HHS Public Access

Author manuscript

Ethics Behav. Author manuscript; available in PMC 2019 February 08.

Published in final edited form as:

Ethics Behav. 2017 ; 27(5): 351–384. doi:10.1080/10508422.2016.1182025.

Are Ethics Training Programs Improving? A Meta-Analytic Review of Past and Present Ethics Instruction in the Sciences

Logan L. Watts, **Kelsey E. Medeiros**, **Tyler J. Mulhearn**, **Logan M. Steele**, **Shane Connelly**, and **Michael D. Mumford**

Department of Psychology, The University of Oklahoma

Abstract

Given the growing public concern and attention placed on cases of research misconduct, government agencies and research institutions have increased their efforts to develop and improve ethics education programs for scientists. The present study sought to assess the impact of these increased efforts by sampling empirical studies published since the year 2000. Studies published prior to 2000 examined in other meta-analytic work were also included to provide a baseline for assessing gains in ethics training effectiveness over time. In total,this quantitative review consisted of 66 empirical studies, 106 ethics courses, 150 effect sizes, and 10,069 training participants. Overall, the findings indicated that ethics instruction resulted in sizable benefits to participants and has improved considerably within the last decade. A number of specific findings also emerged regarding moderators of instructional effectiveness. Recommendations are discussed for improving the development, delivery, and evaluation of ethics instruction in the sciences.

Keywords

ethics; ethics instruction; scientific ethics; ethics training; meta-analysis

Science was once held to be a self-correcting endeavor. That is, through the systematic application of procedures designed to maintain accountability (e.g., blind peer review, replication), it was thought that scientists rarely engaged in research misconduct, and that any instances of foul play would be internally detected and corrected. However, high-profile cases ofmisconduct emerging in the 1970s and 1980s have increased pressure on external sources, such as government agencies and research institutions, to hold scientists accountable (Kalichman, 2013). The National Institutes of Health (NIH), and more recently, the National Science Foundation (NSF), issued mandates to research institutions that rely on federal funding to train their scientists in the responsible conduct of research (RCR). Specifically, RCR education was proposed as an important mechanism for improving the integrity of the scientific enterprise that involves instructing scientists, usually young scientists, in the knowledge and application of professional guidelines with respect to key stakeholders (Steneck, 2006).

Correspondence should be addressed to the Principal Investigator, Dr. Michael D. Mumford, Department of Psychology, The University of Oklahoma, Norman, Oklahoma, 73019 or mmumford@ou.edu.

As a result of this increased attention, RCR resources and programsmultiplied over the last three decades (Kalichman, 2014).Despite this increase in resources, scholars commenting on the current state of RCR education continue to sound pessimistic. For example, Steneck (2013) and Resnik (2014) recently argued that RCR education does little to improve trainee ethicality. Kornfeld (2012) presenteda similarly skeptical tone on the subject stating, “RCR instruction cannot be expected to establish basic ethical standards in a classroom of young adult graduate students” (p. 879). Further, Kalichman (2014) recently argued that “RCR education is in need of rescue” (p. 69). Given the pessimism evident in some scholars’ views of the current state of RCR education, three important questions come to the fore: 1) how effective is RCR education, 2) what types of learning outcomes are influenced by RCR courses, and 3) is RCR educationimproving? The present study seeks to investigate these questions by meta-analyzing the empirical research literature on ethics instruction in the sciences.

Mumford, Steele, and Watts (2015) recently argued that meta-analysis may serve as an important tool for measuring the effectiveness of RCR programs, identifying best practices for improving ethics education, and systematically investigating changes in program effectiveness over time. Prior meta-analytic work by Mumford and colleagues (Antes et al., 2009, 2010; Waples et al., 2009) demonstrated that ethics courses, on average, result in small to moderate effects. However,these initial efforts were limited by the small number of empirical studies of ethics instruction available at the time, and many ethics evaluation studies have been published since these reviews. In addition, these prior effortsshowed considerable variability among assessed programs with regard to what was evaluated, magnitude of program effectiveness, and the types of instructional characteristics employed, suggesting that designing RCR courses to incorporate features associated with more effective courses may prove of particular value. Kalichman (2013) recently noted that while some RCR programs have proven effective, many fail. Thus, a final goal of the present study is to identify instructional characteristics that may help to differentiate between higher and lower performing ethics education programs, such that best practices might be identified for facilitating improvements in the effectiveness of ethics instruction.

