments to inform decision-making and practice (Raudenbush, 2008).

Motivation Theory and Interventions

Despite the benefits of intervention research described above, intervention research in the field of education has been on the decline over the past two decades (Hsieh et al., 2005; Robinson et al., 2007). Motivation research is no exception. Despite the considerable volume of theoretical, qualitative, observational, and correlational studies, there have been fewer experimental tests of motivation theory in the field of education (Wentzel & Wigfield, 2007). This trend has persisted despite calls for increasing intervention and use-inspired research (e.g., Blackwell, Trzesniewski, & Dweck, 2007; Hidi & Harackiewicz, 2000; Maehr & Meyer, 1997; Midgley & Edelin, 1998; Pintrich, 2003; Wentzel & Wigfield, 2007). In addition, it was unclear to what extent different theories had been experimentally tested, what interventions were effective and in what context (e.g., were interventions more effective for one age group over another).

Thus, in order to systematically evaluate what has been done to date, Lazowski and Hulleman (2015) conducted a meta-analysis examining motivation interventions that were conducted in authentic educational field settings (e.g., classrooms, workshops). Prior to this work, a meta-analysis examining motivation interventions conducted in field settings had not yet been conducted. We conducted a comprehensive search of the literature and identified theoretically grounded motivation interventions that had been experimentally tested in educational contexts and examined the extent to which the interventions impacted student outcomes. In summary, the meta-analysis included 66 published and unpublished papers of 84 field studies grounded in motivation theory, accounting for 37,239 participants. The motivation interventions in this meta-analytic review were promising, averaging approximately a half a standard deviation effect size ($d = 0.49$; 95% $CI = [0.42, 0.56]$). Importantly, this average effect size did not significantly vary according to any study characteristics we coded, with the exception of experimental design (randomized experiments demonstrated smaller effect sizes than quasi-experiments). More specifically, there were no statistically significant differences in effect size due to theoretical framework of the intervention, age of participants (elementary through post-secondary students), type of dependent variable (academic performance, behavior, self-reported motivation), or degree of naturalness [whether the intervention was part of the regular academic experience (natural treatment), occurred in a setting outside the laboratory (natural setting), or included a dependent measure that normally occurs within the educational context (e.g., exams, choices about activities; natural behavior)]. Table 1 presents the motivation intervention studies included in this meta-analysis, the corresponding effect sizes, and the sample characteristics described above.

This meta-analysis was the first step in systematically evaluating the effectiveness of these interventions to provide a “state of affairs” for the motivation field and the next steps in developing, testing, and implementing effective interventions in educational contexts. Based on the results of this meta-analysis, we offered comprehensive implications related to theory development as well as implications related to policy, practice, and research. However, although formal meta-analytic techniques can provide a quantitative analysis that can be useful in summarizing the interventions, one limitation is that there is often not enough space to also provide a comprehensive narrative review of the studies included. Thus, a narrative review can offer qualitative insight that can complement the quantitative analyses found via meta-analysis. Toward this end, we offer a more thorough narrative review of the studies included in our meta-analysis.

Jingle Jangle in Motivation Research

There has emerged a growing body of research over the past 50 years demonstrating the impact of motivation on various educational outcomes including, but not limited to, academic achievement, effort and persistence, development of interest, and task engagement (see Elliot & Dweck, 2005 for a comprehensive review of theories and research). As a result, several theoretical perspectives of motivation have been proposed. Although these theories have helped develop substantial knowledge of the factors facilitating or thwarting motivation, the proliferation of theories and constructs has also contributed to some uncertainty about what factors are most salient to student motivation (Glynn, Aultman, & Owens, 2005; Murphy & Alexander, 2000). Subsequently, it can be difficult to interpret motivation theory and research for practical use in the classroom (Schunk, 2000). In addition, there exists some overlap in the constructs and terminology used for seemingly distinct motivation theories. These theories and constructs therein, therefore, are susceptible to “jingle” and “jangle” fallacies (Block, 1995; Marsh, 1994), especially to those without expertise in the field. Specifically, it may be difficult for someone without intimate knowledge of these theories to determine whether two constructs with the same label are in reality quite different (jingle fallacy); or, it may be difficult to determine whether two constructs with different labels are in reality the same thing (jangle fallacy). The jingle and jangle fallacies are likely to thwart, rather than advance, the goal of synthesizing theory and research with practical application (e.g., Hulleman, Schrager, Bodmann, & Harackiewicz, 2010).

For instance, suppose a fifth grade teacher consults with a motivation researcher at the local university asking for assistance in her classroom. The teacher indicates that she has a "motivation" problem in her class and asks the researcher, "What should I do in my classroom based on your research?" This seemingly innocuous and simple question may be particularly difficult for the researcher to answer. First, as noted earlier, there have been few experimental tests of motivation theory in the field of education and thus the researcher may not be able to provide a pointed or exact answer based on previous research. Second, the motivation researcher would have to ask several follow-up questions to determine exactly how the teacher was operationalizing the term "motivation". Does she mean building their self-efficacy? Enhancing the value they attach to the subject matter? Increasing their intrinsic motivation? Perhaps she means persisting with effort? Among these questions also lies the problem of the jingle and jangle fallacies. For instance, suppose the teacher indicates that she would like to build their self-efficacy. Does she mean self-efficacy according to Bandura's (1997) conceptualization as a belief in one's ability to plan and execute the skills necessary to produce certain behaviors? Or does she actually mean perceived competence according to Ryan and Deci's Self-Determination Theory (2002), defined as "feeling effective in one's ongoing interactions with the social environment and experiencing opportunities to exercise and express one's own capacities" (p. 7). Or does she mean both?

In an effort to bring some cohesion to the various similar (or dissimilar) constructs among the theories, I propose here the use of the expectancy-value-cost framework (Barron & Hulleman, in press; Eccles et al., 1983). Eccles and colleagues (1983) shy away from using the term theory. Instead, they refer to their work as an expectancy-value framework or model and adopt an integrative perspective of various constructs from different motivational theories to better understand students' academic performance, persistence, and choice behaviors. Their framework was also meant to be developmental and contained numerous antecedents of expectancies and values. The entire model is complex and integrates the cultural milieu, unique past events, students' perceptions of past events, socializers' behaviors and attitudes, students' perceptions of socializers' attitudes and expectations, and students' goals and self-concept.

However, within this complexity lies an organization that includes many of the constructs contained in separate theories but at the same time is parsimonious and practitioner-friendly. The former indicates that the expectancy-value-cost framework first serves as a conceptual umbrella under which other motivation theories and constructs can easily fit. The latter suggests it offers a practical advantage to linking theoretical constructs to real-world applications by narrowing the focus into three digestible, overarching concepts: a) an individual's anticipated ability to successfully accomplish the task (i.e., Expectancy), b) an individual's perceived importance for the task (i.e., Value), and c) how much an individual perceives that he or she has to sacrifice or give up to accomplish the task (i.e., Cost). Therefore, when the researcher in the above scenario is posed the "motivation" question by the teacher, the answer may lie in one of three questions: a) *Do you want your students to think they can accomplish their tasks?*, b) *Do you want your students to see the value in the subject matter?*, or c) *Do you want to reduce barriers that may be preventing them from investing time, energy, or resources into the class?* Using the expectancy-value-cost framework, these types of questions will first integrate the various constructs from different theories in a more parsimonious manner. In the previous example, Bandura's concept of self-efficacy and Deci and Ryan's concept of perceived competence according to Self-Determination theory would both correspond to the first question (expectancy), reducing the jingle and jangle fallacy. Second, to a practitioner who might not be an expert in motivation theory, the questions offer a more digestible and accessible explanation for motivation – that is, motivation being a function of an individual's expectancies for success, values for the task, and cost associated with attaining that task – without the undue burden of having the requisite knowledge of the various motivation theories and sundry constructs therein.

In a similar fashion and to achieve these same goals, I will organize the following narrative review of motivation theories and corresponding intervention studies depending on whether they were intended to increase student expectancies, increase values, or reduce cost (cf. Hulleman, Barron, Kosovich, & Lazowski, in press). The theories represented in the Lazowski and Hulleman (2015) meta-analysis are presented and described in Table 2. It is important to note that this organization is intended as a parsimonious means to categorize the interventions for people not particularly knowledgeable about the sometimes subtle nuances and distinctions among the various constructs and theories. While I certainly honor the theoretical space each theory commands, the ultimate aim is to provide a useful organization for practitioners not familiar with motivation research.

Prior to presenting the studies related to expectancies, values, and cost, each section will begin with a description of research-based sources, or drivers, of the interventions. Doing so serves to organize and identify pathways for practitioners to enhance student motivation using research-based sources of expectancy, value, and cost that are potentially amenable to (and have been tested through) interventions. I then provide a summary of the studies for each theory that is linked to expectancy, value, or cost, along with a detailed explanation of one exemplar intervention study from each theory. Tables 3, 4, and 5 provide these research-based sources of expectancy, value, and cost, respectively, along with the definition of each source. Next, Table 6 presents expectancy, value, and cost interventions included in the Lazowski and Hulleman (2015) meta-analysis broken down by theory and research-based sources.

Research-based Sources of Expectancy-Related Beliefs

Research indicates that there are various sources or pathways of expectancy in the literature, either from theory or research, which can be targeted by interventions. These sources or pathways refer to the underlying psychological processes that both serve as antecedents of expectancy-related constructs and that are potentially amenable to intervention by educational practitioners, including teachers, parents, and administrators (Hulleman et al., in press). Importantly, these sources can serve as the targets or drivers of interventions aimed at enhancing student outcomes by boosting students' expectancies. Although there are additional sources of expectancies – such as those identified in the

Eccles model, including cultural milieu and socializers' goals and expectations – many of the sources described below have been identified as being the most amenable to change through direct intervention. That is, many have been tested via experimental methods.

One source that can target and potentially alter expectancies is the manipulation of students' *perceptions of their ability and skill* for specific tasks. Research has demonstrated that when students perceive they have a high level of ability and/or skill on a task, they are more likely to experience high expectancy for success on that task (Bandura, 1997; Wigfield & Eccles, 2002). In addition, when students have high perceptions of their ability and skills for a particular task, they are more likely to engage in that task (Eccles et al., 1983). Perceptions of ability and skill are also closely related to another source of expectancy, *success experiences*. When students are successful at an activity, or watch others have success, they are more likely to experience high levels of expectancy (Bandura, 1997; Eccles et al., 1983). Previous performance of a skill is considered a strong source of self-efficacy and expectancy-related beliefs, representing tangible, authentic evidence that an individual can or cannot perform the requisite skill. Not only are students' own perceptions important for expectancies, but *perceptions of others' expectations* are important as well. Research has demonstrated that parents' and teachers' expectancies and attitudes shape expectancies. For example, if teachers have high expectations for their students, these students in turn develop high expectancies for themselves (Bandura, 1997; Dweck & Leggett, 1988; Dweck 1999; Eccles et al., 1983).

Largely based in research regarding self-efficacy theory, two other sources that serve to increase student expectancies are *support and scaffolding* (Bandura, 1997) and *clear expectations* (Pajares, 1996). Support and scaffolding refers to an appropriate amount of support in completing an activity (e.g., through encouragement or having the requisite resources to successfully complete a task). For example, expectancies can increase when an individual receives positive verbal encouragement from a knowledgeable and reliable source (such as a teacher) (Bandura, 1997). Furthermore, studies also indicate that having clear expectations can promote students' expectancies. That is, if students know what is expected of them on an activity and have clearly defined goals, then they are more likely to have high levels of expectancy that they can successfully perform that activity (Pajares, 1996).

Other sources that promote expectancies are related to the difficulty of the task and means to overcome those difficulties. For instance, one means to promote higher levels of expectancy is through changing students' perceptions of the difficulty of a task, activity, or subject, formally termed *perceived task difficulty.* When students perceive a task as not being difficult, they develop higher estimates of their own abilities for the subject or task, which in turn increase expectancies and their motivation to engage in the behaviors necessary to complete the task (Bandura, 1997; Pajares, 1996; Wigfield & Eccles, 2002). Furthermore, when the difficulty of the task or activity matches students' current skill-set, they are more likely to experience high expectancy for success (Eccles et al., 1983). Thus *appropriate challenge* is another source that can be manipulated to promote student expectancies. Additionally, expectancies increase when appropriate challenge is matched with *growth experiences*. If students are provided with learning experiences that challenge them to develop and learn, and subsequently experience growth in their skills and improvement in performance, they are more likely to experience higher expectancies of success in the future (Dweck & Leggett, 1988; Dweck, 1999; Hong et al., 1999).

Intervention research also supports the notion that the type of *feedback* students receive has an impact on their expectancy to successfully complete a task in the future. Implicit theories of intelligence (see Dweck & Leggett, 1988; Dweck 1999) posit that individuals generally possess one of two different theories regarding their intelligence: 1) that intelligence is dynamic, malleable, and amenable to change given sufficient effort and hard work (*incremental view of intelligence*, *or growth mindset*), or 2) that intelligence is fixed, static, and resistant to change regardless of effort and hard work (*entity view of intelligence, or fixed mindset*). Primarily guided by this work, when students receive *feedback* that effort matters, that skills and abilities are amenable to change, and are task-focused (growth mindset, rather than fixed mindset), they are more likely to experience high expectancies for success and that difficult tasks can be overcome and accomplished (Dweck & Leggett, 1988; Dweck, 1999). This *feedback* can also impact students' *effort attributions* to failure or success. When students believe that their effort will lead to learning, they are more likely to attribute success or failure to the effort they have expended. When these attributions are related to something within their control (effort) rather than something they cannot control (fixed ability), they are more likely to experience higher levels of expectancy on tasks (Dweck & Leggett, 1988; Dweck 1999; Weiner, 1974).

