Meta-Analysis in Higher Education: An Illustrative Example Using Hierarchical Linear Modeling

Nida Denson · Michael H. Seltzer

Received: 2 March 2010 / Published online: 16 November 2010

© The Author(s) 2010. This article is published with open access at Springerlink.com

Abstract The purpose of this article is to provide higher education researchers with an illustrative example of meta-analysis utilizing hierarchical linear modeling (HLM). This article demonstrates the step-by-step process of meta-analysis using a recently-published study examining the effects of curricular and co-curricular diversity activities on racial bias in college students as an example (Denson, *Rev Educ Res* 79:805–838, 2009). The authors present an overview of the meta-analytic approach and describe a meta-analysis from beginning to end. The example includes: problem specification; research questions; study retrieval and selection; coding procedure; calculating effect sizes; visual displays and summary statistics; conducting HLM analyses; and sensitivity analyses. The authors also offer guidelines and recommendations for improving the conduct and reporting of research which in turn can provide the information necessary for future and more comprehensive meta-analytic reviews.

Keywords Meta-analysis · Effect size · Random-effects model · Hierarchical linear modeling · Moderator analysis

While meta-analytic techniques have been available for over a century (see Cooper and Hedges (1994) for a history of meta-analysis), meta-analysis as a method of inquiry for integrating quantitative results from a stream of research began to become popular approximately 30 years ago (Glass 2000). While conducting a meta-analysis can be very time consuming when done properly, advances in computing and statistical software have facilitated conducting literature searches and applying meta-analytic procedures. The purpose of this article is to provide higher education researchers with a pedagogical example of a meta-analysis utilizing hierarchical linear modeling (HLM) to encourage

N. Denson (✉)

Centre for Educational Research, University of Western Sydney, Building 19, Bankstown Campus
Locked Bag 1797, Penrith South DC, NSW 1797, Australia
e-mail: n.denson@uws.edu.au

M. H. Seltzer

University of California, Los Angeles, Los Angeles, CA, USA

more widespread use of meta-analytic methods. In this article, we begin by first providing an overview of the meta-analytic approach. We then provide an example of meta-analysis using Denson's *Review of Educational Research* (2009) study of the effects of curricular and co-curricular diversity activities on racial bias in college students. Finally, we provide guidelines and recommendations for improving the conduct and reporting of future research.

Overview of the Meta-Analytic Approach

Put simply, meta-analysis is a *statistical* technique of combining the findings of a set of studies that address common research hypotheses (Cooper and Hedges 1994; Lipsey and Wilson 2001). As Glass (1976) defines it, “meta-analysis refers to the analysis of analyses...the statistical analysis of a large collection of analyses results from individual studies for the purpose of integrating the findings” (p. 3). Conducting a meta-analysis is much like conducting a primary research study. In a meta-analysis, instead of students or participants being the unit of analysis, the primary studies themselves become the unit of analysis. In meta-analysis, the researcher collects research studies from a particular domain (e.g., studies of the effects of curricular diversity activities), codes them, and synthesizes results from them using statistical methods analogous to those used in primary data analysis (e.g., techniques analogous to regression). As will be discussed in detail below, the key quantitative results of each study (e.g., comparisons of those students who participate in a diversity program of interest versus those who do not) are converted to effect sizes. The effect sizes for a sample of studies become the dependent variable, and various study characteristics (e.g., variables that capture the demographic characteristics of each study's sample; variables that capture key features of each study's program) become the independent variables. In this way, a meta-analysis provides an integrated review of the research base that enables us to calculate an overall summary of the effects of a program or policy of interest, and investigate various hypotheses concerning why the effects may be larger in some studies and smaller in others; that is, we can attempt to identify key factors that might magnify or dampen the effects of a program of interest.

Many extremely thoughtful and thorough qualitative research syntheses have been undertaken in education (e.g., Feldman and Newcomb 1969; Pascarella and Terenzini 1991, 2005). Meta-analysis complements qualitative literature reviews by integrating the findings in a systematic *quantitative* way, which we will discuss in detail below. Some examples of meta-analyses in education include: coaching effects on SAT scores (Messick and Jungeblut 1981), teacher expectancy effects on student IQ (Raudenbush 1984), school desegregation effects on African-American students (Wells and Crain 1994), and effects of school reform programs on achievement (Borman et al. 2003).

Thus, meta-analysis can help answer three basic questions:

- (1) *What is the magnitude of the overall effect size for a program or practice of interest?*
- (2) *To what extent do the effect sizes of the studies in a sample vary around the overall effect size estimate?*
- (3) *What are the reasons for the observed differences in effect sizes across studies?* The presence of heterogeneity provides an opportunity to ask additional questions, and to investigate more closely the reasons for the observed differences in effect size across studies. Such investigations have the potential to generate further insight into what

conditions and for whom a program or practice of interest may be particularly beneficial.

The next section illustrates the step-by-step process of conducting a meta-analysis by using Denson's *Review of Educational Research* (2009) study as an example, which focuses on the effects of curricular and co-curricular diversity activities on racial bias in college students.

An Illustrative Meta-Analysis Example

Problem Specification

In the last 50 years, the rapidly changing demographics of the United States have seen a parallel shift in the demographic makeup of our college campuses. Naturally, this change has been met with some resistance and racial tension. As a result, institutions have implemented diversity-related initiatives designed to promote positive intergroup relations. While there certainly have been many *qualitative* literature reviews on this topic (e.g., Engberg 2004), there had yet to be a *quantitative* review that summarized and integrated the findings of this topic thus far. Denson's (2009) study was the first to address this research gap.

Research Questions

Thus, Denson's (2009) study sought to answer the following general research questions:

- (1) What is the *magnitude* of the overall effect size of participation in curricular and co-curricular diversity activities on racial bias? (overall effect)
- (2) Is there *variation* in the effect of curricular and co-curricular diversity activities on racial bias? (heterogeneity in effect sizes)
- (3) Which types of *programs* are most effective? Which types of *studies* show the most effective programs? (factors that amplify or diminish the magnitude of effect sizes)

Study Retrieval and Selection

One of the first steps to conducting a meta-analysis is study retrieval and selection (Cooper 1982). This is an important step as a meta-analysis is only useful if it provides an accurate representation of the research at large. So, a meta-analysis is only as good as the studies it represents. Thus, it is imperative to retrieve every relevant study possible. In addition to published research, a thorough search should also include unpublished studies such as dissertations and conference papers. The keywords and descriptors used as search terms here were: diversity, ethnic studies, women studies, bias, prejudice, stereotypes, discrimination, higher education, and college students. Specifically, the search was implemented as follows:

KEYWORDS = (*diversity or ethnic studies or women studies*)
 AND
 KEYWORDS = (*bias or prejudice or stereotypes or discrimination*)
 AND
 DESCRIPTORS = (*higher education or college students*)

Table 1 Characteristics of the Sample of Studies ($N = 27$) (Copyright 2009 American Educational Research Association. Used by permission)

Category	Number of studies	Percentage
Source of the studies		
Journal article	17	63
Book or book chapter	3	11
Conference paper	3	11
Doctoral dissertation	3	11
Article under review	1	4
Year of publication		
1977–1999	15	56
2000–2006	12	44
Single/multiple institution		
Single institution	13	48
Multiple institution	13	48
Both	1	4
Type of intervention		
Diversity course	8	30
Ethnic studies course	10	37
Women's studies course	6	22
Diversity workshop/training	15	56
Peer facilitated training	5	19
Combination	3	11

Note: Type of intervention does not add up to 27 because many studies examined multiple interventions

In addition, the search was limited to sources written in English only, as well as to only studies based on students from four-year colleges or universities within the United States. These descriptors facilitated a computer-based search of three databases: Education Resources Information Center (ERIC) (1966–2006), PsycINFO (1840–2006), and ProQuest Digital Dissertations (including Dissertation Abstracts International) (1861–2006) through June 2006. In addition, a hand search was conducted over the last 10 years of the following top higher education journals: *Journal of Higher Education*, *Review of Higher Education*, *Research in Higher Education*, *Journal of Social Issues*, and the *Journal of College Student Development*. Finally, references were checked in each of the included studies as well as review articles for other relevant works.

As in all meta-analyses, the present study required specific inclusion rules. A study was included in the meta-analysis if it: (a) investigated the relationship between curricular and/or co-curricular diversity and a racial bias outcome and (b) reported quantitative data in sufficient detail for calculation of a standardized mean difference effect size. These selection criteria resulted in a sample of 27 primary studies examining the relationship between curricular and/or co-curricular diversity activities and racial bias (see Appendix A in Denson 2009). Table 1 lists the characteristics of the sample of coded studies and presents a brief overview of the distribution of studies.

Coding Procedure

The next step is to code these studies for their study characteristics and effect sizes which will become the independent variables and dependent variable, respectively (Lipsey and Wilson 2001). The study characteristics are possible moderators which may help explain

the variation seen in effect sizes across studies. It is likely that there are systematic differences across studies that may account for differences in effect sizes. Effect size estimates are standardized measures of an effect and comprise the key outcome measure in meta-analysis and enable the synthesizing and integrating of findings.

Study Characteristics

The independent variables selected and coded for this meta-analysis were included as possible covariates, and were categorized into three types: study characteristics, student characteristics, and institutional characteristics. The study characteristics consisted of three subgroups: study source characteristics (a through c), diversity initiative characteristics (d through i), and methodological characteristics (j through q).