Potential Moderators of Instructional Effectiveness

Antes et al. (2009) identified and examined several moderators of instructional effectiveness to help explain why some ethics instruction programs prove more effective than others. In order to extend this prior work, the present study examines similar categories of moderators. However, the list of potential moderator categories was also expanded through further examination of the literature bearing on training development, delivery, and evaluation (e.g., Arthur, Bennett, Edens, & Bell, 2003; Goldstein & Ford, 2002; Salas, Tannenbaum, Kraiger, & Smith-Jentsch, 2012). That is, in the present study, 13 categories of potential moderators were examined, including: 1) time of publication, 2) general criterion type, 3) criteria characteristics, 4) study design characteristics, 5) participant characteristics, 6) quality ratings, 7) general instructional parameters, 8) trainer characteristics, 9) instructional development, 10) instructional content, 11) delivery methods and activities, 12) case-based instruction, and 13) practice characteristics.

Time of publication.

Due to the substantial resources that have been devoted toRCR education over the last two decades (Kalichman, 2014), time of publication was included as a moderator of instructional effectiveness. These improvements may have occurred vis-à-vis a number of mechanisms, including improvements in the identification, development, and delivery of instructional content, or improvements in the quality of evaluation designs and criteria used to measure instructional effectiveness, among other characteristics. Thus, we expect more recently published studies to demonstrate larger benefits to participants than prior studies of ethics programs.

General criterion type.

Ethics programs employ a variety of criteria for assessing instructional effectiveness, and these differences in criteria may result in differences in effect sizes observed in meta-analytic studies (Antes et al., 2009). Many ethics programsappear to rely on established theoretical frameworks for guiding criterion selection decisions, such as Rest's (1986) four-stage model of moral behavior which consists of recognizing moral issues, making a judgment, establishing moral intent, and ultimately acting. Along these lines, some of the most common criteria employed in RCR courses include assessments of knowledge (e.g., recall of professional guidelines) and ethical awareness (e.g., sensitivity to, or recognition of, ethical dilemmas)which underlie the ethical issue recognition identified in Rest's first stage of moral behavior. Other common criteria include moral judgment (e.g., evaluating the morality of a character's actions)and moral reasoning (e.g., stage of cognitive moral development)which typically rely on established measures for assessing the second stage of Rest's framework—moral judgment (Jones, 1991). However, some ethics programs have used more global criteria that simultaneously assess multiple stages of the moral behavior process, such as measures of ethical decision making and the use of metacognitive strategies (e.g., forecasting, emotional self-regulation) that aid individualsin making ethical decisions. Finally, although less common, some ethics programs use attitudinal criteria (e.g., perceptions of self and others) or measures of abstract thinking (e.g., conceptual development) to assess the benefits of ethics instruction.

Criteria characteristics.

In addition to the general criteria used in ethics evaluation studies, the specific criterion employed may account for observed differences in the effectiveness of instructional programs. For example, the Defining Issues Test (DIT; Rest, 1979) is a general, off-the-shelf assessment of an individual's moral reasoning ability that is not specific to any particular field. Because fields differ with regard to legal and ethical norms, guidelines, and codes, measures that assess the application of general ethical principles, while efficient, may show less sensitivity in detecting ethics training gains compared with measures that are customized for the profession in which trainees operate. For example, a field-specific version of the DIT has been developed for engineers (Borenstein, Drake, Kirkman, & Swann, 2010). In other words, criteria developed in-house, for a specific field and training context, may be expected to provide more comprehensive, and accurate, assessments of

gains due to instruction (Goldstein & Ford, 2002). Thus, identifying specific criteria, and the characteristics of these criteria, that are more or less useful may help instructors, program designers, and researchers to improve the effectiveness of measurement efforts over time (Mumford et al., 2015).

