ETHICS & BEHAVIOR, *19*(5), 379–402

ISSN: 1050-8422 print / 1532-7019 online
DOI: 10.1080/10508420903035380

Routledge
Taylor & Francis Group

A Meta-Analysis of Ethics Instruction Effectiveness in the Sciences

Alison L. Antes
University of Oklahoma

Stephen T. Murphy
Pearson

Ethan P. Waples
University of Central Oklahoma

Michael D. Mumford, Ryan P. Brown, Shane Connelly, and Lynn D. Devenport
University of Oklahoma

Scholars have proposed a number of courses and programs intended to improve the ethical behavior of scientists in an attempt to maintain the integrity of the scientific enterprise. In the present study, we conducted a quantitative meta-analysis based on 26 previous ethics program evaluation efforts, and the results showed that the overall effectiveness of ethics instruction was modest. The effects of ethics instruction, however, were related to a number of instructional program factors, such as course content and delivery methods, in addition to factors of the evaluation study itself, such as the field of investigator and criterion measure utilized. An examination of the characteristics contributing to the relative effectiveness of instructional programs revealed that more successful programs were conducted as seminars separate from the standard curricula rather than being embedded in existing courses. Furthermore, more successful programs were case based and interactive, and they allowed participants to learn and practice the application of real-world ethical decision-making skills. The implications of these findings for future course development and evaluation are discussed.

Keywords: ethics, scientific ethics, ethics instruction, ethics training, meta-analysis

Cases of scientific misconduct range from extremely serious events, such as blatant fabrication of study findings and harm to research participants (Resnick, 2003), to less serious, yet more prevalent, instances of misbehavior, such as inappropriate assignment of authorship and withholding details of methodology or results in publications (Martinson, Anderson, & de Vries, 2005). Regrettably, instances of misconduct undermine progress in science and, moreover, create a sense of distrust for science among the public and breed distrust within the scientific community (Abbott,

Correspondence should be addressed to Alison Antes, Center for Applied Social Research, The University of Oklahoma, 3100 Monitor, Suite 100, Norman, OK 73072. E-mail: alisonantes@ou.edu

1999; Friedman, 2002; Kalichman, 2007). As the nature of science continues to become increasingly competitive, interdisciplinary, and global, not only do new ethical considerations enter the field, but the implications of scientific misconduct become even more significant. Thus, it is not surprising that the scientific community is paying a great deal of attention to understanding unethical behavior in scientific work and what might be done to manage it.

Among the most commonly suggested remedies for addressing this growing concern is to provide ethics education to scientific researchers and practitioners. In fact, some institutions have implemented mandatory ethics instruction in an attempt to manage scientific misconduct (e.g., Barnes et al., 2006). Moreover, funding agencies, such as the National Institutes of Health, now mandate that scientists complete an instructional program in the responsible conduct of research in order to be eligible for funding under their sponsorship (Dalton, 2000). Given the widespread application of instruction in ethics as a potential solution for misbehavior in the sciences, not to mention the substantial time and resources required for the development and implementation of instructional programs, a critical question arises: Are such programs effective?

Although addressing this question has been of primary concern for researchers in the field of ethics, the approaches to designing and assessing instructional programs in ethics have been quite varied. Consequently, although there appears to be a general consensus about the importance of ethics education for researchers and scientists, there is little agreement about the most effective approach to instruction, or even the most appropriate goals for these programs (Kalichman, 2007; Steneck & Bulger, 2007). Moreover, evaluation studies have reported mixed findings regarding the effectiveness of instruction. Some ethics courses have been shown to induce the desired effects, whereas others indicate little or no effects of ethics instruction on learning outcomes (Kalichman & Plemmons, 2007).

The intent of the present study was to provide a comprehensive examination of ethics instruction effectiveness. Meta-analytic procedures were used to quantitatively assess prior program evaluation efforts. In addition to addressing the general effectiveness of ethics instruction, key characteristics of instructional programs and evaluation efforts that may be associated with degrees of effectiveness were identified. Before turning to the specifics of the present study, we first consider some key issues with respect to approaches to ethics instruction.

ETHICS INSTRUCTION IN THE SCIENCES

As previously noted, researchers have taken several approaches to the design of ethics instruction. Distinctions between instructional approaches are important because they point to a fundamental issue with respect to ethics education. Specifically, alternative approaches reflect differences in the frameworks being applied for understanding ethical behavior. These frameworks lead to different assumptions about how ethical behavior might be improved and ultimately lead to differences in the goals and design of instructional courses (Tannenbaum & Yukl, 1992).

A number of ethics instructional courses in the sciences rely on Kohlberg's (1969) and Rest's (1986) models of moral development and moral reasoning. Although these models are commonly referenced for constructing and conducting instructional programs in ethics, the implementation of the programs conducted under these frameworks vary rather widely. These differences seem to arise as a function of how the models are interpreted and what aspects are specifically emphasized. For instance, one approach to ethics instruction has been to emphasize ethical sensitivity.

Rest (1986) asserted that ethical sensitivity is an awareness of the ethical implications of a situation and involves empathetic understanding of how others might be affected by the situation. Instructional programs based on ethical sensitivity assume that improving ethical behavior rests in enhancing scientists' ability to recognize the presence of an ethical problem, as this is the first step in real-world ethical decision-making (Clarkeburn, 2002; Clarkeburn, Downie, & Matthew, 2002; Myyry & Helkama, 2002).

Another common approach to ethics instruction has emphasized the developmental nature of ethical behavior. In this approach, students may be merely exposed to the regular curriculum with the expectation that general education in health and medicine, for instance, might implicitly advance students' level of moral development (e.g., Bebeau & Thoma, 1994; Duckett et al., 1997; Self, Schrader, Baldwin, & Wolinsky, 1993). Other programs operating under this developmental framework have emphasized the abstract, philosophical nature of moral dilemmas (e.g., Goldman & Arbuthnot, 1979; Penn, 1990). The intent of such courses is to promote students' progress to an advanced stage of moral development by shifting one's thinking about abstract moral dilemmas to a more sophisticated level. In turn, it is believed that a higher level of moral development will translate into improved moral reasoning and ethical behavior.

Other ethics instructional programs, more directly emphasize the cognitive nature of moral reasoning. This approach focuses less on the philosophical nature of moral dilemmas and ascending to a higher level of moral development, and puts greater emphasis on the need to think through and analyze complex ethical problems before responding (e.g., Frisch, 1987; Gaul, 1987). The underlying assumption of this approach is that moral reasoning is a function of how one thinks through an ethical problem. Thus, ethical behavior improves as ethical problem-solving and decision-making skills are enhanced (Gawthrop & Uhlemann, 1992).

In line with this assertion, some researchers have emphasized the importance of understanding the cognitive nature of ethical decision making (e.g., the specific processes underlying it) along with individual, situational, and organizational influences on these processes (Antes et al., 2007; Jones, 1991; O'Fallon & Butterfield, 2005; Treviño, 1986; Treviño, Weaver, & Reynolds, 2006). Recently, scholars have argued that these rational approaches may be made even more complete by focusing not only on cognitive processes of ethical decision making but also on social-psychological processes and the emotional nature of ethical problem solving (Haidt, 2001; Sonenshein, 2007).

It is important to note at this juncture that, although many programs follow the common themes noted above, ethics courses do not always fit neatly into clear categories. Rather, some courses employ a mixture of themes for ethics education (e.g., Ryden & Duckett, 1991). Given that differing frameworks for developing ethics instruction have led to a number of differences in instructional programs, we examined a host of characteristics of instructional design and delivery that might impact the effectiveness of ethics courses. Moreover, aspects of an evaluation study itself that might impact the observed effectiveness of instruction were also examined. In the following section, we outline several plausible moderators of instructional effectiveness.

POTENTIAL MODERATORS OF INSTRUCTIONAL EFFECTIVENESS

Although an understanding of instructional effectiveness in general is of value, identification of moderating variables linked to the effectiveness of instruction provides practical guidance for the design and delivery of instruction. Therefore, based on the aforementioned distinctions in

ethics instruction, in addition to recommendations offered by experts in instructional design and evaluation (cf. Goldstein & Ford, 2002; Wexley & Latham, 2002), seven categories of factors that might account for differences in the effectiveness of ethics instruction were examined. These categories of factors likely to influence the success of ethics instruction included (a) criterion type, (b) study design characteristics, (c) participant characteristics, (d) quality ratings, (e) instructional content, (f) general instructional characteristics, and (g) characteristics of instructional methods.