Interventions Primarily Designed to Promote Student Expectancies

Attribution theory. Another set of interventions aimed at promoting student expectancies has focused primarily on changing cognitive attributions for success and failure (*effort attributions*). Many of these interventions are designed to provide students with training about attributing academic success to things that are within their control (e.g., effort), and that academic difficulties are not uncommon and can be overcome. These interventions have been successful in increasing perceived academic control, and these changes mediate effects on academic motivation and achievement outcomes (e.g., Hall et al., 2004; Ruthig et al., 2004).

There have been several studies grounded in attribution theory demonstrating that changes in causal attributions relate to changes in academic achievement. Many of these intervention studies sought to change the attributions that low performing or at-risk students made regarding their academic achievement. The interventions attempted to shift the cause from low ability to underscoring the importance of effort and the notion that achievement was amenable to change. These shifts in attribution have been demonstrated to improve course grades (Boese et al., 2013; Hall et al., 2007; Hall et al., 2004; Yeager et al., 2013), exam performance (Struthers & Perry, 1996), GPA (Boese et al., 2013; Ruthig et al., 2004; Yeager et al., 2013; Wilson & Linville, 1982; Wilson & Linville, 1985), standardized test scores (Good et al., 2003; Wilson & Linville, 1982; Wilson & Linville, 1985), intrinsic motivation (Hall et al., 2007), and reduce text anxiety and voluntary course withdrawal (Ruthig et al., 2004).

In their seminal study and replication study, Wilson and Linville (1982, 1985) tested effects of an attribution intervention on academic performance. In the original study (1982), students were randomly assigned to either an experimental or control condition. Students in the experimental condition watched videotapes and were shown statistics indicating how students typically struggled academically during their freshman year but improved afterwards. Students in the control condition did not receive this information. As a manipulation check, half of the students in both conditions subsequently wrote lists explaining why there was an increase in grades after freshman year and which of those explanations were relevant to their experience. The other half wrote lists explaining why they thought the divorce rate in some states was decreasing.

The replication studies had slightly different student samples and selection criteria for academically at-risk students, but the intervention was the same. In the first replication study (Wilson & Linville, 1985, Study 1), two separate experimental conditions were included. One contained information indicating that grades generally improve following freshman year, and in another, this information was not provided to the students. However, given very similar responses on dependent measures and manipulation checks between these separate conditions, the two experimental groups were aggregated to form one experimental condition. In the second replication study (Wilson & Linville, 1985, Study 2), the same experimental condition was used, but students in the control condition watched videotapes and were shown “filler” statistics without any information about grades. As in the original study, students in the replication studies were randomly assigned to either an experimental or control condition.

Effects of these interventions were measured through both short-term and long-term academic performance, and retention in school. For short-term academic performance, defined by reading comprehension items from the GRE taken immediately after the intervention, males in the experimental condition performed better than those in the control conditions across all three studies (average $d = 0.45$). There were no significant differences for females. For long-term academic performance, defined as the comparison of grades from the semester prior to the study to the semester after the study, students in the experimental condition also showed gains across all three studies (average $d = 0.27$). The effect was stronger for males ($d = 0.47$) than females ($d = 0.21$). Retention in school one year following the study was also impacted by the intervention, with those in the experimental condition less likely to drop-out (2%) than those in the control condition (10%).

Implicit theories of intelligence. Based on Dweck's (1986, 1999) theory of the malleability of intelligence, implicit theories of intelligence interventions target students' perceptions about their capacity to learn. Specifically, these interventions attempt to change students' beliefs about whether intelligence is fixed (i.e., entity mindset) or is malleable (i.e., incremental mindset) through *feedback* and *effort attributions*. There have been several interventions guided by this theory that have been demonstrated to be effective in enhancing various student outcomes. These studies have been effective in changing students' beliefs about their intelligence (e.g. Aronson et al., 2002; Blackwell et al., 2007; Paunesku et al., 2014; Yeager et al., 2013); increasing enjoyment/interest for and importance of academics (e.g., Aronson et al., 2002; Hong & Lin Siegler, 2011); reducing stereotype threat (e.g., Aronson et al., 2002), stress, anxiety, and negative self-feelings (e.g., Yeager et al., 2014); improving grades and academic performance (e.g., Aronson et al., 2002; Bornine, 1998; Blackwell et al., 2007; Good et al., 2003; Hong & Lin Siegler, 2011; Paunesku et al., 2014; Yeager et al., 2014); and improving classroom motivation as measured by teachers (Blackwell et al., 2007).

As an example, Blackwell, Trzesniewski, and Dweck (2007) tested the effects of an in-depth intervention designed to teach seventh grade students about various facets of implicit theories of intelligence and how the brain can become stronger through effort. Students randomly assigned to the experimental condition participated in eight weekly, twenty-five minute lessons that covered topics such as the structure and function of the brain, incremental theory of intelligence, and discussions about the malleability of the brain and how learning makes students smarter. A control condition participated in the same or similar lessons, excluding those that explicitly covered incremental theory of intelligence, for which alternative lessons were created.

Based on teacher reports, with teachers being blind to condition, students in the experimental condition were reported to have larger gains in motivation by comparison to the students in the control group ($OR = 3.26$). Based on student reports, those in the experimental condition scored significantly higher on items that tested the incremental theory intervention content than did those in the control condition ($d = 0.95$). In addition, at post-test, students in the experimental condition more strongly endorsed an incremental theory of intelligence compared to their own pre-test scores ($d = 0.66$), and compared to students in the control condition at post-test ($d = 0.47$). Trajectories of math grades were compared at three time points, including grades from the previous year (spring term of sixth grade; Time 1), pre-intervention (fall term of seventh grade; Time 2), and post-intervention (spring term of seventh grade; Time 3). Although grades for both conditions declined from Time 1 to Time 2 prior to the intervention (b = -.34), students in the experimental condition improved their math grades between Time 2 and 3 following the intervention (b = .53), whereas students in the control continued in their declining trajectory.

Multiple perspectives. Some studies attempted to integrate various concepts and/or constructs from multiple theories in the design and delivery of the intervention. For those designed to promote expectancies, for example, these interventions may have incorporated different facets of several theories (e.g., Craven et al., 1991; Duckworth et al., in press; Kitsantas et al., 2010; Paunesku et al., 2014). Morisano et al. (2010) provides an example of one such approach, integrating theories of goal setting and possible selves. In this study, the authors tested the effects of an online goal-setting intervention on academic achievement for struggling college students. Participants were chosen if they had a GPA lower than 3.0, were enrolled full-time (at least nine credits), and indicated that they were struggling academically. From this sample, students were randomly assigned to either an experimental or control condition.

Students in the experimental condition completed a comprehensive online program that was grounded in goal-setting theory and possible selves. This intervention included 8 separate steps that required students to think and write about the following: 1) Write about possible, desirable selves and futures; 2) Identify several goals related to these selves and futures; 3) Rank order these goals based on importance and possible attainment, and; 4) Examine the impact on themselves and others should the goal(s) be achieved. Steps 5-7 required students to elaborate in detail their goals and implementation plans. Finally, Step 8 required students to indicate how committed they were to achieve the goal(s). Students in the control group completed online tasks and wrote about topics such as positive psychology, positive experiences in their past, and completed a career-interest inventory. Several dependent measures were used in the analyses including GPA, retention rates, affect, and content of goal (e.g., elaboration/word count for goals).

Although no differences were found at pretest for GPA between the experimental and control groups, GPA at posttest was significantly higher for the experimental group compared to the control group ($d = 0.50$). Furthermore, changes in GPA from pretest to posttest were significant within the experimental group ($d = 0.65$) but no differences were found for the control group. A Fisher's exact test was used to examine retention rates, which was operationalized as maintaining a full course load during the semester following the intervention. Results indicated that there was a significant difference between the two conditions, with no students in the experimental condition enrolled in fewer than nine credits compared to 20% (8 students) in the control condition.

A questionnaire completed at the end of the study revealed that students in the experimental condition scored higher for a reduction in negative affect compared to the control condition ($d = 0.46$). The reduction in negative affect was also correlated with improvement in grades ($r = .19$). Finally, content analyses revealed that elaboration for possible futures (number of words) was correlated with improvement in grades as well ($r = .30$).

Theoretical perspectives with one study. In the Lazowski and Hulleman (2015) meta-analysis, some theories were only represented by a single study. For the Primarily Expectancy Intervention category, one such theory was Self-Confrontation. This theory suggests that motivation to change is elicited when students perceive that their behaviors and values differ from their self-conception (Rokeach, 1973). Greentein (1976) conducted a study targeting *feedback.* In this study, student teachers who received objective feedback concerning their own values and those of good and mediocre teachers exhibited significantly higher value ranks for mature love and loving and lower ranks for self-respect d = 0.61), and showed significantly higher scores on a behavioral measure of teaching ability than did student teachers not receiving such feedback (d = 0.58).

Research-based Sources of Value

Like expectancy-related beliefs, research supports that there are various sources or pathways that serve to promote values. Correspondingly, interventions (and the theories that guide these interventions) designed to promote student values have attempted to do so through these various sources. Based largely in expectancy-value theory (Eccles et al., 1983), self-determination theory (Deci & Ryan, 1985), and interest theory (Hidi & Renninger, 2006), one source to improve the value students attach to tasks is through *intrinsic benefits*. When students find the activities and academic content inherently enjoyable and interesting, they are more likely to experience high value for those activities and the academic content (Eccles et al., 1983; Renninger & Hidi, 2011). These intrinsic benefits yield higher *levels* of motivation (Eccles et al., 1983) and higher *quality* of motivation (Ryan & Deci, 2000). Intrinsic benefits are juxtaposed with another source of value interventions, *extrinsic benefits.* Extrinsic benefits refer to external rewards or incentives that are used to promote motivation. When students receive external rewards and incentives for learning (e.g., prizes, food), they are more likely to experience high value to complete an activity; however, this also leads to low value for producing quality work (Marinak & Gambrell, 2008) and is considered to be of lesser quality compared to intrinsic motivation (Ryan & Deci, 2000).

A host of research indicates that another way to increase value is through *variety and novelty*. Not surprisingly, when students engage in activities that are varied and novel, they are more likely to experience high levels of interest and value (Hidi & Renninger, 2006; Kang et al., 2009). Another, related source that promotes values in students is *enthusiastic models*. That is, when students interact with or observe teachers or other adults who are enthusiastic, interested in, and passionate about a subject area or activity, they are more likely to feel higher value for that subject area or activity themselves (Patrick, Hisley, & Kempler, 2000). Therefore, value can be bolstered not only by the activity, but also by the individuals explaining or teaching those activities.

How the student perceives the task as being useful and meaningful also has bearing on the value attached to the task. For instance, when students are able to connect what they are learning to their personal lives and/or the real world, they are more likely to experience high value for the material (Hulleman & Harackiewicz, 2009). This source is considered *relevance*. Related to relevance, providing *context and rationale* for learning the material can also contribute to higher levels of value. Students are more likely to have higher value for material that has meaning and purpose in their lives (Lepper & Henderlong, 2000). Clearly, getting students to discover on their own how the material is relevant, meaningful, and purposeful to their lives can have a profound impact on the value they attach to the material.

Another source that appears to have an impact on value is *self-affirmation.* Studies targeting self-affirmation primarily ask students to think and write about their most important values in an effort to affirm, reinforce, and strengthen core aspects of themselves (e.g., Cohen et al., 2006, 2009). Self-affirmation can serve as a buffer against potential threats (such as stereotype threat) and subsequently increase academic performance, especially with low-achieving students (e.g., 2006).

Primarily grounded in self-determination theory (Ryan & Deci, 2000), two other sources that drive intervention efforts to increase value are *autonomy* (choice and control) and *relatedness* (positive relationships and a sense of belongingness). Autonomy refers to a sense of control, self-direction, and choice over learning, rather than feeling controlled or forced to comply from an outside source (2000). Personal value for learning tends to be higher when students feel a sense of autonomy over their learning, rather than feeling a sense that it is controlled by others (Reeve, 2009). Finally, *relatedness* refers to a sense of meaningful and caring relationships with others (Ryan & Deci, 2000). In an academic setting, when students experience meaningful student-to-student or student-teacher relationships, they are more likely to experience higher levels of value (Furrer & Skinner, 2003; Walton & Cohen, 2007).

Interventions Designed to Promote Student Value

Expectancy-value framework. Much of the experimental work within the Expectancy-value framework has been aimed at promoting value, primarily *relevance* or *utility value*. Many of the interventions used a brief writing task where students were asked to write about how the course material was useful or relevant to them or someone they knew (e.g., Durik et al., 2014; Hulleman et al., 2010; Hulleman & Harackiewicz, 2009). These interventions have been found to positively impact a number of outcomes, including: course-related interest (e.g., Durik et al., 2014; Hulleman et al., 2010; Hulleman & Harackiewicz, 2009); course performance (e.g., Durik et al., 2014; Harackiewicz et al., 2012; Hulleman et al., 2010; Hulleman & Harackiewicz, 2009); future interest in course-related careers (e.g., Hulleman & Harackiewicz, 2009); future course enrollment (e.g., Harackiewicz et al., 2012); perceptions of utility value for the subject area (e.g., Durik et al., 2014; Harackiewicz, et al., 2012); and increased expectancies for success (e.g., Durik et al., 2014; Hulleman & Harackiewicz, 2009).