Study design characteristics.

Along these lines, the evaluation design employed to study program effectiveness may also influence gains observed due to instruction (Arthur et al., 2003). For example, Antes et al. (2009) found that studies employing stronger designs, with regard to internal validity (e.g., pre–post with control group), demonstrated slightly smaller training effects than alternative designs that lacked such internal controls (e.g., pre–post). Because the Cohen's *d* effect size varies little with respect to sample size (Cohen, 1992), sample size was not expected to have a significant influence on ethics instruction effectiveness. Similarly, as found by Antes et al. (2009),study funding was not expected to influence the effect size observed. Journal impact factor, or an index of the average number of citations received during a specific timespan (e.g., prior three years) for the journal outlet where a study was published, is sometimes employed as a proxy measure for study quality. However, due to a number of criterion contamination issues associated with the useof journal impact factor as a measure of study quality (e.g., Seglen, 1997), we did not expect much variability in effect sizes based on impact factor. Finally, because studies showing statistically significant results tend to get published more often than studies showing no effects (Rosenthal, 1979), we expected peer-reviewed studies to show larger effects than unpublished studies.

Participant characteristics.

In addition to characteristics of study design and the training environment, characteristics of the participants themselves may be expected to influence the effectiveness of instruction (Goldstein & Ford, 2002). Older participants (e.g., early- and mid-career professionals), for example, may be better prepared to benefit from instruction involving complex problem-solving, in comparison to their younger counterparts (e.g., undergraduate students). This may be due to their greater maturity with regard to emotional self-regulation and social interactions (Blanchard-Fields, 2007) and having accumulated more life and professional experiences (Dane & Sonenshein, 2015). However, participants with prior exposure to instruction in ethics may not show the same gains as those who are exposed to such material for the first time.

The gender of participants may also influence the outcomes observed, as many studies have demonstrated that women, on average, tend to score slightly higher than men on measures of ethical judgment and reasoning (O'Fallon & Butterfield, 2005). Thus, because men tend to start lower on these measures, we might expect men to show larger gains than women due to ethics instruction. Finally, fields may differ with regard to the instructional content and delivery approaches employed, indicating that field differences in ethics instruction effectiveness may be observed (Antes et al., 2009). For example, ethics instruction for social scientists may be expected to focus more on research issues involving the protection of human subjects and data sharing compared with training programs targeting engineers,which

might focus more on safety issues, bids, and contracts. Finally, due to the fact thatmany of the instructional programs in ethics and publicly availableresources have been developed in the United States for English-speaking populations, international training programs, which may employ different content and criteria, and occur in different cultures, could also show differences in gains observed due to ethics instruction (Steele et al., 2015). In sum, differences in participants were examined to determine if these trainee characteristics were associated with the effectiveness of instruction.

Quality ratings.

In addition to examining objective characteristics of ethics instruction, subjective assessments might also provide evidence of external validity for program effectiveness (e.g., Antes et al., 2009; Scott, Leritz, & Mumford, 2004; Waples et al., 2009). Thus, quality ratings of instruction, study design, and criteria were assigned to each course examined in this study by three judges familiar with the ethics instruction and evaluation literatures. More details concerning these ratings will be presented in the method description.

General instructional parameters.

Ethics courses vary considerably with regard to their length and delivery format, among other characteristics. For example, courses employing a brief, online delivery format may differ from those employing a semester-long, face-to-face or hybrid delivery format (Sitzmann, Kraiger, Stewart, & Wisher, 2006). Prior meta-analytic work (Antes et al., 2009) has shown that stand-alone courses exceeding one day of instruction delivered in a face-to-face format may be especially beneficial to ethics training participants. In addition, features of the training environment may also influence instructional effectiveness (Goldstein & Ford, 2002). Thus, the present effort examinestotal hours of instruction, stand-alone versus embedded instructional format, face-to-face versus online versus hybrid delivery format, average class size, and the mandatory or voluntary nature of participation.

Trainer characteristics.