Criterion Type

A distinction between types of criteria used to examine the effects of ethics instruction was included in the analysis. The criterion measure selected to assess instruction should reflect the intended outcome of the instructional program (Kraiger & Jung, 1996). For example, if the program is intended to enhance ethical sensitivity, then ethical sensitivity is the most appropriate criterion. Given the widespread application of Kohlberg's (1969) and Rest's (1986) models of cognitive-moral development, their measures of moral reasoning, or moral judgment, have been the most commonly applied instruments for assessing instructional effectiveness. More specifically, Kohlberg (1976) developed a measure of moral judgment called the Moral Judgment Scale (MJS). Rest (1976, 1988) constructed a measure of moral development (the Defining Issues Test [DIT]) intended to address the shortcomings of the complex coding procedure used for Kohlberg's MJS. The DIT requires an individual to select responses to six moral problems. Examining different criteria types, such as ethical sensitivity and moral reasoning, makes possible an examination of the relative differences in the impact of ethics instruction on these outcomes.

Study Design Characteristics

Characteristics of the design of the evaluation study may impact the size of the effects observed for instruction (Goldstein & Ford, 2002). For example, the type of design employed in the study (e.g., pre–post, pre–post with control, post only, or longitudinal) and the sample size can systematically influence the observed effectiveness of ethics programs (Kirk, 1995). Moreover, factors related to the design of the study have implications for the internal validity of the study and thus the value of any conclusions drawn from it. For example, whether the author of the study was involved in instructing the course might introduce bias or demand characteristics (Cook & Campbell, 1979) that ultimately influence observed effects. Along related lines, another variable that might be important for the validity of conclusions drawn from these studies is whether the study was externally funded. In fact, externally funded studies are generally more likely to produce larger effects (Conn, Valentine, Cooper, & Rantz, 2003).

Participant Characteristics

Characteristics of the participants may play a role in the effectiveness of instruction and, therefore, may have implications for the generalizability of findings regarding ethics instruction effectiveness. For instance, participants' career stage and field of study may impact ethical reasoning and responses to ethics instruction (Mumford et al., 2009; Weeks, Moore, McKinney, & Longenecker, 1999). In addition, several studies suggest that gender and age may influence ethi-

cal attitudes and behavior (Borkowski & Ugras, 1998; Ruegger & King, 1992; Weeks et al., 1999). Moreover, whether participants received an incentive to complete the ethics course might be associated with differences in motivation for completing the course and thus reactions to, and outcomes of, ethics instruction (Colquitt & Simmering, 1998).

Quality Ratings

The impact of general quality variables on instructional effectiveness was examined (Scott, Leritz, & Mumford, 2004a). Differences across studies in the overall quality of the instructional program, the overall quality of the study design, and the overall quality of the criteria utilized were captured via three subjective quality ratings assigned by expert raters. A more detailed description of these ratings is provided in the method.

Instructional Content

As noted previously, the approach taken to instruction creates rather significant variations in instructional content. For instance, courses may or may not cover domains of ethical practice (e.g., objectivity, conflicts of interest) and ethical standards (e.g., avoiding harm, maintaining confidentiality). In addition, the focus of skills to be learned may be primarily cognitive (e.g., moral reasoning or ethical decision making) or social-interactional (e.g., ethical sensitivity). Moreover, the domain-specificity of the skills taught may differ (Perkins & Salomon, 1989; Smith, 2002). For instance, skills may be taught in a global manner, focusing on skills that apply generally to ethics across domains, or they may be covered in a domain-specific manner. Domain-specific skills are limited to ethical considerations in a specific domain, such as nursing or psychology. In addition, covering reasoning errors that hinder ethical decision making and reasoning strategies that help people work through the complexities common to real ethical problems may improve ethical decision making (Kligyte et al., 2008; Mumford et al., 2008). Thus, we examined whether there was an association between training effectiveness and the coverage of reasoning errors and/or strategies.

General Instructional Characteristics

In addition to instructional content, general characteristics of the instructional environment might moderate the effectiveness of ethics instruction. For example, organizational support for the program might impact instructional effectiveness (Baldwin & Ford, 1988; Hung & Wong, 2007), thus whether the instructional program was supported by the organization was examined. In addition, we examined whether instruction conducted in a traditional classroom setting differed from instruction administered as a separate activity in a seminar or workshop setting. In addition, the general purpose of the instructional program (e.g., standard education vs. experimental investigation) might be associated with effectiveness.

Characteristics of Instructional Methods

The delivery approach for learning experiences (e.g., opportunities for application, interaction, and involvement) proves a critical influence on instructional effectiveness (Fink, 2003). Thus,

characteristics of learning and practice activities, for instance, whether practice activities included a single type or multiple activities, were examined. In addition, participant interaction during instruction might be limited, for instance, in courses structured primarily around lecture, or extensive, such as in courses utilizing role-play activities and group discussion. The level of participant interaction is likely to influence engagement and thus learning outcomes (Slavin, 1996). We now turn to the meta-analytic procedure applied in the present study to address two overarching research questions: How effective is ethics instruction in the sciences, and What characteristics are associated with the effectiveness of ethics instruction in the sciences?

METHOD

Literature Search

To identify potential studies for inclusion in the meta-analysis, an extensive literature search was conducted. First, we identified any major review articles pertaining to ethics and ethics education in the sciences. Second, journals pertaining to ethics in research and the sciences were searched. These journals included the following: *Accountability in Research, Ethics and Behavior, Science and Engineering Ethics, Journal of Moral Education, Journal of Medical Ethics,* and *Nursing Ethics*. In addition, we identified journals associated with higher education and education in the sciences to search for additional articles pertaining to training and instruction in ethics. Some of these journals included the following: *Teaching Higher Education, Academic Medicine, Studies in Higher Education, Journal of Further and Higher Education, Teaching of Psychology, Medical Education,* and *Journal of Nursing Education*.

Following this search of specific journals, we explored major databases, such as PsycINFO, ERIC, Academic Search Elite, Blackwell-Synergy, Chronicle of Higher Education, EBSCO Collection, Health Source: Nursing/Academic Edition, LexisNexis Academic, MEDLINE, and Professional Development Collection, using targeted search terms including, but not limited to the following: "ethics training," "responsible conduct of research training," "moral development training," "ethics education," "ethics instruction," and "training and professional ethics."

After obtaining the studies identified in these searches, their reference sections were searched for additional ethics training and instruction articles that might be included in this study. To address the file drawer problem (Hunter & Schmidt, 2004; Rosenthal, 1979), we also searched Dissertation Abstracts International, a database of unpublished dissertations. In addition, we posted an announcement on the online message boards of organizations committed to ethics and responsible conduct of research training and education (e.g., Responsible Conduct of Research Education Committee). This announcement asked for individuals with available instructional evaluation data who were willing to participate in a meta-analytic study of ethics instruction effectiveness to contact us via e-mail with their data and materials describing their instructional course or program. This initial search for instructional evaluation articles and unpublished studies resulted in 140 studies that were candidates for potential inclusion in the meta-analysis.

Inclusion Criteria

Several criteria were applied to determine which studies would be included in the meta-analysis. First, each article was required to include an empirical investigation of the effectiveness of some

type of ethics education effort for scientists or researchers. Ethics education was defined as any instructional program or course, including single courses in ethics, multiple courses in a sequence covering ethics, or an entire curriculum, spread over time, that addressed scientific, research, or medical ethics. It is of note that about half of the initially identified studies were not included because they did meet the first inclusion criteria. Although these studies discussed ethics instruction or training in some fashion, they were not empirical investigations of an evaluation effort. Thus, approximately 70 remaining studies were subjected to the remaining two inclusion criteria.

Second, if the study discussed an evaluation effort, the researchers must have included, at a minimal level, descriptions of both the general instructional approach and an ethics related outcome measure. Third, and most important, the article had to report appropriate descriptive (e.g., *M*, *SD*) and/or inferential (e.g., *F*, *t*, χ^2) statistics to calculate the effect size, or *d* statistic. We utilized the effect size formulas recommended by Arthur, Bennett, and Huffcutt (2001) to calculate *d* statistics.

Before calculating the *d* statistics, the independence (or nonindependence) of data points was considered. Here, we first determined if the effect size to be computed from the reported statistics would be distinct (independent) from other effect sizes produced from the same data set. For instance, if an article produced an effect for ethical sensitivity and ethical decision making, these effects were considered independent. Second, we determined if the effect sizes from an article represented one construct or multiple constructs. For example, if an article reported multiple effects for moral reasoning (e.g., one effect for the MJS and one effect for the DIT), these effects were combined to avoid problems caused by data dependency.

In addition to determining the dependency of the data, we corrected, where possible, each effect size for measurement error. For example, where the DIT was used, we used a reliability coefficient of .76 (Rest, 1979) to correct the computed *d* statistic. As suggested by Arthur et al. (2001) and Hunter and Schmidt (2004), the formula used to correct for unreliability specified that the effect size should be divided by the square root of the criterion reliability.

Following the application of these remaining inclusion criteria and the calculation of *d* statistics, we were left with 26 independent effect sizes drawn from 20 empirical studies involving 3,041 individuals. As may be seen in the results tables, however, the total number of effect sizes (*k*) and the sample size (*N*) for the subsequent moderator analyses was typically less than that of the overall effect size estimates. This reduction in *k* and *N* across the moderator analyses can be accounted for by the fact that a number of moderator variables were uncodeable based on the information provided in the articles. Thus, all effect size estimates were included in the overall analysis of instructional effectiveness, but only those with codeable, or nonmissing data for a given moderator, were included in the subsequent moderator analyses.