The study source characteristics included: (a) author identification, (b) year of publication, and (c) type of publication (journal article, book, book chapter, conference paper, doctoral dissertation and other unpublished work). The diversity initiative characteristics included: (d) type of curricular or co-curricular diversity activity (required diversity course, non-required diversity course, ethnic studies, women's studies, diversity workshop or training intervention, peer-facilitated intervention), (e) pedagogical focus (enlightenment only, enlightenment and contact), (f) intensity of cross-racial interaction (random, built-in), (g) duration of intervention (in weeks), (h) intensity of intervention (in hours/week), and (i) year of intervention. The methodological characteristics included: (j) outcome measures (k) reliability of outcome, (l) measurement technology of outcome (homegrown, well-established), (m) target group (general, specific), (n) quality of covariate adjustment (none, one/pretest, additional/multiple), (o) sample size, (p) sample assignment (not random, random), and (q) sample matching (no, yes).

The student characteristics included: (a) age, (b) gender, and (c) ethnicity. Finally, the institutional characteristics included: (a) number of institutions in study (single, multiple), (b) institution name, (c) institutional diversity requirement (no, yes), (d) type of institution (public, private), (e) size of institution (total number of full-time undergraduates), (f) structural diversity (percentage of students of color), and (g) region of country (Western, Southern, Eastern, Midwest). All of the primary studies were coded according to these variables.

In addition, because coding might be subject to human error, two coders were involved in the coding process and inter-rater reliability indices were calculated accordingly. Specifically, Cohen's (1960) kappa was computed for the categorical variables and Cronbach's (1951) alpha was computed for the continuous variables. The inter-rater reliability indices were high, with the Cohen's kappa statistics ranging from 0.80 to 1.00 and the Cronbach's alpha coefficients ranging from 0.95 to 0.99.

Calculating Effect Sizes

The dependent variable was effect size and was estimated as a group contrast in this meta-analysis. A group contrast compares an outcome variable across two or more groups of respondents. For example, the racial bias measures can be compared between students who had taken a required diversity course versus those who had not. The group contrasts were represented by the *standardized mean difference* because the operationalization of racial bias differed to some extent across the studies, and it is thus necessary to standardize the group contrasts so that the values can be meaningfully compared across the sample of studies.

The standardized mean difference effect size statistic for a given study is defined as follows:

$$d = ES_{sm} = \frac{\bar{X}_{G1} - \bar{X}_{G2}}{S_p} \quad (1)$$

where $\bar{X}_{G1}$ is the mean for Group 1 (e.g., the treatment group), $\bar{X}_{G2}$ is the mean for Group 2 (e.g., comparison group), n_{G1} is the number of subjects in Group 1, n_{G2} is the number of subjects in Group 2, and S_p is the pooled standard deviation, defined as:

$$S_p = \sqrt{\frac{(n_{G1} - 1)S_{G1}^2 + (n_{G2} - 1)S_{G2}^2}{(n_{G1} - 1) + (n_{G2} - 1)}} \quad (2)$$

Where S_{G1} is the standard deviation for Group 1 and S_{G2} is the standard deviation for Group 2. In the case of small sample sizes, however, this effect size index has been shown to be upwardly biased, especially for sample sizes less than 20 (Hedges 1981). As a result, a simple correction was applied and all subsequent computations used this corrected (unbiased) effect size estimate as follows (Hedges 1981):

$$g = \left[1 - \frac{3}{4N - 9} \right] ES_{sm}, \quad (3)$$

where N is the total sample size ($n_{G1} + n_{G2}$), and the standard error of the Hedges-corrected effect size estimate is:

$$SE(g) = \sqrt{\frac{n_{G1} + n_{G2}}{n_{G1}n_{G2}}} + \frac{(ES'_{sm})^2}{2(n_{G1} + n_{G2})}, \quad (4)$$

Thus, the effect size estimate (g) and its standard error $SE(g)$ were computed for each of the studies (see Lipsey and Wilson 2001, for a more detailed explanation).

The studies in the sample for this study were essentially quasi-experimental in that students selected into the various diversity programs, and in nearly all cases the authors of these studies attempted to statistically adjust for any pre-program differences in initial bias. When possible, *adjusted* effect sizes were calculated from inferential statistics (e.g., ANCOVA, multiple regression) reported in these articles (for details concerning computing adjusted effect sizes and their error variances, see Borenstein 2009). In these cases, the effect sizes can be viewed as being *adjusted* for differences in covariates (e.g., a pretest). When insufficient information was reported such that it was not possible to obtain the adjusted effect size, the *unadjusted* effect size was calculated from the reported descriptive statistics (e.g., means, standard deviations, correlations) using standard meta-analytic procedures (Hedges and Olkin 1985; Lipsey and Wilson 2001). Thus, whether or not a study's effect size is adjusted or not is coded, and in our analyses we will examine the extent to which the pattern of results tend to differ for the set of studies with adjusted effect sizes and those with unadjusted effect sizes.

In the cases where there was insufficient detail provided in the article to calculate an effect size(s), the author(s) were contacted to obtain the required information. Of the 27 studies eligible to be included in the meta-analysis, 11 studies were excluded due to insufficient information. These authors were contacted via email, but either they did not respond or could not provide me with the requested information. Of the remaining 16 studies, two utilized the same sample, and thus the effect sizes were averaged. If a study included multiple outcomes of the same outcome type, then the effect sizes were averaged.

On the other hand, if a study examined multiple intervention types (e.g., diversity course and ethnic studies course), then separate effect sizes were calculated for each intervention type. In total, 30 effect sizes were computed from the 16 studies, with each study contributing, on average, 1.9 effect sizes (range = 1–9) (see Table 2 for a list of included studies and their associated effect sizes).

Table 2 displays the 30 effect size estimates and their standard errors. As can be seen the effect size estimates range in value from 0.15 to 2.17, and the standard errors of the estimates range from 0.01 to 0.52. To help understand the meaning of these quantities, consider, for example, the results for the Lopez et al. (1998) study. The effect size estimate for this study indicates that the outcome scores for the students in the treatment group were on average approximately 0.87 standard deviations higher than the outcome scores for the students in the control group (with higher scores indicating greater racial understanding or less bias). According to Cohen (1988), effect size estimates around 0.2–0.3 are considered to be small, around 0.5 are moderate, and estimates greater than 0.8 are large.

Just as a treatment effect estimate provides us with an estimate of the true effect of a program or treatment, an effect size estimate (g_j) can be viewed as providing an estimate of the true effect size for a study (δ_j).¹ The standard error gives us a sense of how precise the effect size estimate is, and enables us to construct a confidence interval, which helps convey how large or small the true effect might be. Thus, for the Lopez et al. study, adding and subtracting approximately 2 standard errors (2×0.16) to the effect size estimate (0.87), we obtain a 95% interval whose lower boundary is 0.56 and whose upper boundary is 1.18. Based on this interval, the notion that the true effect might be 0 is clearly not plausible; as can be seen in Fig. 2, the interval lies well above a value of 0. Note also that the interval excludes a value of 0.48, which is the mean value of the effect size estimates in Table 2.

Next consider the effect size estimate from the Antonio (2001) study. The magnitude of the estimate is 0.30, indicating that the outcome scores for the treatment group students in this study were on average approximately a third of a standard deviation higher than the outcome scores for the control group students. Using the standard error of the estimate (i.e., 0.04) to construct a 95% interval, we see that the lower boundary of the interval is 0.22 and the upper boundary 0.38 (see Table 2, Fig. 2). Just as in the case of the Lopez et al. study, the notion that the true effect size for this study is 0 is not very plausible. But as can be seen the intervals for the Antonio and Lopez et al. studies do not overlap at all, i.e., the interval for the Antonio study lies well below the interval for the Lopez et al. study. This provides evidence that the true effect size for the Antonio study is likely appreciably smaller than the true effect size for the Lopez study.

Visual Displays and Summary Statistics

There are many types of different visual displays available to present and summarize information about magnitudes and distributions of effect sizes (Kosslyn 1994; Tufte 1983). For example, the Tukey's stem-and-leaf display is a plot of the data that describes the distribution of results and retains each of the recorded effect sizes (Rosenthal 1995; Tukey 1977). As another example, the schematic plot (i.e., the box plot) records (a) the median effect size, (b) the quartiles, and (c) the minimum and maximum effect sizes (Tukey 1977). The box represents the interquartile range which contains the middle 50% of the effect

¹ For the remainder of the article, the subscript “ j ” will be used to index or denote individual studies in the sample.