Ethical dilemmas are by nature complex, ambiguous, and ill-defined (Mumford et al., 2008). In other words, ethical dilemmas are often “gray.” In such a domain, trainers, or instructors, may be expected to take on a critical role not only as a conveyor of content (e.g., professional guidelines), but as a facilitator that guides participants in the development and application of complex problem-solving skills. This suggests that considerable expertise may be required of ethics instructors. Along these lines, we expected trainees’ benefits from instruction to increase with trainer expertise. Further, exposure to multiple trainers, each with varied experiences and knowledge, may prove beneficial. Lastly, when trainers are rotated, participants are exposed to a greater number of perspectives, which may be expected to improve ethical reasoning. Alternatively, if too little time is devoted to each instructor, trainer rotation may interfere with participant acquisition of instructor mental models (Kligyte et al., 2008). Thus, we expected differences in trainer characteristics to contribute to differences in observed effect sizes.

Instructional development.

Instructional objectives, or explicit statements about the goals of instructional courses, have been found to vary widely among RCR programs (Kalichman & Plemmons, 2007). In order to identify the characteristics of more effective programs, Antes (2014) recently called for a more systematic approach to RCR instruction, including clear specification of objectives that guide instructional development and delivery. For example, programs that fail to specify instructional objectives, or specify too many objectives, may fail to accomplish much (DuBois & Dueker, 2009). Further, overly specific objectives may negatively influence program effectiveness by limiting the applicability of instruction. Finally, instructional content that is customized for the training context in question may be expected to result in greater benefits to participants, compared with off-the-shelf training programs (Goldstein & Ford, 2002).

Instructional content.

Characteristics of instructional content vary considerably between courses and institutions. For example, in a meta-analysis ofethics instruction programs in engineering, Haws (2001) identified a variety of pedagogical approaches employed, including focusing on philosophical theories of morality, case-based instruction, and courses explicitly focused on professional codes of conduct, among other approaches. Many RCR programs choose to draw on the professional guidelines framework presented by the Office of Research Integrity (ORI). The nine core content areas described by ORI (Steneck, 2007) include: 1) research misconduct (e.g., fabrication, falsification, plagiarism), 2) protection of human subjects, 3) welfare of laboratory animals, 4) conflicts of interest, 5) data management practices, 6) mentor/trainee relationships, 7) collaboration, 8) authorship/publication practices, and 9) peer review. Instructional programs that draw on an established framework such as this, applied within a particular field, may be expected to achieve more comprehensive coverage of the domain of potential ethical dilemmas faced by participants (Antes, 2014). Further, courses that emphasize the application of professional guidelines, vis-à-vis practicing the development of cognitive processes, or strategies of decision making, may benefit participants (Mumford et al., 2008). Finally, training participants in the identification of stakeholders (e.g., government, funding institution, research subjects, public) may also benefit participant decision making, as ethical decisions must be made with respect to key stakeholders (Steneck, 2006). Thus, it is expected that differences in instructional content will contribute to differences in observed effect sizes.

Delivery methods and activities.

In addition to the instructional content presented, how this content is delivered may also influence training effectiveness. For example, employing a variety of delivery activities, versus relying on a single approach, may be expected to improve participant engagement with the content and ultimately facilitate knowledge and skill acquisition (Ames, 1992). Similarly, programs that structure opportunities for active trainee participation may also be particularly effective (Antes et al., 2009). Further, courses that separate instructional content

into separate, but coherent, chunks may also support participant learning when domains of learning are particularly complex (Goldstein & Ford, 2002).

Case-based instruction.

Due to the ambiguity and complexity involved in many ethical dilemmas, case-based, or experiential, knowledge has been identified as a critical resource for engaging in reasoning that leads to more ethical decisions (Kolodner, 1992; Mumford et al., 2008). Along these lines, prior meta-analyses by Antes et al. (2009) and Waples et al. (2009) demonstrated the benefits of case-based instruction on ethics training outcomes. Thus, courses that emphasize case-based instruction may be expected to result in larger benefits to participants. However, cases also vary with regard to a number of characteristics that may influence the magnitude of these benefits, includinglength, complexity, emotions, and realism. On the one hand, longer cases that are rich in content may be expected to engage participants in deeper processing,resulting in improved knowledge and skill development (Johnson et al., 2012). On the other hand, cases that are too rich with regard to their complexity, affective content, and realism may overwhelm participant cognitive processes, such that learning is disrupted (Bagdasarov et al., 2013; Thiel et al., 2013).