Content Coding Procedure

To examine the impact of relevant instructional characteristics and study characteristics on instructional effectiveness, all of the articles were content analyzed. Three industrial and organizational psychologists, familiar with the ethics literature and the training and instructional design literature, coded the articles for the meta-analysis. Each coder received approximately 30 hr of training in the coding process and the variable set to be coded. The coders utilized a detailed glossary containing definitions of all variables to be coded for reference throughout the coding process. For all variables, coders provided a rating only if the material was explicitly discussed in the

article or could be reasonably inferred based on information provided. Otherwise, coders provided a missing data code.

After this initial introduction to the coding process, the coders made initial ratings for a set of 10 articles. Next, coders met to discuss any discrepancies in their ratings. Then, after demonstrating proficiency in the coding process, the judges coded the remaining articles independently. To ensure the accuracy of the data, the three coders held consensus meetings to discuss any discrepancies in the ratings obtained to reach consensus on their ratings. Prior to these meetings, the average interrater agreement across the seven broad coding dimensions was approximately 75%. Following these meetings, each data point entered in the analysis reflected almost complete agreement (i.e., interrater agreement of 95%). The specific variables coded in the content analysis consisted of the criterion used in the study, which yielded the obtained effect size (i.e., *d* statistic) and the sets of potential moderators of instructional effect sizes mentioned in the introduction. We describe these variables in more detail next.

Coding Criteria

Previous researchers have used different criteria to evaluate the effectiveness of ethics instructional programs. Thus, in coding the effect size estimates for each independent data point, we recorded the criterion applied to assess the effects of instruction. Due to the limited number of effect sizes available for analysis, we collapsed the criteria into two broad types. Specially, the two types were "moral development" criteria (e.g., the DIT and Kohlberg's MJS) and "ethical analysis" criteria (e.g., ethical decision making and ethical sensitivity). We also report the results specifically for the DIT and MJS, along with an aggregate of the ethical decision-making measures and an aggregate of the ethical sensitivity measures. Moreover, all criterion types were aggregated into one overall effect size estimate for instructional effectiveness. Alongside these considerations in our coding of criteria, we also coded whether the article reported the reliability of the criterion measure, which allowed us to adjust the effect sizes for unreliability.

Coding Moderators

Although the overall effect sizes arising from the meta-analysis are of interest, an examination of potential moderators that might influence these effect sizes is of critical importance. Given the limited number of effect size estimates available for analysis, analysis at the individual criterion level was not feasible. Thus, the relationship of these moderators to the observed effect sizes was computed with respect to the *overall* effect size estimate.

Study design characteristics. The primary purpose of coding characteristics of the study and design as moderators of instructional effectiveness was to examine to what extent aspects of internal validity of the studies included in the meta-analysis might, in part, account for the observed effect sizes. Specifically, first we coded for the type of design used to examine instructional effectiveness (e.g., pre–post, pre–post with control, post with control, or longitudinal). Next, we also identified the size of the sample utilized for the study and whether the author of the study served as the instructor. In addition, we coded the field of the investigator (i.e., health/medicine, philosophy, psychology, and other), the funding status of the study (funded or not funded), the

publication area of the article (ethics, health, medicine, social science, or other), and whether the publication was peer reviewed.

Participant characteristics. This set of moderator variables included characteristics associated with the individuals who participated in the instructional program. The purpose of these variables was to provide some evidence for whether the overall observed effect size for instructional effectiveness might be externally valid. More specifically, these variables provide evidence for whether the effects might generalize across different populations of people. Therefore, first we coded for the audience of the instructional program (i.e., undergraduate students, graduate/medical students, and residents/interns) and the field of study of the participants (i.e., health, medicine, psychology/counseling, or other). Here, we also coded for the participants' gender majority (male, female, or mixed) and age majority (younger than 35 years old, 35 and older, or mixed ages). Finally, we coded for whether the participants received an incentive (e.g., course credit) to complete the instructional program.

Quality ratings. Several ratings of general quality were made by the three trained content coders. The coders judged, and then rated on a 5-point Likert scale, the overall quality of the instructional program, the quality of the study design implemented to test the effectiveness of the course, and the quality of the criterion used to evaluate instructional effectiveness. The quality rating of the instructional program was judged based on an overall assessment of the quality of the content covered, use of delivery media, and the practice and application exercises utilized. The assessment of quality of the study design was based on as assessment of the adequacy of study design (e.g., sample size and inclusion of a control group). Finally, we rated the quality of the criterion measure utilized. This rating was intended to capture whether the criterion measure matched the instructional course and its intended outcomes. Thus, if a course purported to teach ethical decision-making skills but assessed outcomes via the DIT, a measure more appropriate for assessing moral development, that course received a low-criterion quality rating. The interrater reliability of these quality ratings was assessed using an intraclass correlation coefficient (ICC) and showed fairly high consistency (average ICC = .82).

Instructional content. The instructional content moderators included characteristics of course content capturing how courses in ethics might differ. Thus, these moderators provide evidence for which characteristics might lead to more or less effective instruction. First, we coded the overarching instructional objective of the program—specifically, whether the instruction focused on enhancing decision making/problem solving, moral development, or ethical sensitivity. After this general rating of instructional objective, we coded for the overall pedagogical approach. Specifically, we coded whether the instruction was primarily cognitive in nature (i.e., focusing on thinking about and solving ethical problems), or social-interactional in nature (i.e., focusing on the social and interpersonal aspects of ethical problems such as how others might react to the problem or how one's behavior might affect others). Moreover, we coded whether the types of skills learned via the course were global skills that translate to real-world ethical problems across domains and settings or specific skills that are limited to the domain (e.g., nursing) at hand.

In addition to these broader elements of instructional content, we coded a number of specific elements of the instructional content. Our review of the literature uncovered a number of ethical domains, behaviors, and standards that might be discussed in ethics instruction. Thus, we coded for whether the courses included these elements. For example, ethical domains included responsi-

bility, objectivity and fairness, mentor–mentee relationships, conflicts of interest, and peer review and publication. The ethical behaviors taxonomy included, for example, coverage of appropriate data management, informed consent, treatment of human participants and animal subjects, protection of intellectual property, protection of public welfare and environment, fair treatment of staff and collaborators, and appropriate use of physical resources (Helton-Fauth et al., 2003). Coverage of ethical standards included coding for whether courses included material on ethical values considered central to ethics as a researcher or scientist (e.g., avoiding harm, maintaining confidentiality, avoiding personal gain, and confronting ethical issues).

In addition, recent research has stressed the importance of covering common reasoning errors that might be encountered in ethical decision making and strategies for dealing with these errors and the social-cognitive complexities of ethical problems (Kligyte et al., 2008; Mumford et al., 2008). Therefore, we coded whether typical reasoning errors (e.g., personal biases, thinking in simplistic terms) and strategies (e.g., perspective taking, emotion regulation, self-skepticism) were covered in the course. Because each individual content variable was covered only intermittently across courses, after coding for specific characteristics, we had to collapse the coding according to whether *any* elements of these instructional content areas were covered in the ethics course. Thus, we used a present/not present approach to coding these instructional content variables.

General instructional characteristics. This set of moderating variables consisted of more general characteristics associated with ethics courses. These variables included the setting for instruction (e.g., academic classroom or workshop/seminar), whether the organization actively advocated the program by providing resources and encouraging participation, whether the instructional program was mandatory, and the purpose of the program (standard education, professional development, or experimental investigation). Finally, we also coded for whether the course was integrated into the curriculum (e.g., an ethics section in a regular course) or whether it was a stand-alone course (e.g., an ethics course taken by medical students separately from regular coursework).

Characteristics of instructional method. In the final set of moderators, specific instructional methods, such as learning activities and instructional media, were examined. Specifically, this coding dimension included the length of instruction (less than 9 hr, or equal to or greater than 9 hr), the primary delivery method utilized (e.g., traditional classroom approach or case-based approach), the type of learning methods employed (variable or constant), and the type of practice sessions utilized (massed or distributed). We also coded the different types of learning activities used, for example, case-based exercises, essay or diary entries, face-to-face discussion, lecture, textbook readings, and role-plays. Because we could not analyze each type of activity separately because of the lack of consistency in activity use, we collapsed this coding dimension into the number of types of learning activities utilized (less than or equal to 3, or equal to or greater than 4). We coded practice exercises for use of a single type, multiple types, or none. Finally, the level of participant interaction during learning (low, moderate, or high) was coded.

Analysis Plan

Using the procedures recommended by Arthur et al. (2001), we used a SAS PROC MEANS program to conduct the analyses based on the meta-analytic approach recommended by Hunter and

Schmidt (1990). This approach allowed sample-weighted means to be computed, which were corrected for sampling error.