Table 2 Summary of effect sizes by study

Study	Pedagogical approach	Type of intervention	Type of sample	Type of effect size	ES estimates (g)	SE of estimates (SE(g))
Katz and Ivey (1977)	Pedagogical approach	Diversity workshop and training intervention	Whites	Unadjusted	2.17	0.52
Lopez et al. (1998) [1]	Pedagogical approach and CRI	Non-required diversity course and peer facilitated intervention	Mixed	Adjusted	0.87	0.16
Muthuswamy et al. (2006) [1]	Pedagogical approach and CRI	Peer-facilitated intervention	Mixed	Unadjusted	0.86	0.22
Hyun (1994), Milem (1994) [4]	Pedagogical approach	Diversity workshop and training intervention	Whites	Unadjusted	0.83	0.03
Antony (1993) [3]	Pedagogical approach	Diversity workshop and training intervention	Mixed	Unadjusted	0.81	0.02
Taylor (1994) [1]	Pedagogical approach	Combination	Whites	Unadjusted	0.67	0.11
MacPhee et al. (1994) [2]	Pedagogical approach	Non-required diversity course	Mixed	Unadjusted	0.64	0.09
Antony (1993) [1]	Pedagogical approach	Ethnic studies course	Mixed	Unadjusted	0.61	0.02
Hyun (1994), Milem (1994) [3]	Pedagogical approach	Ethnic studies course	Whites	Unadjusted	0.59	0.02
Hogan and Mallott (2005) [2]	Pedagogical approach	Required diversity course	Mixed	Unadjusted	0.53	0.16
Hyun (1994), Milem (1994) [2]	Pedagogical approach	Diversity workshop and training intervention	Students of color	Unadjusted	0.52	0.10
Antony (1993) [2]	Pedagogical approach	Women's studies course	Mixed	Unadjusted	0.48	0.02
Gurin et al. (2004) [1]	Pedagogical approach and CRI	Non-required diversity course and peer facilitated intervention	Mixed	Adjusted	0.48	0.18
Hyun (1994), Milem (1994) [1]	Pedagogical approach	Ethnic studies course	Students of color	Unadjusted	0.46	0.10
Gurin et al. (2002) [1]	Pedagogical approach	Ethnic studies course	Whites	Unadjusted	0.44	0.02
Vogelgesang (2000) [3]	Pedagogical approach	Diversity workshop and training intervention	Mixed	Unadjusted	0.43	0.02
Vogelgesang (2000) [6]	Pedagogical approach	Diversity workshop and training intervention	Students of color	Unadjusted	0.42	0.08

Table 2 continued

Study	Pedagogical approach	Type of intervention	Type of sample	Type of effect size	ES estimates (g)	SE of estimates (SE(g))
Vogelgesang (2000) [9]	Pedagogical approach	Diversity workshop and training intervention	Whites	Unadjusted	0.41	0.02
Vogelgesang (2000) [5]	Pedagogical approach	Women's studies course	Students of color	Unadjusted	0.40	0.08
Stake and Hoffman (2001) [3]	Pedagogical approach	Women's studies course	Mixed	Adjusted	0.39	0.09
Chang (2002)	Pedagogical approach	Required diversity course	Mixed	Adjusted	0.38	0.15
Stake and Hoffman (2001) [2]	Pedagogical approach	Women's studies course	Mixed	Adjusted	0.37	0.09
Vogelgesang (2000) [8]	Pedagogical approach	Women's studies course	Whites	Unadjusted	0.36	0.02
Vogelgesang (2000) [2]	Pedagogical approach	Women's studies course	Mixed	Unadjusted	0.35	0.02
Gurin et al. (2002) [2]	Pedagogical approach	Ethnic studies course	Students of color	Unadjusted	0.33	0.12
Vogelgesang (2000) [1]	Pedagogical approach	Ethnic studies course	Mixed	Unadjusted	0.31	0.01
Antonio (2001)	Pedagogical approach	Diversity workshop and training intervention	Mixed	Adjusted	0.30	0.04
Vogelgesang (2000) [7]	Pedagogical approach	Ethnic studies course	Whites	Unadjusted	0.30	0.02
Vogelgesang (2000) [4]	Pedagogical approach	Ethnic studies course	Students of color	Unadjusted	0.25	0.08
Henderson-King and Kaleta (2000) [2]	Pedagogical approach	Required diversity course	Mixed	Unadjusted	0.15	0.11

Fig. 1 Visual displays and summary statistics of the 30 effect sizes (Copyright 2009 American Educational Research Association. Used by permission)

sizes. The whiskers are lines that extend from the box to the highest and lowest values, excluding outliers. The line across the box indicates the median.

Summary statistics include measures of central tendency as well as measures of variability. Measures of central tendency include the mean (average) and median (the number that divides the distribution into halves). Measures of variability include: the minimum (smallest value), maximum (largest value), Q_1 (25th percentile), Q_3 (75th percentile), $Q_3 - Q_1$ (the range for the middle 50%), and standard deviation (measure of dispersion or variation in a distribution). In addition, examining the distance (e.g., in standard deviation units) of the minimum and maximum effect sizes from the mean, Q_1 , and Q_3 of the full distribution of effect sizes is a useful start in data analysis for outliers.

The stem-and-leaf plot in Fig. 1 shows the distribution of the effect sizes for the 30 standardized mean difference racial bias outcomes. The mean of the sample distribution is = 0.54 (median of 0.44), with 95% of the effect sizes falling between 0.1 and 0.9. According to Cohen's (1988) conventions, there is a medium effect of curricular/co-curricular activities on reducing bias. The shape of this distribution is fairly normal and it appears that there is an extreme outlier with an effect size estimate of 2.17. The one extreme outlier is the only study conducted before 1990 (i.e., Katz and Ivey 1977).

Note that the standard error for the effect size estimate from the Katz and Ivey study (i.e., 0.52; see Table 2) is over three times larger than any of the other standard errors in our dataset, resulting in an extremely wide confidence interval. This suggests that the estimate is quite imprecise. This study is unusual in two other respects. It was the only sample in the study published prior to 1990. Secondly, this was a study with an all White sample and an intense intervention (i.e., over two weekends) designed specifically to reduce racism among Blacks. We omitted this study from the first set of HLM analyses presented below, but included it in a second set of analyses to see whether key results were sensitive to the inclusion/exclusion of this case.

Another kind of display known as a forest plot presents the effect sizes graphically with their associated 95% confidence intervals (Light et al. 1994). In a forest plot, each study is represented by a box whose center symbolizes the effect size estimated from that study (random-effects point estimate) and the lines coming out from either side of the box represents the 95% confidence interval. The contribution of each study to the meta-analysis (i.e., its weight) is represented by the area of the box. The summary treatment effect (average) of all the effect sizes can be found at the very bottom of the forest plot and is

Fig. 2 Forest plot of the 29 effect sizes

shown by the middle of a diamond whose left and right extremes represent the corresponding confidence interval.

Figure 2 presents the forest plot of the 29 standardized mean difference effect sizes for the racial bias outcomes. From the forest plot, the 29 effect sizes taken as a group show a significant medium effect of curricular and co-curricular diversity activities on reducing racial bias outcomes. This can be seen in the fact that the diamond does not include zero. Forest plots are also extremely helpful tools for obtaining a sense of how consistent or how variable the studies in a sample appear to be in terms of effect size. The dotted line in Fig. 2 corresponds to the summary effect size for the sample. Note that the 95% intervals for 15 studies in the sample contain or include the summary effect size (e.g., the intervals for the Muthuswamy et al. 2006; Gurin et al. 2004; and Chang 2002 studies). However, we see that the 95% intervals for a number of other studies lie below the summary effect size (e.g., Antonio 2001), and the intervals for a third group of studies lie above the summary effect size (e.g., Lopez et al. 1998); the non-overlap in 95% intervals for the latter two groups of studies signals that the studies in the sample very likely vary substantially in their true effect sizes. In contrast, if the intervals for nearly all of the studies in the sample contained the summary estimate, this would signal that the studies are likely fairly homogenous in terms of their true effect sizes.

Returning to Table 2, we know that connected with each effect size estimate is a certain amount of error that is captured by the standard error of the estimate. One question that arises is: How much of the variability that we see among the effect size estimates is attributable to such error (e.g., lack of precision in each of the estimates), and how much is attributable to actual underlying differences across studies in effect size (i.e., heterogeneity in effect size)? A second question, as noted earlier, is: What factors underlie this heterogeneity? How do differences in various key study characteristics relate to differences in effect size? A key factor that we will focus on in our analyses is the pedagogical approach

employed in a study's program to reduce racial bias. We will also focus on effect size type, i.e., whether or not an effect size is adjusted for possible differences in covariates (e.g., pretest). The following section explains and illustrates how to use HLM analyses to answer these important meta-analytic questions.

HLM Analyses

HLM is an appropriate, effective, and natural technique for quantitative meta-analysis (Raudenbush and Bryk 2002). The data used in quantitative meta-analysis essentially have a nested or hierarchical structure, i.e., subjects (individuals) nested within studies. As noted above, when we consider the variability in effect size estimates for a sample of studies (see, e.g., Table 1, Fig. 2), the variation that we see stems from two sources: the error or lack of precision connected with each study's effect size estimate, and actual differences across studies in their true effect sizes. HLMs for meta-analysis consist of two interconnected models that enable us to represent and partition these two sources of variation. As will be seen, the information regarding the effect size based on the subjects nested in a study is summarized and represented in a within-study (level 1) model; for each study, key elements of the within-study model are the estimate of the true effect size for a study and a variance term based on the standard error of the estimate. To represent and investigate the amount of variance across studies in their true effect sizes, and to study factors that systematically underlie such heterogeneity, we pose a between-study (level 2) model.