Practice characteristics.

Prior research points to the importance of application and practice opportunities in facilitating knowledge acquisition and the transfer of complex knowledge (Ericsson, Krampe, & Tesch-Romer, 1993). Thus, ethics courses offering frequent opportunities to practice and apply content may be particularly beneficial to participants. However, in addition to frequency of practice opportunities, several other characteristics have the potential to influence the benefits of practice that have not received much empirical attention in an ethics instruction context, including the duration, field-specificity, and realism of practice content (Spiro, Coulson, Feltovich, & Anderson, 1988). Further, practice groups vary in size, and group-based practice may result in different effects than individual-based practice. Given these observations, the following research questions were used to guide the present meta-analytic effort:

Research Question 1: Have ethics instruction programs in the sciences, on average, become more effective over time?

Research Question 2: What moderators, or instructional characteristics, are associated with more effective ethics programs?

Method

Literature Search

An extensive literature search was conducted in order to identify potential studies for inclusion in the meta-analysis. First, key review articles (Craft, 2013; O'Fallon & Butterfield, 2005) and prior meta-analytic studies (e.g., Antes et al., 2009, 2010; Waples et al., 2009) pertaining to the ethics instruction in research, science, and academic and

professional settings were collected. The researchers searched the reference lists of these articles to identify empirical studies of ethics instruction.

Second, 26major databases were selected from a list of over 100 databases available to the researchers through their institution's library resources. Only databases judged relevant to the responsible conduct of research (RCR), ethics in the physical sciences, social sciences, health, engineering, and professional ethics were searched, including Academic Search Elite, American Bibliography of Slavic and East European Studies, Anthropology Plus, Article 1st, ASCE Library, ASTM Standards and Engineering Digital Library, Chronicle of Higher Education, CINAHL Plus, Communication & Mass Media Complete, Communication Abstracts, CompendexWeb, Current Contents Connect, ERIC, Google Scholar, Health Source: Nursing/Academic Edition, IEEE Xplore, INSPEC, JSTOR, MasterFILE Premier, MEDLINE, Military & Government Collection, ProQuest Natural Sciences Collection, PsycARTICLES, PsycINFO, PubMed, and SocINDEX. These databases were searched using targeted search terms, including: "ethics training", "ethics education", "responsible conduct of research", "moral development training", "ethics instruction", and "professional ethics training."

Third, following the search of major databases, 14 journals were targeted with relevance to ethics instruction, including *Academic Medicine*, *Accountability in Research*, *Ethics & Behavior*, *Journal of Empirical Research in Human Ethics*, *Journal of Further and Higher Education*, *Journal of Medical Ethics*, *Journal of Moral Education*, *Journal of Nursing Education*, *Medical Education*, *Nursing Ethics*, *Science and Engineering Ethics*, *Studies in Higher Education*, *Teaching Higher Education*, and *Teaching of Psychology*. The Principal Investigator and Co-Principal Investigator, senior researchers in research ethics, evaluated the lists of databases and journals for completeness.

Fourth, several methods were applied to help reduce the potential for file drawer bias, or the tendency for peer-reviewed journals to disproportionately publish statistically significant resultsthat might upwardly bias effect size estimates observed in meta-analyses (Hunter & Schmidt, 2004; Rosenthal, 1979). Unpublished dissertations relevant to ethics training were searched via the *ProQuest Dissertation Abstracts* database. In addition, conference presentations relevant to ethics instruction were identified by searching the program lists associated with various professional societies, such as the American Psychological Association, International Association for Education in Ethics, Society for Industrial and Organizational Psychology, and the World Conference in Research Ethics. Moreover, we contacted principal investigators of grants related to ethics instruction funded by the National Institutes of Health (NIH) and National Science Foundation (NSF) to request unpublished manuscripts and evaluation reports. Further, we contacted RCR Program Directors and Research Vice Presidents associated with doctorate-granting, Carnegie research institutions requesting unpublished manuscripts or evaluation reports pertaining to ethics training programs. Finally, corresponding authors of articles that met the inclusion criteria, described next, were contacted with the goal of obtaining supplementary materials describing their ethics course (e.g., syllabi) and unpublished evaluation reports. These initial search procedures for ethics training evaluation studies and unpublished manuscripts resulted in the identification of 4,671 studies for potential inclusion in the meta-analysis.