In deciding to test moderators of instructional effectiveness, we employed the 75% rule-of-thumb suggested by Hunter and Schmidt (2004). Thus, if the overall meta-analysis resulted in less than 75% of the variance in studies being accounted for by sampling error (i.e., if correcting for statistical artifacts did not account for nearly all of the observed variation in effect sizes across studies), then there was reason to suspect that the effect size estimates were dependent on moderators. When conducting moderator analyses, most scholars (Arthur et al., 2001; Hunter & Schmidt, 2004) suggest that for best results, analyses should be limited to situations in which large samples of studies are available (i.e., $k \geq 10$). However, based on the already limited k, we examined moderators if there were at least two cases available; this approach is consistent with Arthur and colleagues' (2001). Nonetheless, any interpretation made from analyses with such limited k size should be made with caution.

In our analysis, we calculated 95% confidence intervals for the sample-weighted mean effect sizes (*Md*s). The confidence interval provides an indication of the accuracy of the estimate of the mean effect size by representing the extent to which sampling error may remain in the sample-weighted *Md*. More specifically, the confidence interval provides a range of values that the mean effect size is likely to take if other studies from the population were to be used in the meta-analysis. Furthermore, fail-safe *N* statistics were calculated to provide an estimate of the number of null effect sizes required to reduce a particular *Md* to below .20 (Orwin, 1983). In examining the results of the meta-analysis, it is of note that the analysis occasionally yielded confidence intervals with zero range (e.g., .11–.11). This finding reveals that all of the observed variance in the *Md* for that moderator analysis was due to sampling error; thus, after correcting for sampling error, no additional variance in the effect size estimates remains to be moderated.

RESULTS

Overall Effectiveness

The results of the overall meta-analysis are presented in Table 1. We applied Cohen's (1969, 1992) recommendations for the interpretation of effect size magnitude. More specifically, when $d = .20$, this is considered a small effect; when $d = .50$, this is considered a medium effect; and when $d = .80$, this is considered a large effect. As may be seen in Table 1, the overall instructional effectiveness of ethics courses was moderate, $d = .42$ ($SD = .27$). However, the percentage of variance accounted for by sampling error was also small (33%), thus we investigated the presence of moderators.

Before turning to the moderator results, of note are the findings with respect to the criterion type. Specifically, moral development ($d = .36$, $SD = .26$) and ethical analysis criteria ($d = .61$, $SD = .16$) revealed differences in effect sizes. We also found that when correction for reliability was possible in the calculation of the d statistic, the effect size was only slightly larger ($d = .43$, $SD = .29$) than when it was not possible ($d = .37$, $SD = .00$). Thus, this difference is of limited concern given that it is small and that only three effect sizes were not able to be corrected for the reliability of the measure.

TABLE 1
Overall Meta-Analysis and Criterion Type

	k	*N*	*Sample Weighted Md*	*SD*	*Variance Due to Sampling Error (%)*	*95% CI*		χ^2	N_{fs}
						L	*U*		
Ethics instruction effectiveness									
Overall meta-analysis	26	3,041	.42	.27	33	–.10	.95	78.41	29
General criterion type									
Moral development	17	2,229	.36	.26	31	–.16	.88	55.20	14
Ethical analysis[a]	9	812	.61	.16	65	–.29	.93	13.97	18
Specific criterion measures									
MJS	4	106	–.14	.27	67	–.67	.38	5.74	—
DIT	13	2,123	.38	.24	31	–.09	.85	42.52	12
Ethical sensitivity[b]	6	701	.58	.20	48	.19	.97	12.44	11
Ethical decision making[c]	3	111	.77	.00	100	.77	.77	.80	9
Reliability corrected									
No	3	346	.37	.00	100	.37	.37	.68	3
Yes	23	2,695	.43	.29	30	–.14	1.00	77.43	26

Note. The dash indicates that the effect size is already below .20. *k* = number of effect sizes; *Md* = sample weighted mean effect size (*d*) corrected for measurement error; *SD* = standard deviation of mean effect size; CI = confidence interval; L = lower; U = upper; N_{fs} = Orwin's (1983) Fail safe *N* (number of null effects to reduce *Md* below .20).

[a]Ethical decision making and ethical sensitivity combined. [b]All ethical sensitivity measures combined because of limited sample size. [c]All ethical decision making measures combined because of limited sample size.

Effects of Moderating Variables

Study design characteristics. The results with respect to study and design characteristics are presented in Table 2. We found that effect size estimates differed based on the type of design used in the study. A posttest with control design yielded the largest effects (*d* = .68, *SD* = .13) followed by a pretest, posttest design (*d* = .52, *SD* = .00); longitudinal design (*d* = .39, *SD* = .21); and pretest, posttest with control (*d* = .35, *SD* = .31). This finding highlights the point that conclusions derived from studies of instructional effectiveness are, at least in part, contingent on the type of study design utilized. Of note is the finding that, if the author conducted the ethics course, effect sizes were larger (*d* = .61, *SD* = .12) compared to when the instructor was not the author (*d* = .29, *SD* = .38). The investigator's field was also associated with the obtained effect sizes. The largest effects were observed for investigators in the field of psychology (*d* = .80, *SD* = .00), followed by philosophy (*d* = .54, *SD* = .21), and the health and medical fields (*d* = .38, *SD* = .17). Moreover, the publication outlet was also associated with effects, such that studies published in the social sciences (*d* = .78, *SD* = .00) had the largest effects, followed by ethics (*d* = .58, *SD* = .25), health (*d* = .44, *SD* = .00), and medicine (*d* = .00, *SD* = .23).

Participant characteristics. The moderating effects of participant characteristics are presented in Table 3. It was found that when residents and interns were the audience for instruction in ethics, the largest effects (*d* = .66, *SD* = .18) were produced, followed by undergraduate students (*d* = .40, *SD* = .22) and graduate/medical students (*d* = .33, *SD* = .42). Thus, the audience of the instructional program seems to influence the effectiveness of the course. Perhaps greater

TABLE 2
Study Design Characteristics

	k	*N*	*Sample Weighted Md*	*SD*	*Variance due to Sampling Error (%)*	*95% CI*		χ^2	N_{fs}
						L	*U*		
Design type									
Pre–post w/ control	12	1,380	.35	.31	27	–.27	.96	44.69	9
Longitudinal	5	1,027	.39	.21	33	–.01	.79	15.50	5
Pre–post	3	160	.52	.00	100	.52	.52	.12	5
Post w/ control	6	474	.68	.13	76	.43	.94	7.90	14
Sample size									
Less than 50	7	153	.23	.00	100	.23	.23	5.14	1
50–100	8	526	.52	.46	23	–.38	1.42	34.12	13
101–150	5	642	.29	.30	26	–.31	.88	19.40	2
151+	6	1,720	.46	.13	47	.21	.71	12.63	8
Author served as instructor									
No	13	1,122	.29	.38	25	–.46	1.04	52.67	6
Yes	10	884	.61	.12	77	.37	.85	13.07	21
Investigator field									
Health/Medicine	9	1,313	.38	.17	49	.05	.72	18.42	8
Philosophy	4	648	.54	.21	38	.13	.94	10.52	7
Psychology	6	441	.80	.00	100	.80	.80	4.95	18
Other	7	639	.13	.21	50	–.29	.55	14.08	—
Study funded									
No	14	1,680	.46	.20	47	.07	.85	29.79	18
Yes	10	1,167	.42	.27	33	–.10	.94	29.89	11
Publication outlet									
Medicine	5	170	.00	.23	70	–.46	.45	7.16	—
Health	5	1,155	.44	.00	100	.44	.44	4.01	6
Ethics	6	680	.58	.25	38	.10	1.06	15.72	11
Social science	4	373	.78	.00	100	.78	.78	3.92	11
Other	6	663	.14	.20	49	–.24	.53	12.25	—
Publication type									
Nonpublished	5	396	–.02	.09	87	–.19	.15	5.74	—
Peer reviewed	20	2,512	.49	.23	39	.04	.93	50.99	29

Note. The dash indicates that the effect size is already below .20. *k* = number of effect sizes; *Md* = sample weighted mean effect size (*d*) corrected for measurement error; *SD* = standard deviation of mean effect size; CI = confidence interval; L = lower; U = upper; N_{fs} = Orwin's (1983) Fail safe *N* (number of null effects to reduce *M d* below .20).

experience in one's field, as is the case for residents and interns, promotes the effectiveness of ethics instruction (Ericsson & Charness, 1994). Along related lines, when the participants consisted of mixed ages (i.e., participants both older and younger than 35 years of age) the effect was larger (*d* = .45 *SD* = .46) than when the majority of participants were younger than 35 (*d* = .22, *SD* = .40). However, the variance in these effects was quite large, as indicated by the standard deviations, and somewhat unstable, as reflected in their respective confidence intervals. Also of note is the finding with respect to the participants' field of study. Similar to the findings for the field of investigator, when the participant's field of study was social sciences (i.e., psychology or counseling), the effect size was largest (*d* = .66, *SD* = .29). When trainees were in health fields, instruction showed greater effects (*d* = .44, *SD* = .00) than when participants were