Raudenbush and Bryk (2002) offer the valuable perspective that more traditional applications of HLM can in a sense be viewed as meta-analyses. For example, consider an application in which we have a sample consisting of students nested within different institutions, and interest centers on the relationship between student involvement in college activities and satisfaction with campus life. Based on Raudenbush and Bryk's perspective, we can view each institution as providing a study of the relationship between student involvement and satisfaction with campus life. For example, the student-level data nested within a given institution would provide information regarding the magnitude of the slope capturing the relationship between student involvement and satisfaction for that institution. A between-institution (level 2) model would enable us to investigate differences in the relationship between student involvement and satisfaction across institutions. Analogous to between-study models, the level 2 model would provide a means of investigating the amount of heterogeneity in involvement/satisfaction slopes across institutions, and institutional characteristics that might underlie this heterogeneity.

This section presents four HLM models: (1) an unconditional model in which we attempt to estimate an overall effect size and the extent to which effect sizes vary around the overall average ($N = 29$), (2) a conditional model in which differences in effect size are modeled as a function of pedagogical approach ($N = 29$), (3) a conditional model based in which pedagogical approach and effect size type (i.e., adjusted versus unadjusted) are predictors ($N = 29$), and (4) a conditional model in which pedagogical approach is the predictor, and the model is fit to a subsample of studies ($N = 6$).

Model 1

The HLM random-effects model in meta-analysis specifies two linked equations: a within-study model and a between-study model (Raudenbush and Bryk 2002). The within-study model (or level 1 model) is a measurement model relating the estimated effect sizes from each study to the "true" effect sizes:

$$g_j = \delta_j + e_j \quad e_j \sim N(0, V_j) \quad (5)$$

where g_j is the standardized effect size measure or estimate for study j ; δ_j is a parameter capturing the true effect size for study j ; e_j is an error term reflecting that g_j is an estimate of δ_j ; and V_j is the error variance of g_j as an estimate of δ_j . As discussed above there is a standard error connected with each effect size estimate. The error variance (V_j) is simply the squared standard error. Thus in the case of the Lopez et al. study, the standard error of the effect size estimate is 0.16 and the error variance is $0.16^2 = 0.03$ (see Table 2).

Between-study (or level 2) models enable us to represent the fact that studies may vary in their actual effect sizes, estimate the amount of heterogeneity in effect size, and investigate factors that may be systematically related to differences in the magnitude of effect size. We begin by posing a between-study model in which the true effect sizes for a sample of studies are viewed as varying around a mean effect size:

$$\delta_j = \gamma_0 + u_j \quad u_j \sim N(0, \tau) \quad (6)$$

where γ_0 is the mean effect size across all the studies; and u_j is a level 2 residual (or random effect in the parlance of Raudenbush and Bryk (2002)), that allows for the fact that the true effect for a given study may differ to some extent from the mean effect size, i.e., for some studies the true effect may be close the grand mean, while for other studies that true effect may lie appreciably above or below the grand mean. The variance term τ in Eq. 6 represents the amount of variance connected with actual differences in effect size across studies; this is often termed *random effects variance* or *parameter variance* by Raudenbush and Bryk (2002). Substituting Eq. 6 into Eq. 5 yields the following combined model:

$$g_j = \gamma_0 + u_j + e_j \quad (7)$$

where $g_j \sim N(\gamma_0, \tau + V_j)$ and $\text{Var}(g_j) = \tau + V_j$.

How does HLM use the effect size estimates (g_j) for the set of studies in our sample and the error variances (V_j) connected with these estimates to obtain an estimate of the mean effect size (γ_0) and the variance component τ ? Returning to Table 2, note that if the error variances for our studies were approximately equal to 0 (i.e., if the g_j 's were extremely accurate estimates of the δ_j 's), that would mean that practically all of the variability that we see in plots of the effect size estimates (i.e., Fig. 1) would be attributable to actual, underlying differences in effect size (i.e., parameter variance). Alternatively, if the error variances were extremely large (e.g., if the error variance for each study was as large as the error variance for the Katz and Ivey study), that would signal that much of the variability that we see in Fig. 1 may very likely be attributable to error variance.

Thus, when we inspect plots such as Fig. 2, we know that the total variance of a sample of effect size estimates ($\text{Var}(g_j)$) consists of two components: error variance (V_j) and random effects (parameter) variance (τ). What is crucially important is that we have information about the magnitude of error variance. HLM is able to use this information and the g_j 's to compute a maximum likelihood estimate of τ , which we will term $\hat{\tau}$.

With respect to γ_0 , HLM uses the effect size estimates and the corresponding error variances to compute a weighted estimate of γ_0 . More specifically, those studies whose effect sizes are estimated more precisely (i.e., those studies with smaller error variances) are given more weight. The form of the weights is:

$$w_j = \frac{1}{V_j + \hat{\tau}} \quad (8)$$

As can be seen the larger the value of V_j for a given study, the smaller the weight accorded a study.

As we see in Table 3, the estimate of the overall effect size is 0.47 points with a standard error of 0.033. Note that as the number of effect sizes in a sample increases, the smaller the resulting standard error will be. To help capture the precision with which the overall average is estimated, we can use the standard error to construct a 95% interval, which is approximately equal to 0.47 plus and minus twice the standard error. The lower and upper boundaries of the resulting interval are 0.41 and 0.54 respectively. We wish to point out that our estimate of the mean effect size and the corresponding 95% interval are nearly identical to the estimate and interval for the mean effect size that appears in the forest plot in Fig. 2.

The estimate of τ (i.e., the amount of parameter or random effects variance among the effect sizes) is approximately 0.03. Since standard deviations can sometimes be more interpretable measures of spread than variances, we can take the square root of our estimate of τ to obtain a standard deviation capturing the variation in effect sizes (i.e., 0.16). As noted above, our model assumes that true effect sizes are normally distributed around an average effect size with a certain amount of variation. If the normality assumption holds approximately, then based on our results we can use our estimate of the overall mean (0.47) and our measure of variation (0.16) to get a better sense of the extent to which the studies actually differ in their effect sizes.² For example, a study whose effect size is one standard deviation above the mean would be $0.47 + 0.16 = 0.63$, while a study whose effect size is two standard deviations above the mean would be $0.47 + 0.32 = 0.79$. Similarly, studies whose effect sizes are one and two standard deviations below the mean would be 0.31 and 0.15, respectively. Due to the importance of the estimate of the random effects variance (τ) for assessing whether the studies in a sample may tend to be homogeneous and have a common true effect size, or whether the studies may tend to vary appreciably in their true effect sizes (i.e., whether there is substantial heterogeneity around the average effect size), the null hypothesis ($\tau = 0$) should be tested. HLM's chi-square test of the hypothesis that $\tau = 0$ yields a p -value < 0.001 , i.e., there is strong evidence against the null hypothesis.³

Model 2

All of the programs that were investigated in the studies in our sample involved trying to expand the content-based knowledge that people have of other groups. However, three of the studies also involved a cross-racial interaction component (i.e., Lopez et al. 1998; Muthuswamy et al. 2006; Gurin et al. 2002; see Table 2). Thus we now focus on whether we see systematic differences in effect sizes between those studies that focused only on expanding content-based knowledge (C) versus those that also involved cross-racial interaction (CRI) activities. To do this, we employ the same level 1 (within-study) model as above (see Eq. 5), but expand our level 2 (between-study) model as follows:

$$\delta_j = \gamma_0 + \gamma_1(pgm_type_j) + u_j \quad u_j \sim N(0, \tau) \quad (9)$$

² See Raudenbush and Bryk (2002, pp. 274–275) for a method for assessing the reasonableness of this assumption, and for a discussion of alternative estimation approaches when the normality assumption may appear untenable.

³ This test is conceptually similar to the homogeneity test based on the Q statistic in other meta-analytic procedures.

Table 3 HLM summary of Models 1, 2, 3, and 4

Fixed effect	Model 1 ^a			Model 2 ^a			Model 3 ^a			Model 4 ^b		
	Variance component	χ^2	df	Variance component	χ^2	t-Ratio	Estimate (SE)	t-Ratio	Estimate (SE)	Variance component	χ^2	t-Ratio
Intercept, γ_0												
Pedagogical approach, γ_1												
Effect size type, γ_2												
Random effect	Variance component	χ^2	df	Variance component	χ^2	df	Variance component	χ^2	df	Variance component	χ^2	df
True effect size, δ_j	0.027	1171.05***	28	0.025	1155.21***	27	0.024	1145.65***	26	0.000	3.86	4

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ ^a Models 1, 2, and 3 were based on the total sample of 29 studies^b Model 4 was based on a subsample of 6 studies for which adjusted effect size estimates and their error variances were available

where pgm_type_j is equal to a value of 1 if the pedagogical approach for study j involved both expanding content-based knowledge (C) and cross-racial interaction (CRI) activities, and equal to a value of 0 if the pedagogical approach involved only content-based knowledge (C). Note that γ_0 now represents the expected effect size for programs that focus only on C, and γ_1 , the parameter of primary interest in this model, captures the expected difference in effect size between programs that have C + CRI components versus those that involve only a C component. The variance parameter (τ) now represents that amount of parameter variance in effect sizes that remains after taking into account program type.