Inclusion Criteria

To determine which studies were appropriate to include in the meta-analysis, multiple inclusion criteria were applied. First, studies must have reported an empirical investigation of the effectiveness of ethics instruction or training. A large portion of the studies identified in the initial search did not meet this first inclusion criterion and, thus, were excluded from further review. Many of the excluded papers were theoretical in nature, presented only qualitative results, or presented the results of field-based opinion surveys reporting ethics training needs and preferences. Approximately 190 remaining studies were examined using the following inclusion criteria.

The second criterion for inclusion in the meta-analysis required that studies provide a description of the course characteristics evaluated in each study as well as a clear description of the ethics-related outcome measures used to assess instructional effectiveness. Third, the study must have presented the necessary descriptive (e.g., means, standard deviations) and/or inferential (e.g., *F*-value, *t*-value, χ^2-value) statistics required to calculate an effect size (i.e., Cohen's *d*).

Before calculating effect sizes, the independence of *d* statisticsacross studies and criteria was considered. When manuscripts described more than one course involving ethics instruction, each course was treated as independent. Additionally, when courses presented evaluation data for multiple criteria that could be expected to theoretically map onto the same higher-order construct—such as when effect sizes for multiple, decision-making strategies could be calculated—these criteria were averaged to reduce data dependency issues. Alternatively, when multiple effect sizes could be calculated for a course that represented different inferred constructs (e.g., knowledge and moral reasoning), these effects were treated as independent.

Once these inclusion criteria were applied and *d* statistics were calculated, 66 empirical studies consisting of 106 ethics courses, 150 effect sizes (*k*), and 10,069 training participants (*N*) remained in the final dataset. Approximately 33% of effect sizes were drawn from unpublished sources, including: dissertations ($k = 17$), theses ($k = 2$), conference papers/presentations ($k = 15$), technical reports ($k = 13$), and other unpublished manuscripts ($k = 2$). Data points from all 150 effect sizes were included in the overall meta-analysis. However, given that courses employed varying content and methods, not all studies were applicable in each moderator analysis. Further, course descriptions differed in comprehensiveness, and studies with lacking information bearing on specific moderators were treated as missing data and excluded from applicable moderator analyses. Thus, the number of effect sizes and participants included in each moderator analysis may not include the full dataset.

Coding Procedures

To collect information regarding the influence of instructional characteristics on training effectiveness, the final database of 66 empirical studies were content analyzed. These content ratings were conducted by three trained judges, all doctoral students in Industrial and Organizational Psychology that were familiar with the literatures bearing on ethics instruction, general instructional design, and organizational training. Before conducting content ratings, however, the authors reviewed the training and ethics instruction literatures

and developed a comprehensive list identifying approximately 300 variables of potential interest as moderators of ethics instruction effectiveness. Additionally, operational definitions and rating scales were developed for each variable. The Principal Investigator and Co-Principal Investigator reviewed the initial variable list, operational definitions, and ratings scales, and provided recommendations for grouping variables into general moderator categories which will be described later. Further, six experts in ethics training, external to the research team, reviewed and critiqued the final list of variables and rating scales. These external reviewers were asked to identify any variables missed by the project team that might influence the effectiveness of ethics instruction.Few additional variables were identified, providing some support for the comprehensiveness of this variable list.