TABLE 3
Participant Characteristics

	k	N	Sample Weighted Md	SD	Variance Due to Sampling Error (%)	95% CI L	95% CI U	χ^2	N_{fs}
Audience									
Graduate/Medical students	9	399	.33	.42	36	–.49	1.15	25.27	6
Undergraduate students	13	2,264	.40	.22	33	–.03	.83	39.66	13
Residents/Interns	4	378	.66	.18	30	.30	1.01	6.87	9
Field of study									
Medicine	6	185	.05	.25	70	–.42	.53	8.61	—
Health	5	1,155	.44	.00	100	.44	.44	4.01	6
Psychology/Counseling	8	535	.66	.29	43	.09	1.23	18.48	18
Other	4	685	.24	.24	29	–.23	.71	13.78	1
Participant gender									
70% male	3	347	.02	.13	70	–.23	.26	4.34	—
Mixed gender	3	104	.27	.00	100	.27	.27	.12	1
70% female	9	787	.40	.35	28	–.29	1.09	32.64	9
Participant age									
70% less than 35	7	274	.22	.40	41	–.55	1.00	17.16	1
Mixed ages	4	142	.45	.46	37	–.45	1.34	10.89	5
Participants had incentive									
No	7	1,182	.41	.00	100	.41	.41	6.36	7
Yes	16	1,545	.36	.34	27	–.30	1.03	59.35	13

Note. The dash indicates that the effect size is already below .20. *k* = number of effect sizes; *Md* = sample weighted mean effect size (*d*) corrected for measurement error; *SD* = standard deviation of mean effect size; CI = confidence interval; L = Lower; U = Upper; N_{fs} = Orwin's (1983) Fail safe *N* (number of null effects to reduce *Md* below .20).

in medicine (*d* = .05, *SD* = .25) or other fields (*d* = .24, *SD* = .24). These findings also revealed that, when the participants were a female majority, the effects were largest (*d* = .40, *SD* = .35) compared to a majority of male participants (*d* = .02, *SD* = .13) or mixed gender (*d* = .27, *SD* = .00). However, this finding should be interpreted with caution given the limited number of effect sizes that could be included in this moderator analysis because of inadequate reporting of participant gender across studies.

Quality ratings. The findings with respect to the quality ratings are provided in Table 4. Not surprisingly, as quality ratings increased, the magnitude of the observed effect sizes increased. For quality of the instructional program, above-average instructional programs demonstrated much larger effects sizes (*d* = .72, *SD* = .15) than average (*d* = .39, *SD* = .02) or below average (*d* = .18, *SD* = .25) programs. Similarly, when the quality of the *study* was below average (*d* = .16, *SD* = .23), effect sizes were much smaller than when the study quality was average (*d* = .48, *SD* = .20) or above average (*d* = .65, *SD* = .20). Finally, when the quality of the *criterion* was average (*d* = .51, *SD* = .08) or above average (*d* = .57, *SD* = .20), effect sizes were larger than when the criterion used for assessing instructional effectiveness was below average (*d* = .29, *SD* = .34). Thus, as expected, the effectiveness of ethics instruction is contingent on the quality of the instruction itself, the quality of the study designed to examine the effects of instruction, and the quality of the criterion applied to evaluate the course.

TABLE 4
Quality Ratings

	k	*N*	*Sample Weighted Md*	*SD*	*Variance Due to Sampling Error (%)*	*95% CI*		χ^2	N_{fs}
						L	*U*		
Quality rating of instructional program									
Below average	5	674	.18	.25	32	–.32	.68	15.67	—
Average	8	1,301	.39	.02	99	.35	.42	8.11	8
Above average	6	445	.72	.15	72	.42	1.01	8.30	16
Quality rating of study									
Below average	10	817	.16	.23	50	–.29	.60	20.13	—
Average	12	1,688	.48	.20	44	.10	.86	27.41	17
Above average	4	536	.65	.20	45	.26	1.03	8.83	9
Quality rating of criterion									
Below average	14	1,284	.29	.34	28	–.38	.95	50.22	6
Average	9	1,375	.51	.08	80	.35	.68	11.28	14
Above average	3	382	.57	.20	45	.17	.96	6.71	6

Note. The dash indicates that the effect size is already below .20. *k* = number of effect sizes; *Md* = sample weighted mean effect size *d* corrected for measurement error; *SD* = standard deviation of mean effect size; CI = confidence interval; L = Lower; U = Upper; N_{fs} = Orwin's (1983) Fail safe *N* (number of null effects to reduce *M d* below .20).

Instructional content. The results obtained when instructional content was examined as a moderator of instructional effectiveness are provided in Table 5. An examination of the overarching instructional objective revealed that an ethical decision-making/problem-solving (*d* = .52, *SD* = .15) approach was most effective, followed by ethical sensitivity (*d* = .42, *SD* = .11) and moral development (*d* = .17, *SD* = .28). In addition, courses that focused on skills applicable to ethics in a global sense (i.e., focusing on ethical problems encountered in a number of real-world settings that span across domains and fields) were more effective (*d* = .64, *SD* = .11) than courses that focused on limited skills that are only specific to a particular field (*d* = .35, *SD* = .27). In addition, when the general approach to instruction was cognitive in nature, the effect size was slightly larger (*d* = .44, *SD* = .22) than social-interactional approaches (*d* = .37, *SD* = .22). This finding suggests that cognitive approaches may be most effective but that social-interactional approaches may also be of value. Moreover, these findings are consistent with those obtained for overarching instructional objective, such that decision-making, a cognitive approach, and sensitivity, a social-interactional approach, were both fairly effective.

With respect to ethical domains (e.g., mentor–mentee relationships, authorship and publication), coverage of these key domains was associated with much higher effectiveness (*d* = .48, *SD* = .14) compared to no inclusion of these domains in the instruction (*d* = -.11, *SD* = .14). Similar findings emerged for coverage of ethical behaviors (e.g., maintaining confidentiality, protection of intellectual property) and ethical standards for conducting research and science (e.g., avoiding harm and avoidance of personal gain). Courses covering behaviors and standards revealed larger effect sizes (*d* = .52, *SD* = .17) than those that did not include this material (*d* = .12, *SD* = .13).

Finally, with regard to inclusion of possible reasoning errors in ethical decision making (e.g., thinking in black-and-white terms, making hasty decisions, failing to weigh future consequences), courses that covered this material showed larger effects (*d* = .57, *SD* = .30) compared to courses that did include this information (*d* = .33, *SD* = .17). Furthermore, courses that included strategies

TABLE 5
Instructional Content

	k	N	Sample Weighted Md	SD	Variance Due to Sampling Error (%)	95% CI L	95% CI U	χ^2	N_{fs}
Overarching instructional objective									
Moral development	4	619	.17	.28	25	–.38	.73	16.18	—
Ethical sensitivity	7	780	.42	.11	74	.19	.64	9.48	8
Decision making/Problem solving	9	1,234	.52	.15	59	.23	.81	15.31	14
Overarching instructional approach									
Social-interactional	6	938	.37	.22	35	–.06	.80	16.92	5
Cognitive	12	1,366	.44	.22	43	.01	.87	28.06	14
Type of skills instructed									
Domain specific	15	2,257	.35	.27	27	–.18	.89	55.97	11
Global	9	744	.64	.11	81	.43	.86	11.12	20
Ethical domains coverage									
No	4	460	–.11	.14	66	–.38	.15	6.06	—
Yes	16	2,014	.48	.14	62	.20	.76	26.00	22
Ethical behaviors taxonomy									
No	6	579	.03	.00	100	.03	.03	5.07	—
Yes	12	1,829	.51	.15	55	.21	.80	21.75	19
Ethical standards coverage									
No	6	1,192	.20	.25	24	–.29	.71	24.98	—
Yes	14	1,282	.52	.19	57	.16	.88	24.38	22
Problems in EDM coverage									
No	10	1,845	.33	.17	43	.00	.67	23.33	7
Yes	9	575	.57	.30	43	–.01	1.16	20.91	17
Strategies for EDM coverage									
No	7	1,229	.22	.25	27	–.27	.71	25.93	1
Yes	13	1,245	.52	.20	53	.13	.90	24.20	21

Note. The dash indicates that the effect size is already below .20. *k* = number of effect sizes; *Md* = sample weighted mean effect size *d* corrected for measurement error; *SD* = standard deviation of mean effect size; CI = confidence interval; L = Lower; U = Upper; N_{fs} = Orwin's (1983) Fail safe *N* (number of null effects to reduce *Md* below .20); EDM = ethical decision making.

that can be used by scientists to assist them in working through ethical problems (e.g., asking for help from someone with an outside perspective, considering the perspectives of others, and managing one's own emotions) revealed larger effects ($d = .52$, $SD = .20$) relative to those that did not include such instruction on strategies ($d = .22$, $SD = .25$). Thus, it appears that including content that focuses on how one might address ethical problems and work through decisions may improve ethics instruction effectiveness.