The results for the fixed effects (γ_0 , γ_1) and τ in Model 2 appear in Table 3. Note that the estimates of the fixed effects are obtained via a weighted regression of the effect size estimates for our sample of 29 studies on pgm_type_j , where the weights are $\frac{1}{v_j + \tau}$. Thus those studies with effect sizes that are estimated more precisely receive more weight. As can be seen, the estimate of the expected effect for C studies is 0.45 (SE = 0.03). Moreover, we see that the expected difference between C only studies versus C + CRI studies is approximately 0.28 (SE = 0.11), a difference that is statistically significant. Thus the expected effect size for C + CRI studies is $0.45 + 0.28 = 0.73$ (SE = 0.11).⁴ This estimate is very sensible when we consider the magnitude of the effect size estimates for the Lopez et al. (1998), Muthuswamy et al. (2006) and Gurin et al. (2004) studies in Table 2 (i.e., 0.87, 0.86 and 0.48, respectively). Note, finally, that the estimate of τ is approximately 0.025. When we compare this estimate with the estimate from the previous analysis (i.e., 0.027), we see that the inclusion of pgm_type_j in the analysis has resulted in a reduction in parameter variance of approximately 7%.

Model 3

We now include effect size type (i.e., adjusted versus unadjusted) in the analysis. One aim is to assess the extent to which results for studies with adjusted effect sizes differ systematically from results for those studies with unadjusted effect sizes. A second is to investigate differences in the effects of C versus C + CRI programs holding constant effect size type. For six of the studies in our sample it was possible to compute adjusted effect sizes (e.g., effect size estimates adjusted for differences in baseline characteristics between program and comparison group students). The scatter plot in Fig. 3 shows that the median effect size (represented by the horizontal bars) for the studies with unadjusted effect sizes was slightly larger than those studies with adjusted effect sizes by approximately a tenth of a point. We also see that among the unadjusted effect sizes, the largest is from a study of a C + CRI program (Muthuswamy et al. 2006). Likewise, among the adjusted effect sizes, the two largest are from studies of C + CRI programs (Lopez et al. 1998; Gurin et al. 2004).

We now expand the between-study model as follows:

$$\delta_j = \gamma_0 + \gamma_1(pgm_type_j) + \gamma_2(es_type_j) + u_j \quad u_j \sim N(0, \tau) \quad (10)$$

⁴ The standard error for the expected effect size for C + CRI studies, can be found by squaring the standard error of the estimate for γ_0 (SE = 0.03), squaring the standard error of the estimate of γ_1 (SE = 0.11), and proceeding as follows:

$$\text{SE of the expected C + CRI effect size} = \sqrt{(0.03)^2 + (0.11)^2}$$

Fig. 3 Scatter plot of the 29 effect sizes by effect size type

where es_type_j takes on a value of 1 if the effect size for study j is an adjusted effect size and 0 otherwise; γ_1 is now the expected difference in effect size between C + CRI and C programs holding constant es_type_j ; γ_2 is the difference in effect size between those studies with adjusted effect sizes versus those with unadjusted effect sizes, holding constant program type; and τ represents the amount of parameter variance that remains after taking into account pgm_type_j and es_type_j .

Similar to a weighted multiple regression analysis, the estimates of the fixed effects in Eq. 10 are obtained via a weighted regression of the effect size estimates in our sample on the program type and effect size type, where the weights are $\frac{1}{\sqrt{v_j + \tau}}$. Looking at the results for Model 3 in Table 3, we see that the resulting estimate for γ_2 is -0.12 , i.e., holding constant program type, the expected effect size for those studies with adjusted effect sizes tend to be slightly smaller. Moreover, we see that the estimate for γ_1 (i.e., the expected difference in effect size between C + CRI and C only programs, holding constant effect size type) is slightly larger than the estimate from the previous analysis (i.e., 0.35 ($SE = 0.10$) vs. 0.28).⁵ However, both this analysis and the previous analysis essentially point to an expected difference in effect size of approximately 0.3 between C + CRI and C only studies.

Model 4

As a final analysis, we focus on the subsample of six studies for which we have adjusted effect sizes, two of which employed C + CRI (i.e., Lopez et al. 1998; Gurin et al. 2004)

⁵ The larger estimate of 0.35 can be viewed as a way of statistically compensating for the fact that adjusted effect sizes tend to be smaller than unadjusted ones. While 2 of the 3 C + CRI studies have adjusted effect sizes, only 4 of the 26 C studies have adjusted effect sizes.

and four of which employed C only (i.e., Stake and Hoffman 2001 [2 and 3]; Chang 2002; Antonio 2001). Thus the results for these studies have been controlled for potential confounders such as pretest differences. In addition, while the samples of students in the set of 29 studies varied in terms of racial composition (e.g., some samples were composed entirely of white students, some entirely of students of color, and some were a mix of white students and students of color), the composition of samples in the subset of six studies were similar in that they consisted of a mix of white students and students of color.

In this analysis, we employ the same between-study model used in our second analysis, i.e., effect size is modeled as a function of program type (see Eq. 9). The results for Model 4 in Table 3 show that the estimate of the expected effect size for C only studies is approximately 0.32 (SE = 0.03), and that the expected difference in effect size between the C + CRI studies and C only studies is 0.38 (SE = 0.13), which is quite consistent with the results from the previous analysis. In addition, the expected effect size for the C + CRI program type category is: $0.32 + 0.38 = 0.70$ (SE = 0.12). Finally note that the resulting estimate of τ is approximately equal to 0, and the test of homogeneity yields a p -value > 0.50 . For this subset of studies, effect size, conditional on program type, appears to be fairly homogeneous. Note that this result is probably driven by the fact that the 4 effect size estimates from the C only studies are highly homogenous, ranging between values of 0.30 and 0.39.

HLM Analyses Summary

Models 2, 3, and 4 all point to the expected effect of programs with a CRI component to be substantially larger than those with content-based knowledge only. Also, the average or expected effect size for programs that involve both components is, based on all three C + CRI studies, 0.73 with an SE of 0.11 (Model 2); and based on the two C + CRI studies with adjusted effect sizes, the expected effect size is 0.70 with an SE of 0.12 (Model 4). Taking a closer look at these three studies (i.e., Gurin 2004; Lopez et al. 1998; Muthuswamy et al. 2006), they were all conducted at either the University of Michigan or Michigan State University. These studies all examined diversity “programs” which have been implemented on a single campus with the specific purpose of improving the institutional diversity climate. Thus, it appears that the observed differences may be due to a combination of the CRI component and institutional commitment to diversity which makes these interventions/programs successful in reducing racial bias.

Sensitivity Analyses

Every meta-analysis involves a number of decisions to be made that can affect the conclusions and inferences drawn from the findings (Greenhouse and Iyengar 1994). Should outliers be included or excluded? Should a random effects model be used (e.g., an HLM model), or should a fixed effects model be used (i.e., a model that, in contrast to HLM, does not include a between-study variance component (τ))? Does publication bias (e.g., systematic differences in effect sizes for published versus unpublished studies) pose a threat to the findings? Because of the many possibilities, it is useful to carry out a series of *sensitivity analyses* to assess whether the assumptions or decisions in fact have a major effect on the results of the review. In other words, are the findings robust to the methods used to obtain them (Cochrane Collaboration 2002)? Thus, sensitivity analyses involve comparing two or more meta-analytic findings calculated based on different assumptions or approaches. Three sensitivity analyses were conducted: (1) inclusion versus exclusion of outliers,

(2) random effects versus fixed effects models, and (3) possible publication bias (i.e., published versus unpublished studies).

Inclusion Versus Exclusion of Outliers

Statistical procedures used to determine the significance of an effect size summary statistic are typically based on certain assumptions, for example, that there is a normal distribution of effect sizes surrounding the true mean effect size. If these assumptions are violated, they can influence the validity of the conclusions drawn from the meta-analytic findings. One of the main purposes of the descriptive results was to examine the distribution of effect sizes by presenting visual displays, measures of central tendency, and measures of variability. From these various plots, there was one obvious outlier which was the only study in the meta-analysis that was published prior to 1990 (i.e., Katz and Ivey 1977). While this study had one of the smallest samples ($n = 24$), it had the largest effect size ($d = 2.17$) which was more than twice as large in magnitude as the next largest effect size. Upon closer inspection, however, this finding was not entirely surprising. First, the study was conducted in the 1970s, long before diversity research really took off. Second, the study participants were all White. Thus, it was apparent that this study was an outlier for a variety of reasons and was eliminated from the inferential analyses.

In the first sensitivity analysis, then, this outlier was added back into the effect size set and compared to the original findings in which this outlier was removed. Table 4 presents a comparison of Model 1 (unconditional model) and Model 3 (final model) both without (left panel) and with (right panel) the outlier. In Model 1, the outlier had very little influence on the estimate of the overall effect (grand mean), variance component, and plausible values range which were all practically identical in both analyses. The outlier had no effect on the 95% confidence interval of the estimate of the overall effect. In Model 3 in which pedagogical approach and effect size type were included as predictors, the results again were virtually identical. Including the outlier slightly affected the estimates for γ_0 and γ_2 . Similar to the findings for Model 1, the variance component is slightly larger in the analysis where the outlier was included which corresponds to a slightly larger plausible values range of (0.17, 0.79) as compared to the previous range of (0.17, 0.78). Overall, it appears that the main findings for Models 1 and 3 are robust to the presence of the outlier. Upon closer inspection, it can be seen that the error variance of 0.27 for this study is substantially larger than any other study (see Table 2). As a result of its large error variance, this study was accorded much less weight in the analyses than the other studies in the sample (see Eq. 8), and therefore it exerted much less influence on the results than the other studies. Thus, including or excluding this outlier makes little difference with respect to the results.