Once coding materials were refined to incorporate the feedback of our expert reviewers,the same three doctoral students each received approximately 40 hours of rater training. Thepurpose of this training was to develop a shared frame of reference toensure consistency in rating scale application (Bernardin & Buckley, 1981). After reviewing the operational definitions and scales, the first five studies were coded as a group by the three raters to encourage discussion of variables requiring more clarity or judgment. Next, an additional five studies were coded independently by the three raters. After raters met again to discuss discrepancies and achieve consensus in ratings, the three raters proceeded with coding the remainder of the studies independently. Interrater agreement coefficients for variable categories ranged from .68 to .96, with a mean of.86. Finally, although approximately 300 variables were included in the initial meta-analytic database, inadequate descriptions of study details (i.e., missing data) required many of these variables to be dropped from further analysis.

Criteria

A variety of criteria were employed across studies included in the present effort. Specifically, nine general categories of criteria were identified, including: moral reasoning (e.g.,Defining Issues Test; Rest, 1979; Rest et al., 1999), knowledge (e.g., Heitman et al., 2007), ethical awareness (e.g., ethical sensitivity tests; Clarkeburn, 2002), ethical decision making and metacognitive strategies (e.g., Mumford et al., 2006), moral judgment (e.g., moral judgment tests; Lind, 1998), conceptual development (e.g., paragraph completion method; Hunt, 1977), and perceptions of selfand others (e.g., self-assessed gains in knowledge, ethical attitudes towards others; Wilson et al., 1993). Due to the relatively small number of effect sizes (*k*) available for each criterion type, these nine criterion types were aggregated into a single, overall effect size of instructional effectiveness when examining the influence of the moderators described next.

Moderators

Year of publication.

The purpose of coding for publication year was to provide an overall assessment of changes in the effectiveness of ethics instruction in the sciences over time. The year 2007 was chosen as a cut point because the most recent meta-analysis of ethics instruction in the sciences (i.e., Antes et al., 2009) included studies through the year 2006. This provided an opportunity for

replicating prior meta-analytic work as well as providing a comparison group of more recent studies. Thus, each effect size was coded as published before 2007, or between 2007 and 2015.

Criteria characteristics.

Characteristics of criteria, or measures used to assess instructional effectiveness, were coded to assess the extent to which variability in instructional effectiveness could explain differences in criteria employed. Along these lines, when information was provided regarding the specific criterion, or measure, employed (e.g., DIT, EDM), this information was recorded, along with whether or not an index of reliability was reported for the criterion. Additionally, how each criterion was developed (i.e., off-the-shelf, in-house) and the extent to which criterion content focused on a specific field (i.e., field-specific, combination, field-general) were also recorded.

Study design characteristics.

In order to assess the influence of study design features(e.g., internal validity) on the overall effect size observed, several features were coded. First, design type (i.e., pre–post with control, pre–post, post with control, longitudinal) was coded. When longitudinal studies provided statistics for a post-test as well as a follow-up test, these data were coded (i.e., follow-up test score minus post-test score) as a measure of skill decay. In addition, several other features of the study design were coded, including sample size (i.e., less than 50, 50 −100, 100+), whether or not the study was funded, and the type (i.e., unpublished, peer-reviewed) and impact factor of the publication (i.e., below average, average, above average).

Participant characteristics.

Characteristics of participants were coded to assess the influence of these characteristics on ethics instruction effectiveness. For example, career stage of participants (i.e., undergraduate students, graduate students, professionals/residents, mixed) was coded, along with whether or not the participants reported any prior exposure to ethics instruction. Gender (i.e., more than 70% males, mixed, more than 70% females) and age (i.e., 18 – 25, 26 – 30, 30+) of participants were also coded. Further, researchers coded for participants' field (i.e., health/medicine, engineering, social sciences, mixed) and whether the ethics instruction occurred in the United States or internationally.

Quality ratings.