General instructional characteristics. The findings with respect to general instructional characteristics as moderators of instructional effectiveness are presented in Table 6. Most notably, we found that instructional courses conducted in a separate workshop or seminar format ($d = .52$, $SD = .02$), as compared to being held in a typical academic (i.e., classroom) setting ($d = .38$, $SD = .26$), were more effective. Similarly, courses held in a stand-alone fashion focusing solely on ethics ($d = .51$, $SD = .30$), instead of being embedded in existing courses or curriculum ($d = .37$, $SD =$

TABLE 6
General Instructional Characteristics

	k	*N*	*Sample Weighted Md*	*SD*	*Variance Due to Sampling Error (%)*	*95% CI*		χ^2	N_{fs}
						L	*U*		
Setting of instruction									
Academic setting	19	2,532	.38	.26	31	–.14	.89	61.31	17
Workshop/Seminar	4	195	.52	.02	99	.48	.56	4.02	6
Organization advocates program									
No	11	874	.42	.28	41	–.13	.96	26.87	12
Yes	11	1,838	.37	.25	28	–.12	.86	38.68	9
Instructional program mandatory									
No	10	909	.53	.25	43	.04	1.02	23.51	17
Yes	13	1,818	.32	.23	36	–.13	.76	35.97	8
Primary purpose of program									
Education	19	2,577	.36	.26	31	–.14	.89	61.33	15
Professional development	3	175	.73	.29	48	.17	1.30	6.25	8
Basic experimentation	2	71	.85	.00	100	.85	.85	.57	7
Type of instructional program									
Integrated	11	1,832	.37	.22	33	–.07	.80	33.08	9
Stand-alone	15	1,209	.51	.30	36	–.08	1.11	41.30	23

Note. The dash indicates that the effect size is already below .20. *k* = number of effect sizes; *Md* = Sample weighted mean effect size *d* corrected for measurement error; *SD* = standard deviation of mean effect size; L = Lower; U = Upper; N_{fs} = Orwin's (1983) Fail safe *N* (number of null effects to reduce *Md* below .20).

.22) showed greater effectiveness. In addition, courses that were conducted for the purpose of professional development (*d* = .73, *SD* = .29) showed large effects compared to the modest effects of instruction conducted solely for educational purposes (*d* = .36, *SD* = .26). Experimental courses conducted solely for the purpose of research showed the largest effects (*d* = .85, *SD* = .00); however, there were only two studies included in this analysis. It was also found that instruction that was not mandatory (*d* = .53, *SD* = .25) was more effective than instruction that was required (*d* = .32, *SD* = .23).

Characteristics of instructional methods. The last set of moderators concerned characteristics of instructional methods. These results are shown in Table 7. A case-based approach to instruction yielded larger effects (*d* = .53, *SD* = .14) than a standard lecture approach (*d* = .36, *SD* = .25). Furthermore, the application of a variable learning method (*d* = .52, *SD* = .09), where learning activities (e.g., discussion, cases, journaling, lecture) vary throughout instruction, yielded larger effects than constant learning methods (*d* = .18, *SD* = .24), where a single learning activity is utilized throughout instruction. Along similar lines, using four or more learning activities (*d* = .48, *SD* = .14) compared to three or fewer (*d* = .12, *SD* = .34) revealed greater instructional effectiveness. Furthermore, those courses that applied practice techniques in a distributed approach throughout the instructional course (*d* = .47, *SD* = .18), as opposed to practicing in one massed session (*d* = .18, *SD* = .33) were more effective. In a related vein, multiple types of practice activities (*d* = .52, *SD* = .12) were more effective than a single type of practice activity (*d* = .18, *SD* = .29). Finally, courses that allowed for greater trainee interaction during learning and practice activities

TABLE 7
Characteristics of Instructional Methods

	k	*N*	*Sample Weighted Md*	*SD*	*Variance Due to Sampling Error (%)*	*95% CI* L	*95% CI* U	χ^2	N_{fs}
Length of instruction									
Less than 9 hr	9	867	.35	.25	41	–.13	.84	21.84	7
Equal to or greater than 9 hr	13	1,867	.47	.24	33	.00	.94	38.86	18
Primary delivery method									
Classroom based	8	1,091	.36	.25	33	–.12	.85	24.44	6
Case based	9	1,214	.53	.14	60	.24	.81	14.98	15
Other	3	328	.11	.00	100	.11	.11	1.52	—
Learning method									
Constant	9	926	.18	.24	41	–.29	.64	21.77	—
Variable	10	1,494	.52	.09	76	.35	.70	12.89	16
Learning activity usage									
Less than or equal to 3	8	692	.12	.34	29	–.55	.78	27.63	—
Equal to or greater than 4	13	1,995	.48	.14	59	.21	.75	22.17	18
Practice									
Massed	6	382	.18	.33	38	–.46	.82	15.77	—
Distributed	11	1,655	.47	.18	47	.12	.82	23.37	15
Practice activities									
Single type	8	543	.18	.29	42	–.40	.75	19.20	—
Multiple types	8	1,442	.52	.12	62	.29	.75	12.83	13
None	3	435	.23	.11	71	.02	.44	4.24	0
Level of participant interaction									
Low	4	411	.05	.10	81	–.14	.24	4.91	—
Moderate	6	1,198	.37	.16	43	.05	.70	13.80	5
High	7	722	.63	.09	84	.45	.81	8.32	15

Note. The dash indicates that the effect size is already below .20. *k* = number of effect sizes; *Md* = sample weighted mean effect size *d* corrected for measurement error; *SD* = standard deviation of mean effect size; CI = confidence interval; L = Lower; U = Upper; N_{fs} = Orwin's (1983) Fail safe *N* (number of null effects to reduce *M d* below .20).

were more effective (*d* = .63, *SD* = .09) than courses with moderate (*d* = .37, *SD* = .16) or low levels of trainee interaction (*d* = .05, *SD* = .10).

DISCUSSION

Before turning to the conclusions and implications arising from these findings, several limitations of this study must first be noted. To begin, these meta-analytic findings should be interpreted with some caution given the limited number of studies included in the meta-analysis. Although a large number of studies discussing ethics instruction in general were identified, few studies explicitly *evaluated* ethics instruction in the sciences. Furthermore, after applying the inclusion criteria for evaluation efforts, several studies could not be included because they lacked descriptiveness, or simply because basic statistics, such as standard deviations, necessary for calculation of the *d* statistic were not reported. Nevertheless, these basic statistics are considered essential for reporting the results of empirical research, particularly when studying human subjects (Wilkinson & the

Task Force on Statistical Interference, 1999). Thus, these observations suggest a great deal of interest in ethics instruction, but limited systematic, rigorous evaluation of ethics instruction.

In addition, moderator data could not be provided for every effect size included in the analysis. Instead, coding of moderators was limited to the descriptiveness provided within the studies, which was unfortunately often less than ideal. As a result, conclusions arising from the analyses of moderators of instructional effectiveness were often based on a sample of studies smaller than the overall effect size analysis.

Along similar lines, because of limited sample size, examination of moderators at the level of each individual variable within a particular dimension was not possible. Instead, moderator analyses were typically dichotomized, although three to four unique categories could be created for some variables. Unfortunately, this approach of collapsing into broader dimensions limits the specificity of our conclusions. For example, we cannot conclude which specific types of learning activities (i.e., lecture, group discussion, and journaling) are most effective and how they compare to one another. Moreover, every possible moderator that might influence instructional effectiveness could not be coded for in this study. However, focus was placed on identifying key factors of instructional programs that might represent differences in effectiveness.

Even taking these limitations into consideration, the findings obtained in the present study suggest some noteworthy conclusions regarding the effectiveness of ethics instruction in the sciences and, furthermore, point to issues to be considered in the design and evaluation of ethics courses. To begin, we return to our first question: How effective is ethics instruction in the sciences? The answer appears to be that ethics instruction is at best moderately effective as it is currently conducted. Not surprisingly, however, the findings also suggest that, when the instructional program quality is high, effectiveness is greater. Therefore, it appears that if instructional programs are well designed, they have the potential to be fairly effective. Hence, we posed our second question: What characteristics are associated with the effectiveness of ethics instruction in the sciences? A response to this question emerges from an examination of the pattern of findings with respect to characteristics associated with larger effects.