On a related note, the Vogelgesang (2000) study may also be seen as an outlier of sorts because this study contributed nine effect sizes to the analyses (ranging from 0.25 to 0.43). Thus, it is also a good idea to examine what sort of impact the information from her studies is having on the results. In Model 4 where we focus only on those studies that involved matching or covariate adjustment, the effect sizes from her studies are set aside, but we still see a substantial difference in pedagogical approach between the C + CRI and C only studies. Thus, Model 4 may also be viewed as a sensitivity analysis as well.

Another sensitivity analysis is to compare the original findings which include all nine Vogelgesang (2000) effect sizes to an alternative analysis in which the nine effect sizes are averaged into one mean effect size. Table 5 presents a comparison of Models 1 and 3 both with all nine Vogelgesang effect sizes (left panel) and with the nine effect sizes averaged into one (right panel). In Model 1 (top panel), when the Vogelgesang effect sizes were

Table 4 Comparison of Models 1 and 3 (without and with outlier)

<i>Model 1: Without outlier (n = 29)</i>						<i>Model 1: With outlier (n = 30)</i>					
Fixed effect	Coefficient	S.E.	t-Ratio	95% CI		Fixed effect	Coefficient	S.E.	t-Ratio	95% CI	
Grand mean, γ_0	0.47	0.033	14.21***	(0.41, 0.54)		Grand mean, γ_0	0.48	0.034	14.17***	(0.41, 0.54)	
Random effect	Variance component	df	χ^2	PVR		Random effect	Variance component	df	χ^2	PVR	
True effect size, δ_j	0.027	28	1171.05***	(0.15, 0.79)		True effect size, δ_j	0.028	29	1192.46***	(0.15, 0.80)	
<i>Model 3: Without outlier (n = 29)</i>						<i>Model 3: With outlier (n = 30)</i>					
Fixed effect	Coefficient	S.E.	t-Ratio	95% CI		Fixed effect	Coefficient	S.E.	t-Ratio	95% CI	
Intercept, γ_0	0.47	0.037	12.82***	(0.40, 0.54)		Intercept, γ_0	0.48	0.037	12.82***	(0.41, 0.55)	
Pedagogical approach, γ_1	0.35	0.096	3.70**	(0.17, 0.54)		Pedagogical approach, γ_1	0.35	0.095	3.68**	(0.16, 0.54)	
Effect size type, γ_2	-0.12	0.041	-2.96**	(-0.20, -0.04)		Effect size type, γ_2	-0.13	0.041	-3.10**	(-0.21, -0.05)	
Random effect	Variance component	df	χ^2	PVR		Random effect	Variance component	df	χ^2	PVR	
True effect size, δ_j	0.024	26	1145.65***	(0.17, 0.78)		True effect size, δ_j	0.025	27	1166.90***	(0.17, 0.79)	

CI confidence interval, PVR plausible values range: ** $p < 0.01$; *** $p < 0.001$

Table 5 Comparison of Models 1 and 3 (with all 9 Vogelgesang effect sizes and with an averaged Vogelgesang effect size)

<i>Model 1: With all 9 Vogelgesang effect sizes (n = 29)</i>						<i>Model 1: With all 9 Vogelgesang effect sizes (n = 21)</i>					
Fixed effect	Coefficient	S.E.	t-Ratio	95% CI		Fixed effect	Coefficient	S.E.	t-Ratio	95% CI	
Grand mean, γ_0	0.47	0.033	14.21***	(0.41, 0.54)		Grand mean, γ_0	0.52	0.042	12.56***	(0.44, 0.60)	
Random effect	Variance component	df	χ^2	PVR		Random effect	Variance component	df	χ^2	PVR	
True effect size, δ_j	0.027	28	1171.05***	(0.15, 0.79)		True effect size, δ_j	0.030	20	1288.80***	(0.19, 0.86)	
<i>Model 3: With all 9 Vogelgesang effect sizes (n = 29)</i>						<i>Model 3: With an averaged Vogelgesang effect size (n = 21)</i>					
Fixed effect	Coefficient	S.E.	t-Ratio	95% CI		Fixed effect	Coefficient	S.E.	t-Ratio	95% CI	
Intercept, γ_0	0.47	0.037	12.82***	(0.40, 0.54)		Intercept, γ_0	0.54	0.036	14.91***	(0.47, 0.61)	
Pedagogical approach, γ_1	0.35	0.096	3.70**	(0.17, 0.54)		Pedagogical approach, γ_1	0.33	0.096	3.43**	(0.14, 0.52)	
Effect size type, γ_2	-0.12	0.041	-2.96**	(-0.20, -0.04)		Effect size type, γ_2	-0.19	0.049	-3.77**	(-0.28, -0.09)	
Random effect	Variance component	df	χ^2	PVR		Random effect	Variance component	df	χ^2	PVR	
True effect size, δ_j	0.024	26	1145.65***	(0.17, 0.78)		True effect size, δ_j	0.024	18	1374.30***	(0.23, 0.84)	

CI confidence interval, PVR plausible values range: ** $p < 0.01$; *** $p < 0.001$

averaged into one effect size, the estimate of the overall effect (grand mean) increased from 0.47 to 0.52 and the standard error increased from 0.033 to 0.042. Also, the 95% confidence interval changed from (0.41, 0.54) to (0.44, 0.60) while the plausible values range changed from (0.15, 0.79) to (0.19, 0.86). This increase in the overall effect, standard error, confidence interval, and plausible values range is to be expected given that the number of effect sizes reduced from 29 to 21 and that the nine Vogelgesang effect sizes ranged from 0.25 to 0.43. In Model 3 (bottom panel) in which pedagogical approach and effect size type were included as predictors, the pattern of change in the results was similar to the pattern of change in Model 1 (top panel). Averaging all nine Vogelgesang effect sizes into one average effect size slightly affected the estimates, 95% confidence intervals, and plausible values ranges. However, the overall findings and conclusion remain the same. Thus, it appears that the main findings for Models 1 and 3 are robust to the fact that the Vogelgesang study, in contrast to the other studies in our sample, contributed 9 effect sizes to the sample of 29 effect sizes used in our original analyses.

Random Effects Versus Fixed Effects

Once it is decided which studies are to be included, the next step is to make a decision regarding the method of analysis. For example, should a fixed effects model or random effects model be assumed? A fixed effects model, which does not include a between-study variance component (τ), essentially assumes that all of the variability between effect sizes is due to sampling error, that is, the “luck of the draw” associated with sampling the subjects in each study from the population of potential subjects (Lipsey and Wilson 2001). Thus, a fixed effects model weights each study by the inverse of the error variance:

$$w_j = \frac{1}{V_j} \quad (11)$$

where w_j represents the weight and V_j represents the error variance for study j . On the other hand, a random effects model assumes that the variability between effect sizes is due to error variance *plus* actual differences in effect sizes across studies in the domain of interest (i.e., parameter or random effects variance) (see Eq. 8; Lipsey and Wilson 2001; Raudenbush and Bryk 2002). Note that if there is appreciable between-study variability, as in our analyses based on the sample of 29 effect sizes, then a random effects model (i.e., an HLM model) is to be preferred as it will yield more appropriate standard errors for estimates of overall effect sizes and of the coefficients for moderators of effect size. However, it is still useful to gauge the extent to which the results based on the two approaches might differ.

Thus, the second sensitivity analysis compared the results based on our HLM analyses (original analyses) to results based on a fixed effects model (comparison analyses). Table 6 presents a comparison of Models 1 and 3 assuming a random effects model (left panel) and a fixed effects model (right panel). In Model 1 (top panel), the estimates of the overall effect and standard error were smaller in magnitude in the fixed effects model ($\gamma_0 = 0.45$; $SE = 0.01$) as compared to the random effects model ($\gamma_0 = 0.47$; $SE = 0.03$). Because the standard error was smaller in the fixed effects model, the 95% confidence interval was also much smaller in the fixed effects model (0.45, 0.46) as compared to the random effects model (0.41, 0.54). In Model 3 (bottom panel), some of the differences were similar to those in Model 1 (top panel). Specifically, the estimates of the overall effect and standard error were smaller in magnitude in the fixed effects model ($\gamma_0 = 0.46$; $SE = 0.01$) as compared to the random effects model ($\gamma_0 = 0.47$; $SE = 0.04$). Also, the 95% CI was also

Table 6 Comparison of Models 1 and 3 (random effects versus fixed effects)

<i>Model 1: Random effects model</i>						<i>Model 1: Fixed effects model</i>					
Fixed effect	Coefficient	S.E.	t-Ratio	95% CI		Fixed effect	Coefficient	S.E.	t-Ratio	95% CI	
Grand mean, γ_0	0.47	0.033	14.21***	(0.41, 0.54)		Grand mean, γ_0	0.45	0.005	95.19***	(0.45, 0.46)	
Random effect	Variance component	df	χ^2	PVR		Random effect	Variance component				
True effect size, δ_j	0.027	28	1171.05***	(0.15, 0.79)		True effect size, δ_j	–				
<i>Model 1: Random effects model</i>						<i>Model 1: Fixed effects model</i>					
Fixed effect	Coefficient	S.E.	t-Ratio	95% CI		Fixed effect	Coefficient	S.E.	t-Ratio	95% CI	
Intercept, γ_0	0.47	0.037	12.82***	(0.40, 0.54)		Intercept, γ_0	0.46	0.005	94.48***	(0.45, 0.47)	
Pedagogical approach, γ_1	0.35	0.096	3.70***	(0.17, 0.54)		Pedagogical approach, γ_1	0.39	0.108	3.62***	(0.18, 0.60)	
Effect size type, γ_2	–0.12	0.041	–2.96**	(–0.20, –0.04)		Effect size type, γ_2	–0.14	0.029	–4.77***	(–0.20, –0.08)	
Random effect	Variance component	df	χ^2	PVR		Random effect	Variance component				
True effect size, δ_j	0.024	26	1145.65***	(0.17, 0.78)		True effect size, δ_j	–				

CI confidence interval, PVR plausible values range; ** $p < 0.01$; *** $p < 0.001$

much smaller in the fixed effects model (0.45, 0.47) as compared to the random effects model (0.40, 0.54). The estimates for pedagogical approach and effect size type had varying results. For example, the estimates for pedagogical approach and effect size type were all slightly larger in magnitude in the fixed effects model as compared to the random effects model. The standard errors of these estimates, however, were larger for pedagogical approach but smaller for effect size type in the fixed effects model as compared to the random effects model. However, the overall findings remain unchanged. Thus, it appears that the main findings for Models 1 and 3 are robust to the choice of model assumed (i.e., fixed effects versus random effects model).