As an example, Hulleman and Harackiewicz (2009) demonstrated the impact of the relevance intervention on student expectancies, interest, and academic performance in high school science classes. Students randomly assigned to the experimental condition were asked to write how the information they were learning in their science classes could be personally relevant or connected to their lives. Students in the control condition wrote summaries of the information they were learning in their classes. The effect of the relevance intervention was most profound for students who initially had lower expectancies for success in the class prior to the intervention. There were significant negative interactions between the relevance intervention and success expectancies on interest in science ($\beta = -.11$) and second-quarter grades ($\beta = -.18$), indicating that students with low-success expectancies reported more interest in science and received higher grades at the end of the semester in comparison to students in the control condition. In addition, the relevance intervention indirectly increased continuing interest in science, as student interest in science at the end of the semester was a significant predictor of interest in enrolling in subsequent science-related courses and pursuing science-related careers (β = .58).

Self-determination theory. Compared to other theoretical frameworks included in this study, self-determination theory has produced a large number of experimental field studies. These studies have largely focused on manipulating *autonomy* (choice and control) and *intrinsic benefits* as the primary drivers of the interventions. In some studies, students were the direct recipients of the intervention (e.g., Patall et al., 2010; Vansteenkiste, 2008), whereas in other studies, other sources (such as teachers or parents) were provided the intervention, and the impact of this intervention was subsequently measured through student outcomes (e.g., Guay et al., 2014; Reeve et al., 2004). These interventions have had a substantial impact on a variety of outcomes, including: intrinsic motivation (e.g., Froiland, 2011; Guay et al., 2014; Patall et al., 2010); autonomy (e.g., Froiland, 2011; Patall et al., 2010; Radill, 2012; Vansteenkiste et al., 2005; Vansteenkiste et al., 2004) and autonomy supportive behaviors (e.g., Reeve, 2004); competence (e.g., Guay et al., 2014; Patall et al., 2010); relatedness to the teacher (e.g., Guay et al., 2014); academic achievement (e.g. Guay et al., 2014; Patall et al., 2010; Schanffer & Schiefele, 2007; Vansteenkiste et al., 2008; Vansteenkiste et al., 2005; Vansteenkiste et al., 2004); homework completion rate (e.g., Patall et al., 2010); effort and persistence (e.g., Patall, et al., 2014; Vansteenkiste et al., 2008; Vansteenkiste et al., 2004); student engagement (e.g., Reeve, 2004; Vansteenkiste et al., 2004); interest (e.g., Schaffner & Schiefele, 2007); reduction in test anxiety (e.g., Schaffner & Schiefele, 2007) or stress (e.g., Vansteenskiste et al., 2004; and goal orientation/achievement goals (e.g., Vansteenkiste et al., 2008; Vansteenkiste et al., 2004).

In a series of three randomized field experiments, Vansteenkiste, Simons, Lens, Soenens, and Matos (2005) manipulated information that early adolescent students received when reading a text about nutrition to align with self-determination theory. Specifically, the goal content manipulation contrasted intrinsic (e.g., physical wellness) and extrinsic goals for the task (e.g., physical attraction); the autonomy-support manipulation varied whether the information was delivered in an autonomy-supportive (e.g., choice to follow the nutritional guidelines) or a controlling style (e.g., explicit expectations and/or pressure to follow the guidelines). In the first two studies, students regarded as obese according to the Body Mass Index (BMI) were selected to participate. Students' retention of the information was assessed in the short-term (i.e., immediately after the session) and in the long-term (i.e., four weeks later). Both rote and conceptual learning were assessed over the short- and long-term as well. A control condition was only used in one of the three studies (Study 2). In the third study, students not regarded as obese according to the BMI were selected to participate. The third study also included measures of task involvement (the degree that participants were absorbed in reading the text) and relative autonomy (the degree that participants felt the text was personally relevant, interesting, and enjoyable).

After controlling for level of obesity in Study 1, the autonomy-supportive condition scored higher in short-term (d = 1.33) and long-term (d = 0.90) conceptual learning relative to the external control condition; and higher in short-term (d = 1.48) and long-term (d = 1.21) conceptual learning relative to the internal control condition. The internal control condition scored higher in short-term (*d* = -0.43) and long-term (*d* = -0.52) rote learning compared to the autonomy-supportive condition; and higher in short-term (*d* = -0.42) and long-term (*d* = -0.36) rote learning relative to the external control condition. With regard to goal framing, students in the intrinsic goal framing conditions also scored higher in short-term (η^2 = .06) and long-term (η^2 = .03) conceptual learning relative to the extrinsic goal framing conditions.

After controlling for short-term conceptual learning, the intrinsic goal framing conditions scored higher in long-term conceptual learning compared to the extrinsic goal framing conditions (η^2 = .05); and the autonomy-supportive condition scored higher in long-term conceptual learning compared to the internal control condition (η^2 = .23). After controlling for short-term rote learning, the internal control condition scored higher in long-term rote learning compared to the autonomy-supportive condition (η^2 = .06).

After controlling for obesity in Study 2, the intrinsic goal framing conditions scored higher in short-term (η^2 = .30) and long-term (η^2 = .43) conceptual learning compared to the extrinsic goal framing conditions. By contrast, the extrinsic goal framing conditions scored higher in short term (η^2 = .10) and long term (η^2 = .26) compared to the intrinsic goal framing conditions. The autonomy-supportive conditions scored higher in short-term (η^2 = .16) and long-term (η^2 = .13) compared to the internal control conditions.

Contrast analyses revealed the autonomy-supportive/intrinsic goal condition scored higher in short-term (*d* = 0.72) and long-term (*d* = 0.80) conceptual learning relative to the control condition. Students in the internal control/extrinsic goal condition scored lower in short term (d = -0.80) and long-term (d = -0.94) conceptual learning relative to the control condition. Students in the autonomy-supportive/extrinsic goal condition scored lower in long-term conceptual learning compared to the control condition (d = 0.42). Also, students in the intrinsic goal/internal control condition scored lower in long-term rote learning relative to the control.

After controlling for short-term conceptual learning, the intrinsic goal framing condition scored higher on long-term conceptual learning (η^2 = .30) than the extrinsic goal framing condition; and the autonomy-supportive condition scored higher on long-term conceptual learning (η^2 = .08) than students in the internal control condition. After controlling for short-term rote learning, the internal control condition scored higher in long-term rote learning compared to the autonomy-supportive condition (η^2 = .20).

In Study 3, participants in the intrinsic goal framing condition scored higher in task involvement (η^2 = .15), relative autonomy (η^2 = .26), and conceptual learning (η^2 = .07) than the extrinsic goal framing condition. Participants in the autonomy-supportive condition scored higher in task involvement (η^2 = .51), relative autonomy (η^2 = .85), and conceptual learning (η^2 = .15).

Self-affirmation theory. Self-affirmation theory has largely produced interventions that help students maintain self-integrity by affirming important values via writing exercises. Many of these studies did so by asking students to list their most important value(s) and to write about why their value(s) were important to them (e.g., Cohen et al., 2006). Largely, these interventions have targeted low-achieving students or those from minority populations. By targeting the source of self-affirmation, these interventions have been successful in improving GPA and reducing achievement gaps (e.g., Cohen et al., 2006; 2009; Cook et al., 2012; Miyake et al., 2010; Sherman et al., 2013; Walton & Cohen, 2011) and course grades (e.g., Miyake et al., 2010; Sherman et al., 2013); standardized test scores (e.g., Miyake et al., 2010); self-perceptions (e.g., Cohen et al., 2009); social belonging in school (e.g., Cook et al., 2012; Walton & Cohen, 2011); retention/matriculation (e.g., Cohen et al., 2009); perceptions of daily adversity and identity threat (e.g., Sherman et al., 2013); and well-being (e.g., Walton & Cohen, 2011).

For example, Cohen, Garcia, Apfel, and Master (2006) tested such an intervention using a sample of African American and European American students in the 7th grade from middle to lower middle class families. In these randomized field experiments, students were presented with a list of values. Students in the experimental condition were instructed to choose their most important value (Study 1) or to choose two or three of their most important values (replication). In contrast, students in the control condition were instructed to choose their least important value (Study 1) or to choose two or three of their least important values (replication). Students in the experimental conditions in both studies wrote a passage about why their value(s) were personally important and students in the control condition wrote about why their chosen least important value might be important to another person.

Comparisons were drawn between the two conditions for grades at the end of the term. Class specific grades (for the class in which the intervention occurred) and mean class grades (all classes the students were enrolled) were used to test these effects. Further, to assess cognitive activation of stereotype threat related to race, students completed a validated word completion task. The stems of each word could potentially activate either a stereotype relevant or irrelevant response (_ACE). For this item, a stereotype relevant response would include "RACE" and a stereotype irrelevant response would include "FACE".

Results indicated that African American students in the experimental condition earned significantly higher class specific grades at the end of the term compared to African American students in the control condition (Study 1: $\beta = 0.26$; Replication: $\beta = 0.34$). No significant differences emerged in either study for European Americans. Effects of the intervention were most pronounced for African American students previously identified as low-achievers ($d = 0.89$) and moderate-achievers ($d = 0.88$). For mean class grades, African American students in the experimental condition earned higher overall grades at the end of the term compared to African American students in the control condition (Study 1: $\beta = 0.31$; Replication: $\beta = 0.21$). Once again, no differences emerged in either study for European Americans. Taken together, these results resulted in an approximate 40% reduction in the racial achievement gap for African American students in the experimental condition. The performance gap between African American students in the control condition and European Americans was 0.75, but reduced to 0.30 for African American students in the experimental condition.

After combining results from Study 1 and the Replication study, African American students in the experimental condition produced fewer stereotype relevant responses on the word completion task than African American students in the control condition ($d = -0.49$). No differences emerged between conditions for European Americans.

Interest theory. Interventions grounded in interest theory have primarily attempted to enhance the *variety and novelty* of the tasks. These studies aimed to excite and engage students in topics through expressive writing and by providing stimulating learning activities (e.g., Guthrie et al., 2006). Only a few interest theory-only (i.e., not combined with other theories) field experiments have been conducted; however, these have been successful in improving the following: reading comprehension (e.g., Guthrie et al., 2006); teacher ratings of student motivation (e.g., Guthrie et al., 2006); performance on writing tasks (e.g., Hidi et al., 2002); self-efficacy (e.g., Hidi et al., 2002); and interest (e.g., Hidi et al., 2002).

Guthrie and colleagues (2006) tested an intervention designed to stimulate situational interest and promote long-term interest and intrinsic motivation in the area of reading. Participants were 3rd grade students from four separate classes that varied in the amount of stimulating tasks delivered by their teachers. Although all four classes participated in the intervention, two of the classes were provided a high number of stimulating tasks and two were provided a low number of stimulating tasks.

The intervention was based on Concept-Oriented Reading Instruction (CORI; see Guthrie, Wigfield, & Perencevich, 2006). Fiction and non-fiction reading was aligned to science observations and experiments that were hands-on and interactive to induce excitement and interest for students. The two conditions (high versus low stimulating tasks) differed in the amount of observations, drawings, and opportunities for experiments which also included creating hypotheses and interpreting findings. Effects of the intervention on reading comprehension and motivation for reading were examined. Reading comprehension was measured using students' performance on two separate reading comprehension tasks, one related to the project and one standardized measure. Motivation for reading was measured using self-report and teacher rating scales related to intrinsic motivation and self-efficacy.

Multiple regression analyses were used to test the effect of the intervention. One analysis was conducted controlling for reading comprehension pre-test scores and portfolio scores (graded on rubrics for quality of drawings, questions, hypotheses, tables and graphs, and conclusions). Results indicated that after controlling for these variables, experimental condition was significant for reading comprehension post-test ($\beta = .27$). Means testing revealed that the high-stimulating tasks condition was significantly higher than the low-stimulating tasks condition ($d = 0.71$).

Another analysis was conducted controlling for self-reported motivation pre-test and portfolio scores. Results indicated that after controlling for these variables, experimental condition was significant for teacher ratings of students' motivation ($\beta = .23$). Means testing revealed that the high-stimulating tasks condition was significantly higher than the low-stimulating tasks condition ($d = 0.71$).

Multiple perspectives. Previously noted for expectancies, some studies attempted to integrate various concepts and/or constructs from multiple theories in the design and delivery of the intervention. For those designed to promote values, these interventions may have incorporated different facets of several theories (e.g., Acee & Weinstein 2010; Martin, 2008). Acee and Weintstein (2010) provide an example of one such approach, integrating theories of the expectancy-value framework, value reappraisal, and possible selves through *relevance* and *context and rationale*. They examined the effects of a value-reappraisal intervention for undergraduate students in two separate introductory statistics courses. The intervention was designed to increase the value students placed on developing statistics-related knowledge and skill by reading a series of passages and completing corresponding activities to enhance either attainment value, utility value, or intrinsic value. Two activities without reading passages elicited students to explore costs and benefits associated with learning statistics. Over the course of the study, participants randomly assigned to the experimental condition read six passages and completed eight activities, each related to one of the four topics identified in the value reappraisal intervention. Students in the control condition read four passages and completed four activities related to multicultural education.