First and foremost, it appears that a cognitive decision-making approach to instruction is most effective, followed by ethical sensitivity, which focuses on the social-interactional nature of ethical problems. In fact, a relevant question for future research might be whether, in combination, these approaches would complement one another, producing even higher levels of instructional effectiveness (Sonenshein, 2007). Along these lines, it was found that covering potential reasoning errors (e.g., thinking in black-and-white terms, making hasty decisions, and overlooking key causes) that might hinder thinking through ethical situations was especially valuable. Furthermore, providing cognitive strategies (e.g., considering others' perspectives, considering personal motivations, and anticipating consequences) for thinking through the likely outcomes and social implications of the problem was also especially beneficial for instructional effectiveness. Indeed, it has been shown in past studies that strategy-based instructional interventions are particularly effective for improving people's problem-solving on complex, ambiguous problems (Scott, Leritz, & Mumford, 2004a, 2004b).

In addition, it should be noted that providing specific content, such as ethical domains, standards, and behaviors, appears to be important for constructing effective ethics instruction. This material may provide a basis for framing what constitutes an ethical problem, and thus for applying newly learned strategies for working through these problems (Gick & Holyoak, 1983). In fact, it seems that older participants may benefit the most from ethics instruction. It may be that they

possess the requisite knowledge to serve as a foundation for strategy-based training (Clapham, 1997; Kolodner, 1997; Önkal, Yates, Simga-Mugan, & Öztin, 2003). Thus, it appears that the success of ethics instruction may, in part, be attributed to developing an understanding of, and providing guidance concerning, the application of requisite strategies for confronting real-world ethical problems.

In addition to foundational knowledge provided by covering ethical standards and behaviors, cases also provide knowledge, in the form of contextualized exemplars (Hammond, 1990; Kolodner, 1993, 1997; Patalano & Seifert, 1997), to be applied for addressing ethical problems. Moreover, case examples provide a learning tool for practicing to apply relevant knowledge and strategies to problem scenarios (Erickson & Kruschke, 1998; Jonassen & Hernandez-Serrano, 2002). In line with these observations, case-based instruction produced larger effects than classroom-based, lecture-style instruction. Moreover, student engagement, by means of highly interactive courses and a number of different learning and practice activities, also promoted instructional effectiveness. In fact, because case-based models are often acquired through social experiences, the effectiveness of case-based approaches to instruction is likely to be enhanced by interactive, cooperative learning (Aronson & Patnoe, 1997; Slavin, 1991). These observations regarding the criticality of interaction among course participants are particularly significant given the current trend towards online formats for ethics instruction (Barnes et al., 2006; Braunschweiger & Goodman, 2007; Kiser, 1999). Although online courses have their advantages, they typically fail to involve any degree of social interaction, limiting training to individual-level application of rules to relatively simple, context-free cases. This all-too-common limitation must be considered seriously as institutions consider the best ways to implement effective ethics instruction, whether it is delivered online or face-to-face. Moreover, given the implicit aim of ethics instruction—to foster a community of social responsibility—it seems reasonable to expect that a learning environment that involves social interaction might facilitate such a goal better than one that does not involve social interaction. Whether this expectation is true, of course, is a question that remains open to empirical investigation.

Before turning to our final conclusions, it is important to reiterate the critical importance of careful, thorough evaluation of ethics instruction. The design of the evaluation study must be as complete and systematic as possible, and the criterion measure must match the intended outcomes of the program (Alliger, Tannenbaum, Bennett, Traver, & Shotland, 1997; Kraiger & Jung, 1996). In fact, to reach a coherent understanding of what might constitute effective ethics instruction, a fundamental consideration includes the most appropriate criterion measure to be utilized in evaluation. In the present study, the DIT (Rest, 1979) was the most commonly applied criterion measure in studies of ethics instruction. If, as the present study suggests, less effective programs use a moral development framework and more effective programs involve instruction in the process of ethical decision making—specifically, learning about social-cognitive elements of ethical problems and the application of strategies for working through problems—then it might be necessary to consider whether a different measure of training effectiveness might more completely assess whether these instructional goals have been accomplished. In fact, the DIT, a measure of moral development, may be limited in its ability to address all potential, and desired, outcomes of instruction.

In conclusion, although this study points to several important considerations for the design and delivery of ethics instruction, these recommendations are certainly not suggested as a panacea for ethics instruction. The present study merely skims the surface of a rather extensive issue still requiring a great deal of research. Fortunately, we did find evidence that ethics instruction in the sci-

ences, if carefully designed and evaluated, has the potential to be fairly effective. We hope that the present study will provide some practical guidance for future course development and evaluation. Moreover, we hope that the present effort might provide direction for researchers generally concerned with studying ethics and ethics instruction.

ACKNOWLEDGMENTS

We thank Jason Hill and Jared Caughron for their contributions to this research. This research was supported by grant #5R01-NS049535-02 from the National Institutes of Health and the Office of Research Integrity, Michael D. Mumford, Principal Investigator.

REFERENCES

*References marked with an asterisk indicate studies included in the meta-analysis.

Abbott, A. (1999, March). Science comes to terms with the lessons of fraud. *Nature, 398*, 13–17.

Alliger, G. M., Tannenbaum, S. I., Bennett, W., Jr., Traver, H., & Shotland, A. (1997). A meta-analysis of the relations among training criteria. *Personnel Psychology, 50*, 341–358.

Antes, A. L., Brown, R. P., Murphy, S. T., Waples, E. P., Mumford, M. D., Connelly, S., et al. (2007). Personality and ethical decision-making in research: The role of perceptions of self and others. *Journal of Empirical Research on Human Research Ethics, 2*, 15–34.

Aronson, E., & Patnoe, S. (1997). *The jigsaw classroom: Building cooperation in the classroom* (2nd ed.). New York: Longman.

Arthur, W., Jr., Bennett, W., Jr., & Huffcutt, A. I. (2001). *Conducting meta-analysis using SAS*. Mahwah, NJ: Erlbaum.

*Baldick, T. L. (1980). Ethical discrimination ability of intern psychologists: A function of training in ethics. *Professional Psychology, 11*, 276–282.

Baldwin, T. T., & Ford, J. K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology, 41*, 63–105.

Barnes, B. E., Friedman, C. P., Rosenberg, J. L., Russell, J., Beedle, A., & Levine, A. S. (2006). Creating an infrastructure for training in the responsible conduct of research: The University of Pittsburgh's experience. *Academic Medicine, 81*, 119–127.

*Bebeau, M. J., & Thoma, S. J. (1994). The impact of a dental ethics curriculum on moral reasoning. *Journal of Dental Education, 58*, 684–692.

Borkowski, S. C., & Ugras, Y. J. (1998). Business students and ethics: A meta-analysis. *Journal of Business Ethics, 17*, 1117–1127.

Braunschweiger, P., & Goodman, K. W. (2007). The CITI program: An international online resource for education in human subjects protection and the responsible conduct of research. *Academic Medicine, 82*, 861–864.

*Chase, N. M. (1999). A cognitive-development approach to professional ethics training for counselor education students. *Dissertation Abstracts International, 59*(08), 2865. (UMI No. 9903261)

Clapham, M. M. (1997). Ideational skills training: A key element in creativity training programs. *Creativity Research Journal, 10,* 33–44.

Clarkeburn, H. (2002). A test for ethical sensitivity in science. *Journal of Moral Education, 31*, 439–453.

*Clarkeburn, H., Downie, J. R., & Matthew, B. (2002). Impact of an ethics programme in a life sciences curriculum. *Teaching in Higher Education, 7*, 65–79.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

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

Colquitt, J. A., & Simmering, M. J. (1998). Conscientiousness, goal orientation, and motivation to learn during the learning process: A longitudinal study. *Journal of Applied Psychology, 83*, 654–665.

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

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings.* Chicago: Rand McNally.

Dalton, R. (2000). NIH cash tied to compulsory training in good behaviour. *Nature, 408*, 629.

*Drake, M. J., Griffin, P. M., Kirkman, R., & Swann, J. L. (2005). Engineering ethical curricula: Assessment and comparison of two approaches. *Journal of Engineering Education, 94*, 223–232.

*Duckett, L., Rowan, M., Ryden, M., Krichbaum, K., Miller, M., Wainwright, H., et al. (1997). Progress in the moral reasoning of baccalaureate nursing students between program entry and exit. *Nursing Research, 46*, 222–229.

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist, 49*, 725–747.

Erickson, M. A., & Kruschke, J. K. (1998). Rule and exemplars in category learning. *Journal of Experimental Psychology: General, 127*, 107–140.

Fink, L. D. (2003). *Creating significant learning experiences: An integrated approach to designing college courses*. San Francisco: Jossey-Bass.

Friedman, P. J. (2002). The impact of conflict of interest on trust in science. *Science and Engineering Ethics, 8*, 413–420.

*Frisch, N. C. (1987). Value analysis: A method for teaching nursing ethics and promoting the moral development of students. *Journal of Nursing Education, 26*, 328–332.

*Gaul, A. L. (1987). The effect of a course in nursing ethics on the relationship between ethical choice and ethical action in baccalaureate nursing students. *Journal of Nursing Education, 26*, 113–117.