File Drawer Problem: Publication Bias Problem

The final sensitivity analysis concerns the “file drawer problem”—a problem that threatens every meta-analytic review (Lipsey and Wilson 2001; Rosenthal 1979, 1991). The file drawer problem is a criticism of meta-analysis in that the available studies to the researcher(s) may not be representative of all the studies ever conducted on a particular topic (Becker 1994; Greenhouse and Iyengar 1994; Rosenthal 1979, 1991). The most likely explanation for this is that researchers may have unpublished manuscripts tucked away in their “file drawers” because their results were not statistically significant and were therefore never submitted for publication. Thus, publication bias arises when studies reporting statistically significant findings are published whereas studies reporting less significant or nonsignificant results are not. As a result, many meta-analyses may systematically overestimate effect sizes because they rely mainly on published sources. Of greatest concern is whether the file drawer problem is large enough in a given meta-analysis to influence the conclusions drawn (Lipsey and Wilson 2001). Consequently, it is suggested that mean comparisons, sample size and effect size comparisons, as well as calculation of a fail-safe N were employed in order to examine this possibility (Light and Pillemer 1984; Lipsey and Wilson 2001).

Mean comparisons. The first approach to assessing the potential magnitude of the file drawer problem is to compare the mean effect sizes for the published (e.g., journal articles) versus unpublished studies (e.g., conference presentations, dissertations) included in a meta-analysis (Lipsey and Wilson 2001). If publication bias did exist, it is expected that the mean effect size for published studies would be higher than the mean effect size of the unpublished studies. In Fig. 4, the mean effect size of the published studies was identical to the mean of the unpublished studies (i.e., both were 0.48), suggesting no publication bias.

Sample size and effect size comparisons. A second approach involves examining the relationship between sample size and effect size via correlation and funnel plots. Nonsignificant correlations and symmetrical funnel plots would each suggest minimal publication bias. The observed correlation between sample size and effect size was nonsignificant, indicating that the magnitude of the effect size does not depend on sample size. This relationship (or lack thereof) can also be examined visually via funnel plots. Figure 5 presents the funnel plot of effect sizes. If no bias is present, the scatterplots should be shaped like an upside down funnel, with the spout pointing up. In other words, the bottom of the funnel should have a wider spread corresponding to the smaller studies and the top of the funnel should have a narrower spread corresponding to the larger studies. Also, the mean effect size should be the same regardless of sample size. That is, the shape of the upside down funnel should be symmetrical. If there is publication bias against studies showing small effect sizes, for example, a bite will be taken out of the lower left-hand part of the funnel plot. Figure 5

Fig. 4 Schematic plot of the 29 effect sizes for unpublished and published studies

Fig. 5 Funnel plot of the 29 effect sizes by sample size

displays the scatter plot of effect size and sample size and roughly resembles a funnel, suggesting that publication bias was not a major problem.⁶

⁶ Technically the logic of funnel plots is rooted in the notion that the studies in a sample are homogeneous, i.e., that they vary only in terms of error variance and share a common, overall effect size. It is often the case, however, that effect sizes are heterogeneous. In such situations, there is the possibility that a funnel plot may have what appears to be a “bite” taken out of the lower left portion of the funnel, but the pattern may not be due to publication bias. For example, if programs are implemented more carefully or more intensively in studies with small sample sizes, such studies may tend to have larger effect sizes compared with large-scale studies; in that case, the distribution of effect sizes for small studies would be shifted toward the right of the plot, leaving what appears to be a “bite” in the lower left portion of the plot (Shadish, personal communication, February 2010; Sutton 2009). Thus, more generally it is important to check if study sample size is systematically related to study characteristics that might moderate the magnitude of effects.

Table 7 Summary of Orwin's (1983) fail-safe N

	Number of weighted ES (k)	Weighted mean ES ($\overline{ES}_k$)	Criterion ES level (ES_c)	Orwin's (1983) fail-safe N (k_0)
Racial bias outcomes	29	0.47	0.20	39

Rosenthal's (1979)/Orwin's (1983) fail-safe N . A final approach is to test publication bias empirically. Rosenthal (1979) developed a fail-safe N that can be calculated in order to estimate possible publication bias. This fail-safe N estimate represents the number of additional studies with nonsignificant results that would have to be added to the sample in order to change the combined p from significant (at the 0.05 or 0.01 level of confidence) to nonsignificant (Rosenthal 1979). This fail-safe N was developed for use with Rosenthal's method of cumulating z -values across studies. This fail-safe N formula estimates the number of additional studies needed to lower the cumulative z below the desired significance level, for example, a z equal to or less than 1.645 ($p \geq 0.05$). However, Rosenthal's (1979) fail-safe N is limited to probability levels only.

In response, Orwin (1983) adapted Rosenthal's (1979) approach and introduced an analogous statistic that is applicable to effect sizes. The major difference between the two approaches is that there is no "standard" or "minimally acceptable" effect size analogous to the "standard" confidence levels (p values) of 0.05 and 0.01. As a result, Orwin (1983) has suggested the use of Cohen's d (1988) effect size classification as small (0.2), medium (0.5), and large (0.8) for comparison purposes. Thus, Orwin's (1983) fail-safe N formula determines the number of studies with an effect size of zero needed to reduce the mean effect size to a specified or criterion level and is as follows

$$k_0 = k \left[\frac{\overline{ES}_k}{\overline{ES}_c} - 1 \right], \quad (12)$$

where k_0 is the number of effect sizes with a value of zero needed to reduce the mean effect size to $\overline{ES}_c$, k is the number of studies in the mean effect size, $\overline{ES}_k$ is the weighted mean effect size, and $\overline{ES}_c$ is the criterion effect size level.

The fail-safe N 's utility is similar to that of the confidence interval (Carson et al. 1990). The fail-safe N provides additional information on the degree of confidence that can be placed in a particular meta-analytic result, in this case, the overall mean effect of curricular and co-curricular diversity activities on racial bias in college students. An important difference, however, is that the fail-safe N assumes that the unobserved studies (i.e., those tucked away in a file drawer somewhere) are assumed to have a null result. For example, if the fail-safe N is relatively small in comparison to the number of studies in the meta-analysis, caution should be used when interpreting meta-analytic results. On the other hand, if the fail-safe N is relatively large in comparison to the number of included studies, then more confidence is assured regarding the stability of the results. Rosenthal (1979) suggests as a general rule a reasonable tolerance level of $5k + 10$. If the fail-safe N exceeds the $5k + 10$ benchmark, then a file-drawer problem is unlikely.

Table 7 presents the summary of Orwin's (1983) fail-safe N for the effect sizes, and displays the number of effect sizes, weighted mean effect sizes, criterion level effect sizes, fail-safe N , as well as Rosenthal's (1979) tolerance level. The fail-safe N is 39, which is the number of studies with an effect size of zero that is needed to reduce the

mean effect size from 0.47 (a moderate effect size) to 0.20 (a small effect size). This fail-safe N does not exceed Rosenthal's tolerance of 155, suggesting that the file-drawer problem is of potential concern. However, the probability that there are 39 studies with an effect size of zero for which we have failed to obtain is unlikely. For an overview of various techniques for investigating publication bias, including recently developed techniques, see Sutton (2009).

Study Limitations and Implications

One of the important outcomes of every meta-analysis is a set of implications and recommendations for the conduct and reporting of future work in a given domain. As such, this section presents some guidelines and recommendations for reporting of results and designing of studies to enable and enhance future meta-analytic reviews.

Limitations Connected with the Reporting of Results

A relatively small number of effect sizes (29) were included in the study, which was due in part to limitations in the reporting of results. While many studies made an effort to statistically control for prior differences in various covariates (e.g. pretest), there was a lack of adequate information for computing adjusted effect sizes and/or their error variances. In our analyses, we saw that the unadjusted effect sizes tended to be only slightly larger, on average, than the adjusted effect sizes. Furthermore, we saw that the programs with both content-related knowledge and cross-racial interaction appear to have larger effect sizes than the content-related knowledge only studies when we control for effect size type, and when we focus only on the subset of studies for which we have adjusted effect sizes. Yet it is important that studies try to report the information that would be necessary for computing adjusted effect sizes and their error variances.