The dependent variables included self-report instruments measuring task value, self-efficacy, and endogenous instrumentality. The authors noted that endogenous instrumentality refers to the usefulness of learning specific course content, which is conceptually different than utility value, which refers to the usefulness of completing the course (2010). The authors also noted that items measuring endogenous instrumentality (revised from Husman, Derryberry, Crowson, and Lomax, 2004) used in the study all made reference to the future, and items measuring task value (MSLQ; Pintrich, 1989) did not. Additional dependent variables included pre-and post-intervention exam performance and a measure of choice-behavior. The measure of choice-behavior included optional websites related to statistics that were recommended, but not required, by the course instructors and posted to the course website. Using a feature on the course website, the researchers were able to track those students who accessed these optional websites. A repeated-measures design was utilized, with a pretest, a posttest immediately following the intervention, and another posttest two weeks after the intervention.

Results indicated that the experimental condition demonstrated significant gains in task value from pretest to immediate posttest ($d = 0.54$) and for pretest to delayed posttest ($d = 0.36$). The results were also significant for endogenous instrumentality from pretest to immediate posttest ($d = 0.84$) and from pretest to delayed posttest ($d = 0.50$), suggesting the intervention was effective in increasing perceptions about the usefulness of statistics knowledge and skill in attaining future goals. There were no significant results on either measure in the control condition. In direct comparisons between the two conditions, the experimental condition was more likely than the control condition to access the optional choice-behavior websites, suggesting the intervention generated greater continued interest in statistics ($OR = 9.23$). After controlling for pre-intervention exam performance, students in the experimental condition scored significantly higher than those in the control condition on post-intervention exam performance in one of the classes in the study, but not in the other.

Theoretical perspectives with one study. Like expectancy interventions, there was a theoretical perspective with an intervention targeting value with only one study. In particular, one Transformative Experiences Theory intervention was included in the Lazowski and Hulleman (2015) meta-analysis. According to Transformative Experiences Theory, reframing a student's learning experience as a real life application of the content can enhance everyday value for the material (Pugh, 2011). Pugh (2011) conducted an intervention that targeted *relevance* and *context and rationale* sources of value. One teacher was instructed in the Teaching for Transformative Experience in Science model, which focused on three principles to scaffold learning: frame the content as ideas to be imagined about rather than as concepts to be learned, re-seeing objects as new ideas, and modeling transformative experience. A control teacher in the same school was not exposed to these principles. Students in the experimental condition scored higher on self-reports of transformative experiences of the content (d = 0.48) and on an assessment of science knowledge (d = 0.85).

Research-based Sources of Cost

Stated earlier, cost refers to how much an individual perceives that he or she has to sacrifice or give up in order to accomplish a task. By comparison to expectancy and value, cost has received less attention in theoretical, correlational, and experimental research (Wigfield & Cambria, 2010). However, the literature is emerging and we are beginning to understand more about cost. Based on previous research (Flake, 2012), there may be different components of cost and the research-based sources of cost are aligned with these different components.

First, perceptions of cost increase when students feel that the workload is unreasonable (e.g., 3 hours/night) and/or unnecessary (e.g., busy work) (Parsons et al., 1980; Perez et al., 2014). In this case, source of cost is the *effort and time needed for the activity*. Cost can also increase due to the *effort and time needed for other competing activities* when students have too many other demands on their time or do not know how to effectively manage their time (Barron & Hulleman, in press; Flake, 2012).

Another source of cost demonstrated through research is the *loss of valued alternatives.* If students feel that the learning activity is not worth their time compared to other things they might do (e.g., socializing), they are more likely to experience high cost (Conley, 2012; Perez et al., 2014). Finally, if students feel unsafe and uncomfortable, either physically or psychologically (e.g., nervous, anxious, bored, tired), they are more likely to experience high cost (Eccles et al., 1983; Ramirez & Beilock, 2011). This source is considered *psychological and physical reactions to the activity*.

Interventions Primarily Designed to Decrease Cost

Social belongingness. Social belongingness theory provides the most intervention studies that are designed to decrease cost. This theory examines the degree to which students perceive they belong and are connected to others, and subsequently, how this influences various learning outcomes (Baumeister & Leary, 1995). Thus, interventions grounded in social belongingness theory have largely been aimed at reducing *psychological reactions* by helping students perceive stronger connections between themselves and important others in the learning context. Some studies have attempted to build a sense of connection and belonging between students and teachers (e.g., Gelhbach et al., 2014), and others have focused more on minority populations in academic settings (e.g., Hausmann et al., 2009; Walton & Cohen, 2007).

The interventions have impacted a variety of outcomes, including: perceptions of similarity (e.g., Gehlbach et al., 2014); ratings of teacher-student interactions (e.g., Gehlbach et al., 2014); homework completion (e.g., Gehlbach et al., 2014); classroom attendance (e.g., Gehlbach et al., 2014); achievement behavior (e.g., Walton & Cohen, 2007); grades (e.g., Gehlbach et al., 2014; Walton & Cohen, 2007); perceived social and academic integration (e.g., Hausmann et al., 2009) and academic fit (e.g., Walton &

Cohen, 2007); perceived cohesion (e.g., Hausmann et al., 2009); goal commitment (e.g., Hausmann et al., 2009); intentions to persist (e.g., Hausmann et al., 2009); institutional commitment (e.g., Hausmann et al., 2009); and challenge-seeking in course selection (e.g., Walton & Cohen, 2007).

Walton and Cohen (2007) examined the impact of potential stigmatization associated with belonging uncertainty in a sample of undergraduate students. The authors hypothesized that Black students would be more susceptible to decreases in motivation and achievement when faced with belonging uncertainty due to the negative characteristics of this group in academic settings.

In this study, the experimental intervention was designed to mitigate feelings of belonging uncertainty in college. Students in the experimental condition were first provided information suggesting that most college students experience some sense of worry or doubt about belonging on campus but that these feelings diminish over time. Students in the control condition were provided information suggesting college students' social-political beliefs become more developed over time. All students subsequently completed self-report scales measuring the following: social fit, academic identification, enjoyment of academic work, self-efficacy, and potential to succeed in college, possible academic selves, and anxiety. To assess levels of academic challenge-seeking, they were also asked to indicate courses that they would be interested in taking. Each course had descriptions such as difficulty level as rated by other students, as well as, the amount of information students reported they learned in the class. Half were rated as difficult, yet highly educational and half were rated as easy and moderately educational.

Over the course of the next seven days, students were then asked to complete two questionnaires each day. The first questionnaire included the measures of social fit, self-efficacy, and potential to succeed in college. The second questionnaire again included these measures but also asked to students to report whether they engaged in academic activities such as emailing professors, attending review/study sessions, hours spent studying, and participating in class. Finally, the students were asked to rate the level of adversity they experienced that day.

Results both immediately after the information was provided to students (feelings of academic fit improving over time vs. social-political beliefs) and over the course of seven days were reported. Immediately after, Black students in the experimental condition reported higher levels of academic fit (d = 1.37), potential to succeed in college (d = 1.63), and selected more challenging courses (d = 1.11) compared to Black students in the control condition. White students, on the other hand, demonstrated the opposite pattern for academic fit, and indicated that White students in the control condition reported higher levels of academic fit compared to White students in the experimental condition (d = 1.22).

After seven days, no differences emerged between Black students in the experimental and control conditions for levels of academic fit. Once again, White students in the control condition reported higher levels of academic fit compared to White students in the experimental condition (d=1.32). Black students in the experimental condition reported higher potential to succeed compared to Black students in the control condition (d = 0.75).

The intervention also appeared to sustain Black students' sense of academic fit, specifically on days they reported high levels of adversity. Black students in the experimental condition reported less variation in their sense of academic fit relative to level of adversity compared to Black students in the control condition ($d = 1.02$). While sense of academic fit remained stable for Black students in the experimental condition, sense of academic fit declined on days when adversity was rated high versus moderate ($d = 0.51$) and low ($d = 0.63$) for Black students in the control condition. Furthermore, Black students in the experimental condition reported engaging in more academic behaviors ($d = 1.47$), studying longer ($d = 1.54$), and sent more emails to professors ($d = 1.70$) compared to Black students in the control condition. Finally, Black students in the treatment condition demonstrated higher gains in GPA compared to Black students in the control condition ($d = 1.10$) and Black students across the entire campus ($d = 0.72$). On the other hand, White students in the control condition demonstrated higher gains in GPA compared to White students in the treatment condition ($d = 0.88$), although changes in GPA for White students in the treatment condition did not differ from White students across campus.

Multiple perspectives. Like expectancy and value, there have been some interventions that have combined multiple theoretical perspectives to design an intervention. For example, Jameison and colleagues (2010) combined components of self-affirmation theory with components of achievement emotions (namely, anxiety which refers to worrying about the consequences of performance, subsequently undermining working memory and other performance outcomes; Ramirez & Beilock, 2011). This intervention served to target the cost source of *physical reactions to an activity*.

In this randomized laboratory experiment, Jameison and colleagues (2010) assigned undergraduates who were preparing to take the Graduate Record Examination (GRE) to either a reappraisal experimental condition or a control group. Prior to taking a practice GRE, students in the reappraisal condition read instructions that feeling anxious while taking standardized tests was not only normal, but that research showed that this arousal was not detrimental and could in fact improve performance on the tests. They were also instructed that if they did feel anxious, they should remind themselves that this anxiety could actually be helping them do well on the test. Student in the control condition read instructions only that feeling anxious during standardized tests was a normal response. The dependent variables in the study were performance on the practice GRE test, physiological reactions to testing, subjective experiences after taking the actual GRE, and actual GRE performance (the students in the study completed a survey and provided a copy of their actual GRE scores following the study during a follow-up session 1-3 months later).

Results indicated that students in the reappraisal condition scored significantly higher on the math section of the practice test compared to the control group (d = 0.82). However, no significant differences emerged on verbal section of the practice test. Physiological reactions to testing were assessed by a saliva test measuring sAA levels (a measure of SNS activation, indexing engagement and challenge orientation; 2010). sAA levels taken before and after testing revealed that students in the reappraisal condition exhibited a significant increase in sAA levels compared to the control group (d = 1.01).

After taking the actual GRE, students in the reappraisal group reported higher endorsement that arousal helped their performance (*d* = 0.99), worried less about feeling anxious (*d* = 0.67), and reported feeling less unsure of themselves (*d* = 0.97) compared to the control condition. These results suggested that the intervention generalized to the actual testing situation. Finally, similar to the practice test, students in the reappraisal condition performed significantly higher compared to the control condition on the math section of the actual GRE (*d* = 1.03). However, no significant differences were found on the verbal section.

Interventions Designed to Promote Both Expectancy and Value

Achievement goal theory. Some theoretical perspectives have produced interventions that target sources related to both expectancy and value. That is, sources targeted by interventions guided by these theories do not primarily fit into either expectancy or value only. Achievement Goal Theory represents once such theoretical perspective. Achievement Goal Theory (Elliot, 2005) suggests that students' goals for engaging in a task shape how they approach, experience, and react to achievement situations. Specifically, the theory differentiates between mastery goals (comparison to one's own self) and performance goals (comparison to others). Further, the theory differentiates between approach goals (to obtain a positive outcome) and avoidance goals (to avoid a negative outcome). Together, the combination of these goals produces a 2X2 framework consisting of the following goal orientations: Mastery Approach, Performance Approach, Mastery Avoidance, and Performance Avoidance. The goal pursuit to either draw comparisons to one's self (mastery goal) or to draw comparisons to others (performance) can be driven based on the value beliefs one holds. Put simply, it depends on what is most important to the individual. By contrast, the goal pursuit to either attain a positive outcome (approach) or avoid a negative outcome (avoidance) can be driven based on one's expectancy for success.

One set of interventions aimed at promoting student expectancies (*growth experiences, effort attributions*, and *feedback*) and values (*context and rationale*) has been conducted using achievement goal theory. Many of these interventions were designed to increase students' adoption of mastery goals (comparisons to oneself, rather than others) and focus on individual improvement by using different activities and assignments (e.g., Hoyert & O'Dell, 2006). Some studies focused primarily on feedback provided by the teacher stressing the importance of learning and improvement (e.g., Muis et al., 2013) whereas others provided education about achievement goals (e.g., Hoyert & O'Dell, 2006). Achievement goal interventions have impacted outcomes such as: mastery and performance goal change (e.g., Hoyert & O'Dell, 2006; Muis et al., 2013; Ranellucci et al., 2013); test grades (e.g., Hoyert & O'Dell, 2006); course grades (e.g., Hoyert & O'Dell, 2006; Muis et al., 2013; Ranellucci et al., 2013); self-efficacy (e.g., Muis et al., 2006); metacognitive self-regulation (e.g., Muis et al., 2013); test anxiety (e.g., Muis et al., 2013); interest-based studying (e.g., Ranellucci et al., 2013); and perceived task difficulty (e.g., Ranellucci et al., 2013).

As an example, Hoyert and O'Dell (2006) conducted two intervention studies aimed at altering achievement goal orientations among struggling college students enrolled in an introductory psychology course. In the first study, a guest speaker came to the class and provided a lecture and discussion about how to set mastery goals and to examine various meanings behind failure. Students then completed exercises and a writing assignment to influence their adoption of mastery goals. Exercises covered topics such as goal setting and study strategies. The writing assignment required students to write about defining goals, describe individuals who illustrate the traits associated with the goal type, and consider personal experiences related to the goal orientation. In the second study, a similar intervention protocol was used, with the exception that instead of a guest lecturer, the students completed a tutorial delivered through a CD-ROM program. The intervention took place following the first examination in each of the studies.

Comparisons were made between students who either participated or did not participate in the intervention. An additional, at-risk group was also identified as a target sample based on high scores in the area of performance goal orientation and poor performance on the first examination. The researchers examined the effects of the intervention on goal orientation change, examination grades, and final grades at the end of the semester.