*Gawthrop, J. C., & Uhlemann, M. R. (1992). Effects of the problem-solving approach in ethics training. *Professional Psychology: Research & Practice, 23*, 38–42.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1–38.

*Goldman, S. A., & Arbuthnot, J. (1979). Teaching medical ethics: The cognitive-developmental approach. *Journal of Medical Ethics, 5*, 170–181.

Goldstein, I. L., & Ford, J. K. (2002). *Training in organizations.* Belmont, CA: Wadsworth.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*, 814–834.

Hammond, K. J. (1990). Case-based planning: A framework for planning from experience. *Cognitive Science, 14,* 385–443.

Helton-Fauth, W. B., Gaddis, B., Scott, G., Mumford, M. D., Devenport, L. D., Connelly, S., et al. (2003). A new approach to assessing ethical conduct in scientific work. *Accountability in Research, 10*, 205–228.

Hung, H., & Wong, Y. H. (2007). The relationship between employer endorsement of continuing education and training and work and study performance: A Hong Kong case study. *International Journal of Training & Development, 11,* 295–313.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Newbury Park, CA: Sage.

Jonassen, D. H., & Hernandez-Serrano, J. (2002). Case-based reasoning and instructional design: Using stories to support problem solving. *Educational Technology Research and Development, 50*, 65–77.

Jones, T. M. (1991). Ethical decision making by individuals in organizations: An issue-contingent model. *Academy of Management Review, 16*, 366–395.

Kalichman, M. W. (2007). Responding to challenges in educating for the responsible conduct of research. *Academic Medicine, 82*, 870–875.

Kalichman, M. W., & Plemmons, D. K. (2007). Reported goals for responsible conduct of research courses. *Academic Medicine, 82*, 846–852.

Kirk, R. E. (1995). *Experimental design: Procedures for behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Kiser, K. (1999). 10 thinks we know so far about online training. *Training, 36*, 66–68.

Kligyte, V., Marcy, R. T., Waples, E. P., Sevier, S. T., Godfrey, E. S., Mumford, M. D., et al. (2008). Application of a sensemaking approach to ethics training for physical sciences and engineering. *Science and Engineering Ethics, 14*(2), 251–278.

Kohlberg, L. (1969). Stage and sequence: The cognitive development approach to socialization. In D. A. Goslin (Ed.), *Handbook of socialization theory* (pp. 347–480). Chicago: Rand McNally.

Kohlberg, L. (1976). Moral stages and moralization: The cognitive-developmental approach. In T. Lickona (Ed.), *Moral development and behavior: Theory, research, and social issues* (pp. 31–53). New York: Holt, Rinehart & Winston.

Kolodner, J. L. (1993). *Case based reasoning.* San Mateo, CA: Morgan Kaufmann.

Kolodner, J. L. (1997). Educational implications of analogy. *American Psychologist, 52,* 57–67.

Kraiger, K., & Jung, K. M. (1996). Linking training objectives to evaluation criteria. In M. A. Quiñones & A. Ehrenstein (Eds.), *Training for a rapidly changing workplace: Application of psychological research* (pp. 151–175). Washington, DC: American Psychological Association.

*Major-Kincade, T. L., Tyson, J. E., & Kennedy, K. A. (2001). Training pediatric house staff in evidence-based ethics: An exploratory controlled trial. *Journal of Perinatology, 21*, 161–166.

Martinson, B. C., Anderson, M. S., & de Vries, R. (2005). Scientists behaving badly. *Nature, 435*, 737–738.

*McKellar, K. A. (1999). Ethical decision-making: Does practice make a difference? *Dissertation Abstracts International, 60*(02), 574. (UMI No. 9920900)

Mumford, M. D., Connelly, S., Brown, R. P., Murphy, S. T., Hill, J. H., Antes, A. L., et al. (2008). A sensemaking approach to ethics training for scientists: Preliminary evidence of training effectiveness. *Ethics and Behavior, 18,* 315–339.

Mumford, M. D., Connelly, S., Murphy, S. T., Devenport, L. D., Antes, A. L., Brown, R. P., et al. (2009). Field and experience influences on ethical decision making in the sciences. *Ethics and Behavior, 19,* 263–289.

*Myyry, L., & Helkama, K. (2002). The role of value priorities and professional ethics training in moral sensitivity. *Journal of Moral Education, 31*, 35–50.

O'Fallon, M. J., & Butterfield, K. D. (2005). A review of the empirical ethical decision-making literature: 1996–2003. *Journal of Business Ethics, 59*, 375–413.

Önkal, D., Yates, J. F. Simga-Mugan, C., & Öztin, S. (2003). Professional vs. amateur judgment accuracy: The case of foreign exchange rates. *Organizational Behavior & Human Decision Processes, 91,* 169–186.

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

Patalano, A. L. & Siefert, C. M. (1997). Opportunistic planning: being reminded of pending goals. *Cognitive Psychology, 34,* 1–36.

*Patenaude, J., Niyonsenga, T., & Fafard, D. (2003). Changes in students' moral development during medical school: A cohort study. *Canadian Medical Association Journal, 168*, 840–844.

*Penn, W. Y. (1990). Teaching ethics: A direct approach. *Journal of Moral Education, 19*, 124–139.

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context-bound? *Educational Researcher, 18*, 16–25.

*Powell, S. T., Allison, M. A., & Kalichman, M. W. (2007). Effectiveness of a responsible conduct of research course: A preliminary study. *Science and Engineering Ethics, 13*, 249–264.

Resnick, D. L. (2003). From Baltimore to Bell Labs: Reflections on two decades of debate about scientific misconduct. *Accountability in Research, 10*, 123–135.

Rest, J. (1976). New approaches in the assessment of moral judgment. In T. Lickona (Ed.), *Moral development and behavior: Theory, research, and social issues.* New York: Holt, Rinehart & Winston.

Rest, J. R. (1979). *Development in judging moral issues*. Minneapolis: University of Minnesota Press.

Rest, J. R. (1986). An overview of the psychology of morality. In J. R. Rest (Ed.), *Moral development and behavior: Theory, research, and social issues* (pp. 133–175). New York: Praeger.

Rest, J. R. (1988). *DIT manual: Manual for the defining issues test* (3rd ed.). St. Paul: University of Minnesota Center for the Study of Ethical Development.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null result. *Psychological Bulletin, 85*, 638–641.

Ruegger, D., & King, E. W. (1992). A study of the effect of age and gender upon student business ethics. *Journal of Business Ethics, 11*, 179–186.

*Ryden, M. B., & Duckett, L. (1991). Ethics education for baccalaureate nursing. *Technical report for the Improvement of Post Secondary Education grant.* Washington, DC: U.S. Department of Education.

Scott, G. M., Leritz, L. E., & Mumford M. D. (2004a). The effectiveness of creativity training: A meta-analysis. *Creativity Research Journal, 16*, 361–388.

Scott, G. M., Leritz, L. E., & Mumford M. D. (2004b). Types of creativity: Approaches and their effectiveness. *The Journal of Creative Behavior, 38,* 149–179.

*Self, D. J., Schrader, D. E., Baldwin, D. C., Root, S. K., Wolinsky, F. D., & Shadduck, J. A. (1991). Study of the influence of veterinary medical education on the moral development of veterinary students. *Journal of the American Veterinary Medical Association, 198*, 782–787.

*Self, D. J., Schrader, D. E., Baldwin, D. C., & Wolinsky, F. D. (1993). The moral development of medical students: A pilot study of the possible influence of medical education. *Medical Education, 27*, 26–34.

Slavin, R. E. (1991). Synthesis of research on cooperative learning. *Educational Leadership, 48,* 71–81.

Slavin, R. E. (1996). Research on cooperative learning and achievement: What we know, what we need to know. *Contemporary Educational Psychology, 21*, 43–69.

Smith, G. (2002). Are there domain-specific thinking skills? *Journal of Philosophy of Education, 36*, 207–227.

Sonenshein, S. (2007). The role of construction, intuition, and justification in responding to ethical issues at work: The sensemaking-intuition model. *The Academy of Management Review, 32*, 1022–1040.

Steneck, N. H., & Bulger, R. E. (2007). The history, purpose, and future of instruction in the responsible conduct of research. *Academic Medicine, 82*, 829–834.

Tannenbaum, S. I., & Yukl, G. (1992). Training and development in work organizations. *Annual Review of Psychology, 43*, 399–441.

Treviño, L. K. (1986). Ethical decision making in organizations: A person-situation interactionist model. *The Academy of Management Review, 11*, 601–617.

Treviño, L. K., Weaver, G. R., & Reynolds, S. J. (2006). Behavioral ethics in organizations: A review. *Journal of Management, 32*, 951–990.

Weeks, W. A., Moore, C. W., McKinney, J. A., & Longenecker, J. G. (1999). The effects of gender and career stage on ethical judgment. *Journal of Business Ethics, 20*, 301–313.

Wexley, K. N., & Latham, G. P. (2002). *Developing and training human resources in organizations* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Wilkinson, L., & the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604.