There is also a lack of detailed information regarding the diversity-related interventions themselves. Many of the characteristics of the curricular and co-curricular diversity activities themselves were chosen to be coded such as intensity of cross-racial interaction (random, built-in), duration of intervention (in weeks), and intensity of intervention (in hours/week). These characteristics were chosen in the hopes of exploring why some diversity-related activities are more effective than others. However, many of the single-institution studies did not report this information in sufficient detail and many of the multi-institution studies relied on survey data which only provided information about whether or not they participated in a certain type of diversity activity. Thus, unfortunately not many characteristics of the curricular and co-curricular diversity activities themselves could be explored as possible moderators in this meta-analysis.

In both of their books on *How College Affects Students*, Pascarella and Terenzini (1991, 2005) attempted to conduct “mini” meta-analyses whenever possible. However, they quickly realized that a substantial percentage of studies simply did not report sufficient information in order to be able to compute effect sizes. Smart (2005) describes seven attributes that distinguish exemplary manuscripts from other manuscripts that utilize quantitative research methods. One of Smart's (2005) recommendations for exemplary manuscripts is to report evidence that permits others to reproduce the results, such as means, standard deviations, and correlations among all the variables included in the analyses at the very minimum. These simple but informative descriptive statistics allow calculation of effect sizes for future meta-analytic reviews.

Limitations Connected With the Design of Studies

In the case of this illustrative meta-analysis of the effects of diversity programs, programs that incorporated both content-related knowledge and cross-racial interaction emerge as being very promising. Given the small number of such studies, additional research on the effects of these programs is needed. In connection with this, it would be valuable for researchers to conduct studies that directly compare content-based knowledge programs with programs that include the additional cross-racial interaction component.

Many of the studies could also benefit from increased research rigor. Due to the nature of this research, it is unsurprising that the current literature is overwhelmingly correlational. As such, there is concern over threats to internal validity (Shadish et al. 2002). In other words, how do we know that the curricular or co-curricular diversity activity is responsible for the change(s) seen in students' racial bias? One of the major shortcomings of these studies is that there is clearly self-selection of the students into these curricular and co-curricular diversity activities. While randomly assigning students to content only or content plus cross-racial interaction programs might not be feasible, efforts should be made to construct these two groups of students that are as comparable as possible at the outset (e.g., Shadish et al. 2002). For example, greater care should be taken to choose comparison groups that are similar in important ways to the groups of program participants and to utilize a rich set of covariates to adjust for initial differences.

Conclusion

In this article, we have demonstrated how this powerful analytical technique can be used to quantitatively synthesize the research findings on a specific research topic. HLM plays a key role as a powerful and natural methodology for a broad class of meta-analytic phenomena. Complementing qualitative literature reviews, meta-analysis—and particularly meta-analysis utilizing HLM—provides a quantitative “big picture” view of the available knowledge base as well as explanations for the similarities and differences among the various studies (Rosenthal and DiMatteo 2001). While originally developed to assess the overall average effectiveness of a program or intervention, what is exciting are the moderator analyses within meta-analysis that can explain the factors that diminish or amplify the effect sizes. No matter what the topic—from the effects of service learning to the effects of university outreach programs—as the research accumulates oftentimes it becomes increasingly difficult to make sense of the literature. It is at this stage when an HLM meta-analysis would be valuable. By providing a brief overview and an illustrative example of meta-analysis utilizing HLM in higher education, we hope to encourage more widespread use of this methodological approach in the field.

Acknowledgments The authors wish to thank Ernest T. Pascarella and the two anonymous reviewers for their valuable feedback and thoughtful comments in the writing of this manuscript.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Antony, J. (1993, November). *Can we all get along? How college impacts students' sense of the importance of promoting racial understanding*. Paper presented at the Annual Meeting of the Association for the Study of Higher Education, Pittsburgh. (ERIC Document Reproduction Service No. ED 365174).
- Antonio, A. L. (2001). The role of interracial interaction in the development of leadership skills and cultural knowledge and understanding. *Research in Higher Education*, 42, 593–617.
- Becker, B. J. (1994). Combining significance levels. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 215–230). New York: Russell Sage Foundation.
- Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 221–235). New York: Russell Sage Foundation.
- 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.
- Carson, K. P., Schriesheim, C. A., & Kinicki, A. J. (1990). The usefulness of the “fail-safe” statistic in meta-analysis. *Educational and Psychological Measurement*, 90, 233–243.
- Chang, M. J. (2002). The impact of an undergraduate diversity course requirement on students' racial views and attitudes. *Journal of General Education*, 51(1), 21–42.
- Cochrane Collaboration. (2002). *Further issues in meta-analysis: Sensitivity analysis*. Retrieved December 3, 2006, from <http://www.cochrane-net.org/openlearning/HTML/mod14-2.htm>.
- Cohen, J. (1960). A coefficient for agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cooper, H. (1982). Scientific guidelines for conducting integrative research reviews. *Review of Educational Research*, 52, 291–302.
- Cooper, H., & Hedges, L. V. (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.
- Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- Denson, N. (2009). Do curricular and co-curricular diversity activities influence racial bias? A meta-analysis. *Review of Educational Research*, 79, 805–838.
- Engberg, M. E. (2004). Improving intergroup relations in higher education: A critical examination of the influence of educational interventions on racial bias. *Review of Educational Research*, 74(4), 473–524.
- Feldman, K. A., & Newcomb, T. M. (1969). *The impact of college on students*. San Francisco: Jossey-Bass.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Research*, 5 (10), 3–8.
- Glass, G. V. (2000). *Meta-analysis at 25*. Retrieved April 30, 2009, from <http://glass.ed.asu.edu/gene/papers/meta25.html>.
- Greenhouse, J. B., & Iyengar, S. (1994). Sensitivity analysis and diagnostics. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 383–398). New York: Russell Sage Foundation.
- Gurin, P., Dey, E. L., Hurtado, S., & Gurin, G. (2002). Diversity and higher education: Theory and impact on educational outcomes. *Harvard Educational Review*, 72, 330–366.
- Gurin, P., Nagda, B. A., & Lopez, G. E. (2004). The benefits of diversity in education for democratic citizenship. *Journal of Social Issues*, 60, 17–34.
- Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6, 107–128.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- Henderson-King, D., & Kaleta, A. (2000). Learning about social diversity: The undergraduate experience and intergroup tolerance. *Journal of Higher Education*, 71(2), 142–164.
- Hogan, D. E., & Mallott, M. (2005). Changing racial prejudice through diversity education. *Journal of College Student Development*, 46, 115–125.
- Hyun, M. (1994, November). *Helping to promote racial understanding: Does it matter if you're Black or White?* Paper presented at the Annual Meeting of the Association for the Study of Higher Education, Tucson. (ERIC Document Reproduction Service No. ED 375710).
- Katz, J. H., & Ivey, A. (1977). White awareness: The frontier of racism awareness training. *Personnel and Guidance Journal*, 55, 485–489.
- Kosslyn, S. M. (1994). *Elements of graph design*. New York: Freeman.
- Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.
- Light, R. J., Singer, J. D., & Willett, J. B. (1994). The visual presentation and interpretation of meta-analyses. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 439–453). New York: Russell Sage Foundation.

- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications, Inc.
- Lopez, G. E., Gurin, P., & Nagda, B. A. (1998). Education and understanding structural causes for group inequalities. *Political Psychology*, 19, 305–329.
- MacPhee, D., Kreutzer, J. C., & Fritz, J. J. (1994). Infusing a diversity perspective into human development courses. *Child Development*, 65, 699–715.
- Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin*, 89, 191–216.
- Milem, J. F. (1994). College, students, and racial understanding. *Thought and Action*, 9(2), 51–92.
- Muthuswamy, N., Levine, T. R., & Gazel, J. (2006). Interaction-based diversity initiative outcomes: An evaluation of an initiative aimed at bridging the racial divide on a college campus. *Communication Education*, 55, 105–121.
- Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, 8, 157–159.
- Pascarella, E. T., & Terenzini, P. T. (1991). *How college affects students: Findings and insights from twenty years of research*. San Francisco: Jossey-Bass.
- Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students: A third decade of research*. San Francisco: Jossey-Bass.
- Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology*, 76, 85–97.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.
- Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin*, 86, 638–641.
- Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage Publications, Inc.
- Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin*, 118, 183–192.
- Rosenthal, R., & DiMatteo, M. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, 52, 59–82.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin Company.
- Smart, J. C. (2005). Attributes of exemplary research manuscripts employing quantitative analyses. *Research in Higher Education*, 46, 461–477.
- Stake, J. E., & Hoffman, F. L. (2001). Changes in student social attitudes, activism, and personal confidence in higher education: The role of women’s studies [Electronic version]. *American Educational Research Journal*, 38(2), 411–436.
- Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 435–452). New York: Russell Sage Foundation.
- Taylor, S. H. (1994). Enhancing tolerance: The confluence of moral development with the college experience. *Dissertation Abstracts International*, 56(01), 114A. UMI No. 9513291.
- Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- Vogelgesang, L. J. (2000). *The impact of college on the development of civic values and skills: An analysis by race, gender and social class*. Unpublished doctoral dissertation, University of California, Los Angeles.
- Wells, A. S., & Crain, R. L. (1994). Perpetuation theory and the long-term effects of school desegregation. *Review of Educational Research*, 64, 531–555.