Results of study 1 indicated that students in the experimental condition did not show significant changes in either performance or mastery goal orientation. Students in the control condition, on the other hand, demonstrated decreases in mastery goal orientation ($d = 0.52$) and increases in performance goal orientation ($d = 0.44$) from the beginning of the semester to the end of the semester. Furthermore, students in the experimental condition generally improved over the course of the semester on examination grades while students in the control condition generally scored lower on subsequent examinations. The largest difference was found for the last examination, which had approximately a 20-point difference ($d = 1.68$). When final course grades were analyzed, students in the experimental condition had higher average grades in comparison to students in the control condition (d = 1.33).

When only at-risk students were included in analyses, a similar pattern of results emerged. At-risk students in the experimental condition did not show significant changes in either performance or mastery goal orientation; however, at-risk students in the control condition demonstrated decreases in mastery goal orientation (d = 0.50) and increases in performance goal orientation (d = 0.43).

Results of study 2 focused largely on the at-risk students. At-risk students in the experimental condition increased in mastery goal orientation (d = 0.75) and decreased in performance goal orientation (d = 1.40). At-risk students in the experimental condition scored higher on each subsequent examination throughout the semester but effect sizes could not be calculated due to insufficient data reported in the original paper. However, the largest difference was found for the last examination, which was approximately a 25-point difference. At the end of the semester, 76% of the students in the control condition failed the course compared to only 37% of the students in the experimental condition, χ^2 (4, N = 222) = 12.38.

Possible selves. In addition, Possible selves theory (Markus & Nurius, 1986) also blends aspects of expectancy and value. Possible selves theory suggests that students' conception of what they might become (both desired and feared) serve as incentives for future behavior and a way to evaluate their current behavior. When students think about what they might become in the future, they are forming some type of expectancy that they can actualize this future self. The incentive for future behavior, however, will be based on what that student values and these values will also drive behaviors toward that self.

Possible selves theory has inspired interventions that help students draw connections between successful future selves and current school involvement through interactive activities and written reflections (e.g., Oyserman et al., 2002, 2006). As such, they encourage students to think about expectancies (future *perceptions of their ability and skill* and future *success experiences*) as well as values (via *context and rationale*). Interventions grounded in this theory have been successful at enhancing self-reports about academic possible selves (e.g., Oyserman et al., 2002; Oyserman et al., 2006) and connection to school (e.g., Oyserman et al., 2002); recognizing the value of education for reaching career goals (e.g., Day et al., 1994); plausible strategies for attaining possible selves (e.g., Oyserman et al., 2002); and various behavioral outcomes such as effort (e.g., 2002), attendance (e.g., 2002; Oyserman et al., 2006), and reductions in disruptive behaviors (e.g., 2006).

Oyserman, Terry, and Bybee (2002), for example, tested an intervention targeted at possible selves for African American middle school students. In this mixed methods field study, students in the experimental condition received an intervention consisting of a 9 week after school program designed to enhance students' abilities to see themselves as successful adults. Further, they were encouraged to draw connections between these selves and their current school involvement. The intervention program consisted of: creating a group, adult images, time lines, possible selves and strategies boards, solving everyday problems, wrapping up/moving forward, building an alliance and developing communication skills, and jobs, careers, and informational interviewing. A control condition that did not receive the intervention was also used in the study.

Data were collected over three years, resulting in three cohorts of students and included comparisons in the following areas: connection and bonding to school, concern for doing well, "balanced" academic possible selves (qualitative, sentence completion items assessing student reports of positive and feared possible selves), plausible strategies (qualitative responses for attaining school-oriented possible selves), and school behavior (misbehavior/discipline referrals and attendance).

After controlling for gender, cohort, and baseline levels for dependent measures prior to the intervention, results indicated that students in the experimental condition demonstrated a greater sense of bonding to school and concern for doing well academically compared to the control condition ($d = 0.36$ and 0.25, respectively). Students in the experimental condition also reported greater balance between positive and feared academic selves and identified more plausible strategies for attaining academic possible selves relative to the control condition ($d = 0.28$ and 0.25, respectively). Regarding school behavior, males in the experimental condition self-reported less behavior problems/discipline referrals in comparison to males in the control condition ($d = 0.33$). Finally, students in the experimental condition also had better attendance patterns compared to students in the control condition ($d = 0.45$).

Theoretical perspectives with one study. One Need for Achievement theory intervention was included in the Lazowski and Hulleman (2015) meta-analysis. Need for Achievement theory (McClelland et al., 1976) suggests the importance of achieving mastery, high achievement, and out-doing others to reach one's full potential are important factors in student motivation. Quintanilla (2007) conducted a need for achievement study that targeted both expectancies via *support and scaffolding, clear expectations,* and *feedback* and values via *relevance, choice and control,* and *context and rationale*. In this intervention, undergraduate students were assigned to a ten week intervention consisting of motivation strategies infused within a first-year student class. The 45-minute experiential lessons occurred weekly and emphasized risk-taking strategies, goal-setting, planning, and reflection. The control condition was enrolled in the first-year class without the experiential lessons. Students in the experimental condition self-reported higher levels of intrinsic and extrinsic goals, academic beliefs and self-efficacy, text anxiety, critical thinking, and self-regulation (average d = 0.36).

Conclusion

In this paper, I have attempted to make an argument for the importance of systematic reviews (e.g., narrative reviews and/or meta-analyses) as they relate to educational research. When done correctly, these techniques have the potential to offer a greater amount of validity and reliability, as they are based on multiple studies rather than one study alone. These findings can also bridge the researcher-practitioner divide by summarizing cumulative, research-based knowledge in a format that may be more digestible to those not intimately involved in the research. The field of achievement motivation research is no different. At the time of this writing, there have been no meta-analyses conducted to examine motivation interventions in authentic educational settings. This is important as there are currently more correlational or laboratory studies examining the impact of motivation on educational outcomes by comparison to field experiments. It is thus critical to provide researchers and practitioners alike with information about the effectiveness of these interventions as well as to identify the characteristics of the studies that appear to have bearing on the interventions' effectiveness.

As the research in the area of achievement motivation continues to expand and related theories and constructs continue to proliferate, it is also important to provide some form of cohesion to the field, lest we risk challenges. For example, although it is important to maintain distinctions among the various motivational theories and respect theoretical space, a challenge is that there appears to be some overlap in the theories and the constructs therein. At times, the overlap occurs when different constructs have the same name or label; at other times, the same theoretical construct may have a different name or label (e.g., Hulleman et al., 2010; Marsh, 1994). Thus, without some organization and cohesion, we may increase the likelihood of committing a jingle and jangle fallacy. Organization and cohesion can also serve to bridge the researcher-practitioner divide by providing the practitioner with a more accessible way to make sense of the quite detailed nature of motivation research, theory, and constructs. In this way, practitioners can use an organized framework with which to base possible interventions and/or educational practice.

Toward this end, I used the expectancy-value framework as a means to capture and organize the various theoretical perspectives that have tested interventions through experimental designs in authentic educational settings. By doing so, the theories and the interventions were aligned with Eccles' and her colleagues' (1983) over-arching constructs of expectancy, value, or cost.. In addition, I identified the various sources or drivers of the interventions within the categories of expectancy, value, or cost. If we are to leverage the relationships between the expectancy-value framework and learning outcomes, it is critical to identify the sources of expectancy, value, and cost that are amenable to change and potentially accessible to educational practitioners. By targeting motivation gaps, educational practitioners, policy-makers, and researchers have a potentially powerful tool to further close achievement gaps and inspire more students to persist academically, both in the short- and long-term.

References

References marked with an asterisk indicate studies included in the Lazowski and Hulleman (2015) meta-analysis.

* Acee, T. W., & Weinstein, C. E. (2010). Effects of a value-reappraisal intervention on statistics students' motivation and performance. *The Journal of Experimental Education*, *78*, 487-512.

* Aronson, J., Fried, C. B., & Good, C. (2002). Reducing the effects of stereotype threat on african american college students by shaping theories of intelligence. *Journal of Experimental Social Psychology*, *38*, 113-125.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Barron, K.E., & Hulleman, C.S. (In press). The expectancy-value-cost model of motivation. In J. Eccles & K. Salmela (Eds.), *International encyclopedia of the social and behavioral sciences (2nd edition)*. Elsevier Publishing.

Baumeister, R., & Leary, M. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin, 117*, 497–529.

*Bernacki, M., Nokes-Malach, T., Richey, J. E., & Belenky, D. M. (2014). Science diaries: A brief writing intervention to improve motivation to learn science, *Educational Psychology: An International Journal of Experimental Educational Psychology,* doi:10.1080/01443410.2014.895293.

Berwick, D. M. (2008). The science of improvement. *Journal of the American Medical Association, 299*, 1182-1184.

* Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development*, *78*, 246-263.

Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin, 117,* 187-215.

*Boese, G. D., Stewart, T. L., Perry, R. P., & Hamm, J. M. (2013). Assisting failure-prone individuals to navigate achievement transitions using a cognitive motivation treatment (attributional retraining). *Journal of Applied Social Psychology*, *43*, 1946-1955.

*Bordine, J. J. (1998). *Improving reading, self-concept, and motivation in at-risk kindergarten students: An intervention* (Unpublished doctoral dissertation). The University of Texas at Austin, Austin, TX.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research, 73*, 125-230.

Brophy, J. E. (2004). *Motivating students to learn*. (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of the Learning Sciences, 2*, 141-178.

Bryk, A. S., Gomez, L. M., & Grunow, A. (2011). *Getting ideas into action: Building networked improvement communities in education*. In Hallinan, M. T. (Ed.), Frontiers in Sociology of Education (pp. 127-162). The Netherlands: Springer.

Casillas, A., Robbins, S., Allen, J., Kuo, Y. L., Ann Hanson, M., & Schmeiser, C. (2012). Predicting early academic failure in high school from prior academic achievement, psychosocial characteristics, and behavior. *Journal of Educational Psychology, 104*, 407-420.

* Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science*, *313*, 1307-1310.

* Cohen, G. L., Garcia, J., Purdie-Vaughns, V., Apfel, N., & Brzustoski, P. (2009). Recursive processes in self-affirmation: Intervening to close the minority achievement gap. *Science, 324*, 400-403.

Conley, A. M. (2012). Patterns of motivation beliefs: Combining achievement goal and expectancy-value perspectives. *Journal of Educational Psychology, 104(1)*, 32–47.

Connell, J. P., Klem, A. M., Lacher, T., Leiderman, S., & Moore, W., with Deci, E. L. (2009). *First Things First: Theory, research, and practice*. Toms River, NJ: Institute for Research and Reform in Education. Available online: http://www.irre.org/publications

* Cook, J. E., Purdie-Vaughns, V., Garcia, J., & Cohen, G. L. (2012). Chronic threat and contingent belonging: Protective benefits of values affirmation on identity development. *Journal of Personality and Social Psychology*, *102*, 479-496.

*Craven, R. G., Marsh, H. W., & Debus, R. L. (1991). Effects of internally focused feedback and attributional feedback on enhancement of academic self-concept. *Journal of Educational Psychology*, *83*, 377-392.

*Cueva, C. (2006). *An achievement motivation program for young minority students utilizing cognitive-behavioral techniques* (Unpublished doctoral dissertation). Texas A & M, Corpus Christi, TX.

Daniel, D.B. (2012). Promising principles: Translating the science of learning to educational practice. *Journal of Applied Research in Memory and Cognition, 1*, 251-253.

Datnow, A., & Stringfield, S. (2000). Working together for reliable school reform. *Journal of Education for Students Placed at Risk, 5*, 183-204.

* Day, J. D., Borkowski, J. G., Punzo, D., & Howsepian, B. (1994). Enhancing possible selves in mexican american students. *Motivation and Emotion*, *18*, 79-103.

Deci, E. L., & Ryan, R. M. (1985). *Self-Determination*. John Wiley & Sons, Inc.

Deci, E. L., & Ryan, R. M. (2002). Overview of self-determination theory: An organismic dialectical perspective. *Handbook of self-determination research*, 3-33.

Design-based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher, 32*, 5–8.

*Duckworth, A.L., Kirby, T., Gollwitzer, A., & Oettingen, G. (in press). From fantasy to action: Mental contrasting with implementation intentions (MCII) improves academic performance in children.

Dudley, D., Okely, A., Pearson, P., & Cotton, W. (2012). A systematic review of the effectiveness of physical education and school sport interventions targeting physical activity, movement skills and enjoyment of physical activity. *European Physical Education Review, 17*, 353-378.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*, 1428–1446.

*Durik, A. M., Schmidt, J., Shumow, L., Rodenbeck, B. (April, 2014). *Self-generated utility among a diverse sample of adolescent students: An analysis of grade level and gender.* Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41*, 1040-1048.

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development*. Philadelphia, PA: Taylor & Francis.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological review, 95*(2), 256-273.

Easton, J.Q. (2013, June). *Using measurement as leverage between developmental research and educational practice*. Center for the Advanced Study of Teaching and Learning: Charlottesville, Virginia.

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation* (pp. 74-146). San Francisco, CA: W.H. Freeman.

Elliot, A. J. (2005). A conceptual history of the achievement goal construct. *Handbook of Competence and Motivation*, *16*, 52-72.

Flake, J. K. (2012). *Understanding cost: The forgotten component of expectancy-value* *theory. Unpublished Master's Thesis. James Madison University: Harrisonburg, VA.*

*Fordyce, M. W. (1983). A program to increase happiness: Further studies. *Journal of Counseling Psychology*, *30*, 483-498.

*Fordyce, M. W. (1977). Development of a program to increase personal happiness. *Journal of Counseling Psychology*, *24*, 511-521.

Fredericks, J., McColskey, W., Meli, J., Mordica, J., Montrosse, B., & Mooney, K. (2011). *Measuring student engagement in upper elementary through high school: A description of 21 instruments* (Issues & Answers Report, REL 2011–098). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast.

* Froiland, J. M. (2011, April). Parental autonomy support and student learning goals: A preliminary examination of an intrinsic motivation intervention. In *Child & Youth Care Forum* (Vol. 40, No. 2, pp. 135-149). Springer US.

Furrer, C., & Skinner, E. (2003). Sense of relatedness as a factor in children's academic engagement and performance. *Journal of Educational Psychology*, *95*, 148-162.

*Gehlbach, H., Brinkworth, M. E., King, A., Hsu, L., McIntyre, J., & Rogers, T. (2014). *Creating birds of similar feathers: Leveraging similarity to improve teacher-student relationships and academic achievement.* Manuscript submitted for publication.

Global Science Forum. (2006). *Encouraging Student Interest in Science and Technology Studies*. OECD Publishing.

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist, 54*, 493-503.

*Good, C., Aronson, J., & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology*, *24*, 645-662.

* Greenstein, T. (1976). Behavior change through value self-confrontation: A field experiment. *Journal of Personality and Social Psychology*, *34*, 254-262.

Grissmer, D., Grimm, K. J., Aiyer, S. M., Murrah, W. M., & Steele, J. S. (2010). Fine motor skills and early comprehension of the world: Two new school readiness indicators. *Developmental Psychology, 46*, 1008-1017.

* Guthrie, J. T., Wigfield, A., Humenick, N. M., Perencevich, K. C., Taboada, A., & Barbosa, P. (2006). Influences of stimulating tasks on reading motivation and comprehension. *The Journal of Educational Research*, *99*, 232-246.

* Hall, N. C., Perry, R. P., Goetz, T., Ruthig, J. C., Stupnisky, R. H., & Newall, N. E. (2007). Attributional retraining and elaborative learning: Improving academic development through writing-based interventions. *Learning and Individual Differences*, *17*, 280-290.

* Hall, N. C., Hladkyj, S., Perry, R. P., & Ruthig, J. C. (2004). The role of attributional retraining and elaborative learning in college students' academic development. *The Journal of Social Psychology*, *144*, 591-612.

Harackiewicz, J.M., & Barron, K.E. (2004). Conducting social psychological research in educational settings: "Lessons we learned in school". In C. Sansone, C. Morf, & A.

Panter (Eds.) *Handbook of methods in social psychology* (pp. 471-484). Thousand Oaks, CA: Sage Publications.

Harackiewicz, J.M., & Hulleman, C. S. (2010). The importance of interest: The role of achievement goals and task values in promoting the development of interest. *Social and Personality Psychology Compass, 4,* 42-52.

* Harackiewicz, J. M., Rozek, C. S., Hulleman, C. S., & Hyde, J. S. (2012). Helping parents to motivate adolescents in mathematics and science: An experimental test of a utility value intervention. *Psychological Science*, *23*, 899-906.

*Hausmann, L. R., Ye, F., Schofield, J. W., & Woods, R. L. (2009). Sense of belonging and persistence in White and African American first-year students. *Research in Higher Education*, *50*, 649-669.

Headden, S. (2014). *Beginners in the classroom: What the changing demographics of teaching mean for schools, students, and society*. Carnegie Foundation for the Advancement of Teaching: Stanford, CA. *Journal of Mathematical and Statistical Psychology, 36*, 123-131.

Hidi, S. (1990). Interest and its contribution as a mental resource for learning. *Review of Educational Research, 60*, 549-571.

*Hidi, S., Berndorff, D., & Ainley, M. (2002). Children's argument writing, interest and self-efficacy: An intervention study. *Learning and Instruction*, *12*, 429-446.

Hidi, S. & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research, 70*, 151-179.

Hidi, S. & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist, 41*, 111-127.

*Hong, H., & Lin-Siegler, X. (2011). How learning about scientists' struggles influences students' interest and learning in physics. *Journal of Educational Psychology*. Advance online publication. doi:10.1037/a0026224.

* Hoyert, M. S., & O'Dell, C. D. (2006). A brief intervention to aid struggling students: A case of too much motivation? *Journal of Scholarship of Teaching and Learning*, *6*(1), 1-13.

Hsieh, P., Acee, T., Chung, W.-H., Hsieh, Y.-P., Kim, H., Thomas, G.D., & You, J. (2005). Is educational intervention research on the decline? *Journal of Educational Psychology, 97*, 523-529.

Hulleman, C. S., & Barron, K. E. (In press). Motivation interventions in education: Bridging theory, research, and practice. To appear in: L. Corno & E. M. Anderman (Eds.), *Handbook of Educational Psychology*, 3rd Ed. (2016). Routledge, Taylor and Francis: New York, NY.

Hulleman, C. S., Barron, K. E., Kosovich, J. J., & Lazowski, R. A. (in press). Achievement Motivation. To appear in A. Lipnevich, F. Preckel, & R. Roberts (Eds.), *Psychosocial skills and school systems in the Twenty-first century: Theory, research, and applications*. Springer.

Hulleman, C. S., Barron, K. E., Lazowski, R. A., & Harackiewicz, J. M. (2011, April). *Relevance is motivating: Lessons learned from applying expectancy-value theory in the real-world.* Paper presented at the annual conference of the American Educational Research Association, New Orleans, LA, April 8-12.

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness*, *2*, 88-110.

* Hulleman, C. S., Godes, O., Hendricks, B. L., & Harackiewicz, J. M. (2010). Enhancing interest and performance with a utility value intervention. *Journal of Educational Psychology*, *102*, 880-895.

* Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science*, *326*, 1410-1412.

Hulleman, C. S., Schrager, S. M., Bodmann, S. M., & Harackiewicz, J. M. (2010). A meta-analytic review of achievement goal measures: Different labels for the same constructs or different constructs with similar labels? *Psychological Bulletin*, *136*, 422-449.

*Jamieson, J. P., Mendes, W. B., Blackstock, E., & Schmader, T. (2010). Turning the knots in your stomach into bows: Reappraising arousal improves performance on the GRE. *Journal of Experimental Social Psychology*, *46*, 208-212.

Kaplan, A., Katz, I., & Flum, H. (2012). Motivation theory in educational practice: Knowledge claims, challenges, and future directions. In K. Harris et al. (Eds) *Educational psychology handbook, Vol 2: Individual differences and cultural and contextual factors* (pp. 165-194). Washington, DC, US: American Psychological Association.

Kaplan, A., & Maehr, M. L. (2007). The contributions and prospects of goal orientation theory. *Educational Psychology Review*, *19*, 141-184.

* Kitsantas, A., Robert, A. R., & Doster, J. (2004). Developing self-regulated learners: Goal setting, self-evaluation, and organizational signals during acquisition of procedural skills. *The Journal of Experimental Education*, *72*, 269-287.

Lazowski, R. A., Hulleman, C. S. (2015, April). *Motivation interventions in education: A meta-analytic review. Paper presented at the annual conference of the American Educational Research Association, Chicago, IL.*

Lepper, M. R., & Henderlong, J. (2000). Turning" play" into" work" and" work" into" play": 25 years of research on intrinsic versus extrinsic motivation. In C. Sansone & J. M. Harackiewicz (Eds.), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance* (pp. 257-307). New York: Academic Press.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Locke, E.A., & Latham, G.P. (1990). *A theory of goal setting and task performance*. Englewood Cliffs, NJ: Prentice-Hall.

Locke, E.A., & Latham, G.P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist, 57*, 705–717.

Maehr, M. L., & Midgley, C. (1999). Creating optimum environments for students of diverse sociocultural backgrounds. In J. Block, S. T. Everson, & T. R. Guskey (Eds.), *Comprehensive school reform: A program perspective* (pp. 355–-375). Dubuque, IA: Kendall/Hunt.

Maehr, M.L., & Meyer, H.A. (1997). Understanding motivation and schooling: Where we've been, where we are, and where we need to go. *Educational Psychology Review, 9*, 371-409.

Markus, H., & Nurius, P. (1986). Possible selves. *American psychologist, 41*, 954-969.

Marsh, H. W. (1994). Sport motivation orientations: Beware of jingle-jangle fallacies. *Journal of Sport & Exercise Psychology*, *16*, 365-380.

Marsh, H. W., & Shavelson, R. (1985). Self-concept: Its multifaceted, hierarchical structure. *Educational Psychologist, 20*, 107-123.

Marshall, M., Pronovost, P., & Dixon-Woods, M. (2013). Promotion of improvement as a science. *Lancet, 381*, 419-421.

* Martin, A. J. (2008). Enhancing student motivation and engagement: The effects of a multidimensional intervention. *Contemporary Educational Psychology*, *33*, 239-269.

Martin, A. J., & Dowson, M. (2009). Interpersonal relationships, motivation, engagement, and achievement: Yields for theory, current issues, and educational practice. *Review of Educational Research, 79*, 327-365. doi:10.3102/0034654308325583.

McClelland, D., Atkinson, J., Clark, R., & Lowell, E. (1976). *The achievement motive*. New York: Irvington Publishers, Inc.

*McGinley, J. J., & Jones, B. D. (2014). A brief instructional intervention to increase students' motivation on the first day of class. *Teaching of Psychology, 41*, 158-162.

Messick, S. (1990). *Validity of test interpretation and use.* Research Rep. No. 90-11. Princeton, NJ: Educational Testing Service.

Metcalf, B., Henley, W., & Wilkin, T. (2013). Effectiveness of intervention on physical activity of children: Systematic review and meta-analysis of controlled trials with objectively measured outcomes. *British Journal of Sports Medicine, 47*, 226-226.

Midgley, C., & Edelin, K. C. (1998). Middle school reform and early adolescent well-being: The good news and the bad. *Educational Psychologist, 33*, 195-206.

* Miyake, A., Kost-Smith, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L., & Ito, T. A. (2010). Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science*, *330*, 1234-1237.

* Morisano, D., Hirsh, J. B., Peterson, J. B., Pihl, R. O., & Shore, B. M. (2010). Setting, elaborating, and reflecting on personal goals improves academic performance. *Journal of Applied Psychology*, *95*, 255-264.

Mortensen, C. R., & Cialdini, R. B. (2010). Full-cycle social psychology for theory and application. *Social and Personality Psychology Compass*, *4*, 53-63.

Mueller, C. M., & Dweck, C. S. (1998). Praise for intelligence can undermine children's motivation and performance. *Journal of Personality and Social Psychology*, *75*, 33-52.

*Muis, K. R., Ranellucci, J., Franco, G. M., & Crippen, K. J. (2013). The interactive effects of personal achievement goals and performance feedback in an undergraduate science class. *The Journal of Experimental Education*, *81*, 556-578.

National Research Council and Institute of Medicine. (2004). *Engaging schools: Fostering high school students' motivation to learn*. Washington, DC: The National Academies Press. Available: www.nap.edu/books/0309084350/html/

Nicholls, J. G. (1979). Quality and equality in intellectual development: The role of motivation in education. *American Psychologist, 34*, 1071-84.

Nicholls, J. G. (1984). Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. *Psychological review, 91*, 328-346.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research*, *78*, 33-84.

*Oyserman, D., Bybee, D., & Terry, K. (2006). Possible selves and academic outcomes: How and when possible selves impel action. *Journal of Personality and Social Psychology, 91*, 188-204.

* Oyserman, D., Terry, K., & Bybee, D. (2002). A possible selves intervention to enhance school involvement. *Journal of Adolescence*, *25*, 313-326.

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research, 66*(4), 543-578.

Pajares, F. (2002). Gender and perceived self-efficacy in self-regulated learning. *Theory Into Practice, 41*(2), 116-125.

Parsons, J. S., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1980). *Self-perceptions, task perceptions and academic choice: origins and change* (pp. 1–98). Ann Arbor, MI.

* Patall, E. A., Cooper, H., & Wynn, S. R. (2010). The effectiveness and relative importance of choice in the classroom. *Journal of Educational Psychology*,*102*, 896-915.

Patrick, B. C., Hisley, J., & Kempler, T. (2000). "What's everybody so excited about?": The effects of teacher enthusiasm on student intrinsic motivation and vitality. *The Journal of Experimental Education*, *68*, 217-236.

*Paunesku, D., Walton, G., Romero, C., Smith, E., Yeager, D., Dweck, C. (2014). *Mindset interventions are a scalable treatment for academic underachievement.* Manuscript submitted for publication.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18*, 315-341.

Perez, T., Cromley, J. G., & Kaplan, A. (2014). The role of identity development, values, and costs in college STEM retention. *Journal of Educational Psychology, 106(1)*, Feb 2014, 315-329.

Perry, N. E., Turner, J. C., & Meyer, D. K. (2006). Classrooms as contexts for motivating learning. P. Alexander & P. H. Winne (Eds.), *Handbook of Educational Psychology: 2nd Edition* (pp. 327-348). Lawrence Erlbaum Associates: Mahwah, NJ.

Pintrich, P.R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology, 95,* 667-686.

Pugh, K. J. (2011). Transformative experience: An integrative construct in the spirit of Deweyan pragmatism. *Educational Psychologist*, *46*, 107-121.

*Pugh, K. J. (2014). *Supporting deep engagement: Development and evaluation of the teaching for transformative experiences in science (TTES) model.* Manuscript under review.

*Quintanilla, Y. T. (2007). *Achievement motivation strategies: An integrative achievement motivation program with first year seminar students* (Unpublished doctoral dissertation),

Texas A & M University, Corpus Christi, TX.

*Radil, A. I. (2012). *An autonomy support motivation intervention with pre-services teachers: Do the strategies that they intend to use change?* (Unpublished master's thesis), University of Alberta, Ottawa, ON, Canada.

*Ramirez, G., & Beilock, S. L. (2011). Writing about testing worries boosts exam performance in the classroom. *Science*, *331*, 211-213.

*Ranellucci, J., Hall, N. C., K. R., & Lajoie, S. P. (2014). *Mastery, maladaptive learning behaviors, and academic achievement: An intervention approach.* Manuscript under review.

Raudenbush, S. W. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher, 34,* 25-31.

Raudenbush, S. W. (2008). Advancing educational policy by advancing research on instruction. *American Educational Research Journal, 45,* 206-230.

* Reeve, J., Jang, H., Carrell, D., Jeon, S., & Barch, J. (2004). Enhancing students' engagement by increasing teachers' autonomy support. *Motivation and Emotion*, *28*, 147-169.

Renninger, K. A., & Hidi, S. (2011). Revisiting the conceptualization, measurement, and generation of interest. *Educational Psychologist*, *46*, 168-184.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130*, 261-288.

Robinson, D.H., Levin, J.R., Thomas, G.D., Pituch, K.A., & Vaughn, S. (2007). The incidence of "causal" statements in teaching-and-learning research journals. *American Educational Research Journal, 44*, 400-413.

Rokeach, M. (1973). *The nature of human values*. New York: Free Press.

* Ruthig, J. C., Perry, R. P., Hall, N. C., & Hladkyj, S. (2004). Optimism and attributional retraining: Longitudinal effects on academic achievement, test anxiety, and voluntary course withdrawal in college students. *Journal of Applied Social Psychology*, *34*, 709-730.

Ryan, R. M., & Brown, K. W. (2005). Legislating competence: High-stakes testing policies and their relations with psychological theories and research. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 354-372). Guilford Press.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*, 68-78.

* Schaffner, E., & Schiefele, U. (2007). The effect of experimental manipulation of student motivation on the situational representation of text. *Learning and Instruction*, *17*, 755-772.

Schunk, D. H., Pintrich, P. R., & Meece, J. L. (2008). *Motivation in education: Theory, research, and applications* (4th Edition). Pearson.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin Company.

*Sherman, D. K., Hartson, K. A., Binning, K. R., Purdie-Vaughns, V., Garcia, J., Taborsky-Barba, S., & Cohen, G. L. (2013). Deflecting the trajectory and changing the narrative: How self-affirmation affects academic performance and motivation under identity threat. *Journal of Personality and Social Psychology.* Advance online publication.

*Siegle, D., & McCoach, D. B. (2007). Increasing student mathematics self-efficacy through teacher training. *Journal of Advanced Academics, 18*, 278-312.

Silva, E., & White, T. (2013). *Pathways to improvement: Using psychological strategies to help college students master developmental math*. Carnegie Foundation for the Advancement of Teaching: Stanford, CA.

*Silva, J. P., White, G. P., & Yoshida, R. K. (2011). The direct effects of principal–student discussions on eighth grade students' gains in reading achievement: An experimental study. *Educational Administration Quarterly*, *47*, 772-793.

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 21, pp. 261-302). San Diego, CA: Academic Press.

Stipek, D. (2002). *Motivation to learn: Integrating theory and practice* (4th edition). Boston: Allyn & Bacon.

*Struthers, C. W., & Perry, R. P. (1996). Attributional style, attributional retraining, and inoculation against motivational deficits. *Social Psychology of Education*, *1*, 171-187.

*Sverdlik, A., & Hall, N. C. (2013, August). *"Be realistic": A mixed method analysis of a motivational intervention.* Poster presented at the meeting of the American Psychological Association, Honolulu, Hawaii. *withdrawn

Tunnell, G. B. (1977). Three dimensions of naturalness: an expanded definition of field research. *Psychological Bulletin, 84*, 426-437.

*Vansteenkiste, M., Simons, J., Lens, W., Soenens, B., & Matos, L. (2005). Examining the motivational impact of intrinsic versus extrinsic goal framing and autonomy-supportive versus internally controlling communication style on early adolescents' academic achievement. *Child Development*, *76*, 483-501.

*Vansteenkiste, M., Simons, J., Lens, W., Sheldon, K. M., & Deci, E. L. (2004). Motivating learning, performance, and persistence: The synergistic effects of intrinsic goal contents and autonomy-supportive contexts. *Journal of Personality and Social Psychology*, *87*, 246-260.

*Vansteenkiste, M., Simons, J., Lens, W., Soenens, B., Matos, L., & Lacante, M. (2004). Less is sometimes more: Goal content matters. *Journal of Educational Psychology*, *96*, 755-764.

*Vansteenkiste, M., Timmermans, T., Lens, W., Soenens, B., & Van den Broeck, A. (2008). Does extrinsic goal framing enhance extrinsic goal-oriented individuals' learning and performance? An experimental test of the match perspective versus self-determination theory. *Journal of Educational Psychology*, *100*, 387-397.

Walton, G. M. (in press). The new science of wise psychological interventions. *Current Directions in Psychological Science*.

* Walton, G. M., & Cohen, G. L. (2007). A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology*, *92*, 82-96.

* Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science*, *331*, 1447-1451.

Washburn, I. J., Acock, A., Vuchinich, S., Snyder, F., Li, K.-K., Ji, P., Flay, B. R. (2011). Effects of a social-emotional and character development program on the trajectory of behaviors associated with social-emotional and character development: Findings from three randomized trials. *Prevention Science, 12*, 314-323.

Weiner, B. (1980). The role of affect in rational (attributional) approaches to human motivation. *Educational Researcher, 9*, 4-11.

Weissberg, R. P., Goren, P., Domitrovich, C., & Dusenbury, L. (2013). *2013 CASEL guide: Effective social and emotional learning programs, Preschool and elementary school edition*. Austin, TX: Collaborative for Academic, Social, and Emotional Learning.

Wentzel, K. R., & Wigfield, A. (2007). Motivational interventions that work: Themes and remaining issues. *Educational Psychologist*, *42*, 261-271.

Wigfield, A., & Eccles, J. (2002). *Development of achievement motivation.* San Diego, CA: Academic Press.

*Wilson, T. D., & Linville, P. W. (1982). Improving the academic performance of college freshmen: Attribution therapy revisited. *Journal of Personality and Social Psychology*, *42*, 367-376.

* Wilson, T. D., & Linville, P. W. (1985). Improving the performance of college freshmen with attributional techniques. *Journal of Personality and Social Psychology*, *49*, 287-293.

*Yeager, D. S., Johnson, R., Spitzer, B. J., Trzesniewski, K. H., Powers, J., Dweck, C. S. (2014). The far-reaching effects of believing people can change: Implicit theories of personality shape stress, health, and achievement during adolescence. *Journal of Personality and Social Psychology.* Advanced online publication. doi:10.1037/a0036335.

*Yeager, D. S., Miu, A. S., Powers, J., & Dweck, C. S. (2013). Implicit theories of personality and attributions of hostile intent: A meta-analysis, an experiment, and a longitudinal intervention. *Child Development,* 1-17.

*Yeager, D. S., Purdie-Vaughns, V., Garcia, J., Apfel, N., Brzustoski, P., Master, A., & Cohen, G. L. (2013). Breaking the cycle of mistrust: Wise interventions to provide critical feedback across the racial divide. *Journal of Experimental Psychology: General*.

Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research, 81*, 267-301.

*Yeager, D. S., Walton, G. M., Brady, S. T., Akcinar, E. N., Paunesku, D., Keane, L., . . . Dweck, C. S. (2014). *Brief psychological interventions can reduce post-secondary achievement gaps at scale.* Manuscript submitted for publication.

Table 1

Summary Table of Motivation Intervention Studies

Study	Theory	Avg. *d*	n_e, n_c [a]	Grade[b]	DV Type[c]	Exp. Design	Naturalness[d]
Fordyce (1983)							
Study 1	Ach. Emotions	0.36	64, 39	PS	SR	Quasi	S, I
Study 2	Ach. Emotions	0.34	50, 21	PS	SR	Quasi	S, I
Study 3	Ach. Emotions	0.68	44, 13	PS	SR	Quasi	S, I
Fordyce (1977)							
Study 1	Ach. Emotions	0.52	68, 27	PS	SR	Quasi	S, I
Study 2	Ach. Emotions	0.70	39, 29	PS	SR	Randomized	S, I
Study 3	Ach. Emotions	1.01	42, 26	PS	SR	Quasi	S, I
Ramirez & Beilock (2011)	Ach. Emotions	0.57	27, 26	HS	P	Randomized	S, I, DV
Hoyert & O'Dell (2006)	Ach. Goal Theory	0.93	69, 68	PS	SR, P	Randomized	S, I, DV
Muis et al. (2013)	Ach. Goal Theory	0.12	198, 52	PS	SR, P	Randomized	S, I, DV
Ranellucci et al. (unpublished)	Ach. Goal Theory	-0.04	135, 42	PS	SR, P	Randomized	I, DV
Boese et al. (2013)	Attribution	0.77	84, 42	PS	SR, P	Randomized	S, I, DV
Good et al. (2003)							
Attribution vs. Control	Attribution	1.11	34, 35	MS	P	Randomized	S, I, DV
Hall et al. (2007)	Attribution	0.28	374, 375	PS	SR, P	Quasi	S, I, DV
Hall et al. (2004)	Attribution	0.43	101, 102	PS	SR, P	Quasi	S, DV
Ruthig et al. (2004)	Attribution	0.59	118, 118	PS	SR, P, B	Quasi	S, I, DV
Struthers & Perry (1996)	Attribution	0.41	108, 150	PS	SR, P	Randomized	S, I, DV
Wilson & Linville (1985)							
Replication 1	Attribution	0.25	20, 20	PS	SR, P	Randomized	I, DV
Replication 2	Attribution	0.13	20, 20	PS	SR, P	Randomized	I, DV
Wilson & Linville (1982)	Attribution	0.73	20, 20	PS	SR, P, B	Randomized	S, I, DV
Yeager et al. (2013)							
Study 1	Attribution	0.76	22, 22	M	SR, B	Randomized	S, I, DV
Study 2	Attribution	0.78	22, 22	M	SR, P, B	Randomized	S, I, DV
Study 3	Attribution	0.44	38, 38	HS	P	Randomized	S, I, DV
Durik et al. (unpublished)	Expectancy-Value	0.10	158, 157	MS	SR, P	Randomized	S, I, DV

Study	Theory	Avg. d	n_e, n_c [a]	Grade[b]	DV Type[c]	Exp. Design	Naturalness[d]
Harackiewicz et al. (2012)	Expectancy-Value	0.32	94, 94	HS	SR, B	Randomized	S, I, DV
Hulleman et al. (2010), Study 2	Expectancy-Value	0.38	160, 158	PS	SR, P	Randomized	S, I, DV
Hulleman & Harackiewicz (2009)	Expectancy-Value	0.27	136, 126	HS	SR, P	Randomized	S, I, DV
Silva et al. (2011)	Goal Setting	0.71	20, 21	HS	P	Randomized	S, I, DV
Sverdlik & Hall (unpublished)	Goal Setting	0.67	25, 27	PS	P	Randomized	DV
Aronson et al. (2002)	Impl. Theories of Int.	0.53	37, 37	PS	SR, P	Randomized	DV
Blackwell et al. (2007)	Impl. Theories of Int.	0.69	49, 50	MS	SR, P, B	Randomized	S, I, DV
Good et al. (2003)							
Implicit Theories vs. Control	Impl. Theories of Int.	0.76	34, 35	MS	P	Randomized	S, I, DV
Yeager et al. (2014)							
Study 2	Impl. Theories of Int.	0.43	39, 39	HS	SR, P, B	Randomized	S, I, DV
Study 3	Impl. Theories of Int.	0.36	75, 75	HS	SR, P, B	Randomized	S, I, DV
Yeager et al. (2013)							
Study 3	Impl. Theories of Int.	0.65	39, 39	HS	SR	Randomized	S, I
Guthrie et al. (2006)	Interest	0.71	49, 49	ES	P, B	Quasi	SI, I, DV
Hidi et al. (2002)	Interest	0.67	90, 90	MS	SR, P	Quasi	SI, I, DV
Acee & Weinstein	Multiple Perspectives	0.63	41, 41	PS	SR, P, B	Quasi	SI, DV
Bernacki, et al. (2014)	Multiple Perspectives	0.54	26, 27	MS	SR	Randomized	SI, I
Bordine (unpublished)	Multiple Perspectives	0.49	16, 15	ES	SR, P	Randomized	S, I, DV
Craven et al. (1991)	Multiple Perspectives	0.08	81, 79	ES	SR	Randomized	S, I, DV
Cueva (unpublished)	Multiple Perspectives	0.18	33, 31	ES	SR, P	Quasi	S, I, DV
Duckworth et al. (in press)	Multiple Perspectives	0.51	38, 39	ES	P, B	Randomized	S, I, DV
Good et al. (2003)							
Combined vs. Control	Multiple Perspectives	0.79	34, 35	MS	P	Randomized	S, I, DV
Hong & Lin-Siegler (2011)	Multiple Perspectives	0.41	88, 93	HS	SR, P	Randomized	S, I, DV
Jamieson et al. (2010)	Multiple Perspectives	0.87	14, 14	PS	SR, P, B	Randomized	I, DV
Kitsantas et al. (2004)	Multiple Perspectives	1.14	48, 48	HS	SR, P	Randomized	S, I, DV
Martin (2008)	Multiple Perspectives	0.48	26, 27	HS	SR	Quasi	S, I
McGinley & Jones (2014)	Multiple Perspectives	0.37	58, 53	PS	SR	Randomized	S, I, DV
Morisano et al. (2010)	Multiple Perspectives	0.75	43, 42	PS	SR, P, B	Randomized	S, I, DV
Paunesku et al. (unpublished)	Multiple Perspectives	0.14	1196, 398	HS	SR, P	Randomized	S, DV

Study	Theory	Avg. d	n_e, n_c[a]	Grade[b]	DV Type[c]	Exp. Design	Naturalness[d]
Siegle & McCoach (2007)	Multiple Perspectives	0.33	430, 442	ES	SR, P	Randomized	S, I, DV
Yeager et al. (unpublished)							
Study 1	Multiple Perspectives	0.24	310, 274	PS	B	Randomized	S, I, DV
Study 2	Multiple Perspectives	0.13	5281, 16278	PS	B	Randomized	S, I, DV
Study 3	Multiple Perspectives	0.23	1186, 406	PS	P	Randomized	S, I, DV
Quintenilla (unpublished)	Need for Ach.	0.36	47, 46	PS	SR	Randomized	S, I
Day et al. (1994)	Possible Selves	0.91	42, 41	ES	SR	Randomized	S
Oyserman et al. (2006)	Possible Selves	0.35	141, 123	HS	SR, P, B	Randomized	S, I, DV
Oyserman et al. (2002)	Possible Selves	0.37	62, 146	MS	SR, B	Quasi	S, I, DV
Cohen et al. (2009)	Self-Affirmation	0.52	192, 193	MS	SR, P	Randomized	S, I, DV
Cohen et al. (2006)	Self-Affirmation	0.37	104, 104	MS	SR, P	Randomized	S, I, DV
Cook et al. (2012), Study 2	Self-Affirmation	0.36	61, 60	MS	SR, P	Randomized	S, DV
Miyake et al (2010)							
Male Comparison	Self-Affirmation	-0.16	178, 105	PS	P	Randomized	S, I, DV
Female Comparison	Self-Affirmation	0.21	69, 47	PS	P	Randomized	S, I, DV
Sherman et al. (2013)							
Study 1	Self-Affirmation	0.34	41, 40	MS	P	Randomized	S, DV
Study 2	Self-Affirmation	0.64	26, 29	MS	SR, P	Randomized	S, DV
Walton & Cohen (2011)	Self-Affirmation	0.52	49, 43	PS	SR, P	Randomized	DV
Greenstein (1976)	Self-Confrontation	0.54	87, 84	PS	SR, B	Randomized	S, DV
Froiland (2011)	Self-Determination	0.71	15, 15	ES	SR	Quasi	S, I
Patall et al. (2010)	Self-Determination	0.12	193, 194	HS	SR, P, B	Randomized	S, I, DV
Radil (unpublished)	Self-Determination	0.43	25, 37	PS	SR	Quasi	S, I, DV
Reeve et al. (2004)	Self-Determination	1.94[e]	10, 10	HS	B	Randomized	S, I, DV
Schaffner & Schiefele (2007)	Self-Determination	0.46	188, 187	HS	SR, P	Randomized	S, I
Vansteenkiste et al. (2008)	Self-Determination	0.70	68, 70	MS	SR, P, B	Randomized	S, I, DV
Vansteenkiste et al. (2005)							
Study 1	Self-Determination	0.83	65, 65	MS	SR, P	Randomized	S, I, DV
Study 3	Self-Determination	0.74	57, 56	MS	SR, P	Randomized	S, I, DV
Vansteenkiste et al. (2004)							

Study	Theory	Avg. *d*	n_e, n_c[a]	Grade[b]	DV Type[c]	Exp. Design	Naturalness[d]
Study 1	Self-Determination	1.57[e]	100, 100	PS	SR, P, B	Quasi	S, I, DV
Study 2	Self-Determination	1.49[e]	189, 189	PS	SR, P, B	Quasi	S, I, DV
Vansteenkiste et al. (2004)	Self-Determination	0.42	123, 122	PS	SR, P, B	Randomized	S, I, DV
Gehlbach et al. (unpublished)	Social Belongingness	0.15	194, 60	HS	SR, P, B	Randomized	S, DV
Hausmann et al. (2009)							
Exp. vs. Control (White)	Social Belongingness	0.26	70, 67	PS	SR	Randomized	S, I, DV
Exp. vs. Control (Afr. Amer)	Social Belongingness	-0.04	41, 42	PS	SR	Randomized	S, I, DV
Walton & Cohen (2007)							
Study 1	Social Belongingness	0.91	18, 18	PS	SR, B	Randomized	I, DV
Study 2	Social Belongingness	1.57[e]	18, 18	PS	SR, P	Randomized	I
Pugh (unpublished)	Transformative Exp.	0.67	76, 82	MS	SR, P	Randomized	S, I, DV
Total		0.49[f]	14200, 23039				

Note: Ach. Emotions = Achievement Emotions; Impl. Theories of Int. = Implicit Theories of Intelligence.
[a]The sample size for the experimental condition (n_e) is reported first, followed by the sample size for the control condition (n_c).
[b]Grade included Elementary School (ES), Middle School (MS), High School (HS), and Post-Secondary (PS).
[c]Types of dependent variables included Self-Report (SR), Behavioral Indicator (B), and Performance Indicator (P).
[d]Types of naturalness included Setting (S), Intervention (I), and Dependent Variable (DV).
[e]Extreme outliers were Windsorized and adjusted to 3 standard deviations from the effect size mean.
[f]Mean Effect Size calculated via macro (meanes.sps) provided by Lipsey and Wilson (2001).

Table 2

Theoretical Frameworks of Student Motivation

Theory/Framework	Description	Overview(s)
Achievement Emotions	Emotional experiences in school emanate from students' perception of control and value for academics	Pekrun (2006)
Anxiety	Worrying about the consequences of performance, which undermines working memory and outcomes	Ramirez & Beilock (2011)
Happiness	An overriding emotional sense of wellbeing	Fordyce (1977)
Achievement Goal	Students' goals for engaging in an activity shape how they approach, experience, and react to achievement situations	Elliot (2005) Kaplan and Maehr (2007)
Attribution	Students' explanations for success or failure influence subsequent achievement behavior	Weiner (1980)
Expectancy-Value	Student motivation is determined most proximally by success expectancies and perceived task value	Eccles et al. (1983)
Goal Setting	Specific, difficult task goals produce higher commitment and performance than vague goals that are easy to attain	Locke and Latham (1990; 2002)
Implicit Theories of Intelligence	Students' beliefs about whether intelligence is fixed (i.e., entity mindset) or is malleable (i.e., incremental mindset) influence goal striving, persistence, and performance	Dweck (1986, 1999)
Interest	The development and deepening of interest in specific topics and academics is influenced by situational and individual difference factors	Hidi and Renninger (2006)
Need for Achievement	The importance of achieving mastery, high achievement, and out-doing others to reach one's full potential	McClelland et al. (1976)
Possible Selves	Students' conception of what they might become (both desired and feared) serve as incentives for future behavior and a way to evaluate current behavior	Markus and Nurius (1986)
Self-Affirmation	Students' who perceive that they are in danger of confirming a stereotype about their group experience increased anxiety and reductions in performance	Steele (1988)
Self-Confrontation	Students' perception that their behaviors and values differ from their self-conception motivates change	Rokeach (1973)
Self-Determination	Satisfying students' three core needs (autonomy, relatedness, competence) are essential for promoting motivation and well-being	Deci and Ryan (1985)
Social Belongingness	The degree to which students perceive they belong and are connected to others can influence their learning outcomes	Baumeister & Leary (1995)
Transformative Experience	Reframing the learning experience as an application of the content in a way that enhances everyday value	Pugh (2011)

Table 3

Research-based Sources of Expectancy-Related Beliefs

Expectancy Source	**Definition**
Perceptions of ability/skill	When students perceive they have a high level of ability and/or skill at an activity, they are more likely to experience high expectancy (Bandura, 1997; Wigfield & Eccles, 2002)
Effort Attributions	When students believe that their effort will lead to learning, they are more likely to experience high expectancy (Dweck & Leggett, 1988; Dweck 1999; Weiner, 1974)
Success Experiences	When students are successful at an activity, or watch others have success, they are more likely to experience high expectancy (Bandura, 1997; Eccles et al., 1983)
Support and Scaffolding	When students are appropriately supported in completing an activity (e.g., through encouragement and having the resources necessary to complete the task), they are more likely to experience high expectancy (Bandura, 1997)
Clear expectations	When students know what is expected of them on an activity, and have clearly defined goals, they are more likely to experience high expectancy (Pajares, 1996)
Appropriate challenge	When the difficulty of the task or activity matches students' skill levels, they are more likely to experience high expectancy (Eccles et al., 1983)
Feedback	When students receive feedback that effort matters, skills are amenable to change, and are task-focused (rather than ability-focused), they are more likely to experience high expectancy (Dweck & Leggett, 1988; Dweck, 1999)
Growth experiences	When students engage in learning activities that challenge them to grow and learn, experience growth in their skills and performance improvements, they are more likely to experience both high expectancy and value (Dweck & Leggett, 1988; Dweck, 1999; Hong et al., 1999)
Perceptions of others' expectations	When students perceive their parents and teachers have high or low expectancies, their own expectancies are shaped accordingly; for instance, if teachers have high expectations for their students, these students in turn develop high expectancies (Bandura, 1997; Dweck & Leggett, 1988; Dweck 1999; Eccles et al., 1983)
Perceived task difficulty	When students perceive a subject or task as being not difficult, they develop higher estimates of their own abilities for the subject or task (Bandura, 1997; Pajares, 1996; Wigfield & Eccles, 2002)
Stability attributions	When students attribute success to a stable factor (ability), then they will have higher expectations for future success; if they attribute it to an unstable factor (good luck), they will be uncertain about future success and have lower expectations for future success (Weiner, 2010)

Table 4

Research-Based Sources of Value

Value Source	**Definition**
Intrinsic Benefits	When students find the activities and academic content enjoyable and interesting, they are more likely to experience high value (Renninger & Hidi, 2011)
Relevance	When students are able to connect what they are learning to their personal lives and/or the real world, they are more likely to experience high value (Hulleman & Harackiewicz, 2009)
Context and Rationale	When students understand that an activity is meaningful and has a purpose, they are more likely to experience high value (Lepper & Henderlong, 2000)
Variety and Novelty	When students engage in activities that are varied and novel, they are more likely to experience high value (e.g., catch and hold interest; Berlyne; Hidi & Renninger, 2006; Kang et al., 2009)
Enthusiastic Models	When students interact with teachers and other adults who are enthusiastic and passionate about learning, they are more likely to experience high value (Patrick, Hisley, & Kempler, 2000)
Choice and Control	When students feel a sense of control and choice over their learning, they are more likely to experience high value (Reeve, 2009)
Positive Relationships and Sense of Belongingness	When students experience meaningful student-student and student-teacher relationships, they are more likely to experience high value (Furrer & Skinner, 2003; Walton & Cohen, 2007)
Extrinsic Benefits	When students receive external rewards and incentives for learning (e.g., prizes, food), they are more likely to experience high value to complete an activity but low value to produce quality work (Marinak & Gambrell, 2008)

Table 5

Research-Based Sources of Cost

Cost Source	**Definition**
Effort and Time Needed for the Activity	When students feel that the workload is unreasonable (e.g., 3 hours/night) and/or unnecessary (e.g., busy work), they are more likely to experience increased cost (Parsons et al., 1980; Perez et al., 2014)
Effort and Time Needed for Other Competing Activities	When student have too many other demands on their time or do not know how to effectively manage their time, they are more likely to experience high cost (Barron & Hulleman, in press; Flake, 2012)
Loss of Valued Alternatives	When students feel like the learning activity is not worth their time compared to other things they might do (e.g., socializing), they are more likely to experience high cost (Conley, 2012; Perez et al., 2014)
Psychological and Physical Reactions to the Activity	When students feel unsafe and uncomfortable, either physically or psychologically (e.g., nervous, bored, tired), they are more likely to experience high cost (Eccles et al., 1983; Ramirez & Beilock, 2011)

Table 6

Expectancy, Value, and Cost Interventions by Theory and (Targeted Sources)

Primarily Expectancy Interventions	**Primarily Value Interventions**	**Expectancy and Value Interventions**	**Primarily Cost Interventions**
Attribution Theory (effort attributions)	Expectancy-value Framework (relevance, utility value)	Achievement Goal Theory (expectancies: growth experiences, effort attributions, feedback) and (values: context and rationale)	Social Belongingness (psychological reactions)
			Achievement Emotions (physical reactions)
Implicit Theories of Intelligence (feedback and effort attributions)	Self-Determination Theory (autonomy, intrinsic benefits)	Possible Selves Theory (expectancies: perceptions of ability and skill, success experiences) and (values: context and rationale)	
Self-Confrontation Theory (feedback)	Self-Affirmation Theory (self-affirmation)		
	Interest Theory (variety and novelty)	Need for Achievement (expectancies: support and scaffolding, clear expectations, feedback) and (values: relevance, choice and control, context and rationale)	
	Transformative Experience (relevance, context and rationale)		