Coders were also trained to provide judgments of quality—instructional quality, study design quality, and criterion quality—to assess the influence of these features on instructional effectiveness. Instructional quality was assessed by considering breadth and depth of topic coverage, the use of multiple delivery methods and activities, including activities requiring active trainee participation (e.g., practice, application exercises). Markers of study design quality included assessment at multiple time points (e.g., pre and post, follow-up test), inclusion of a control group, random assignment of participants, and employing multiple, appropriate criteria. Finally, criterion quality was assessed vis-à-vis the presentation of information regarding reliability and validity, the use of multiple items, and

the alignment of the criterion with instructional objectives. A 5-point Likert scale was applied by the coders for each of these three quality variables, where a five represented high quality and a one represented low quality.

General instructional parameters.

A number of general instructional parameters were coded to assess the extent to which characteristics of the instructional setting influenced ethics program effectiveness. For example, when studies presented information regarding the total hours of ethics instruction, this information was recorded and coded into one of three categories (i.e., less than 8 hours, 8 – 16 hours, 16+ hours). Type of instructional format (i.e., integrated/embedded, stand-alone) was also coded, along with the delivery format of the instruction (i.e., face-to-face, hybrid, online). Finally, researchers coded for average class size (i.e., up to 20, 21 – 40, 40+) and whether or not participation in the instructional course was mandatory.

Trainer characteristics.

Because differences in trainers, or instructors, might be expected to influence the effectiveness of ethics instruction, trainer characteristics were coded. Specifically, we coded for the number of trainers (i.e., 1, 2, 2+), level of trainer expertise (i.e., below average, average, above average), and whether or not the trainers were rotated throughout the course. When information regarding trainer expertise was presented, a 5-point Likert scale was applied by the coders, where a five represents extensive expertise and a one represents little to no expertise.

Instructional development.

The purpose of coding for instructional development characteristics was to assess the influence of training design decisions on instructional effectiveness. Along these lines, we coded for the source of instructional content (i.e., mostly off-the-shelf, mixed, mostly in-house). Additionally, when studies presented information bearing on the instructional objectives that were employed to guide program implementation, these instructional objectives were coded for number (i.e., 1 – 3, 3+) and specificity (i.e., below average, average, above average). A 5-point Likert scale was used to code for specificity of instructional objectives, where a five represents high specificity and a one represents low specificity.

Instructional content.

Features of the instructional content, or training material, may also be expected to influence instructional effectiveness. However, not all studies presented detailed information regarding instructional content. To assess if studies presenting no content details differed systematically from those that presented such details, we coded for whether or not instructional content information was presented. When content details were presented, a number of characteristics were coded. For example, researchers coded for the breadth of coverage of the 9 Office of Research Integrity (ORI) guidelines (i.e.,1 – 3, 3+) as well as the breadth of general topic areas covered (i.e., 1 – 6, 7 – 12, 12+). For example, if a course was described as covering authorship and publication guidelines, the Common Rule,

whistleblowing, data management, scientific misconduct, professionalism, and decision-making strategies, this was coded as covering 7 topics, or moderate breadth of content coverage. The extent to which content focused on a particular field was also coded (i.e., field-specific, balanced, field-general). Five-point Likert scales were used to assess the extent to which the ethics course emphasized decision making processes(5 = *great emphasis*, 1 = *little to no emphasis*) and the number of stakeholders mentioned (5 = *many stakeholders*, 1 = *few to no stakeholders*). Finally, we coded for specific course detailsconcerning the presence of approximately 70 content areas, or topics, including process-based content.

Delivery methods and activities.

Characteristics of the delivery methods and activities employed might also be expected to influence the effectiveness of ethics instruction. Once again, to assess the extent to which studies with missing data differed systematically from those that presented delivery details, we first coded for whether or not delivery method and activity information was presented. For those studies that presented delivery details, we coded for breadth of learning activities (i.e., 1–3, 4–6, 6+). Additionally, we coded for several features of the learning approachemployed, such as a whole versus part-learning, guided versus self-directed learning, and individual versus group-based learning. A five-point Likert scale was used to code the extent to which activities promoting active trainee participation were emphasized, where a five represented high active participation and a one represented low active participation. Finally, we coded for the use of approximately 25 specific delivery methods and activities.

Case-based instruction.

Given that cases are commonly employed to deliver instructional content in ethics courses, several features of these cases were coded to assess the impact of case characteristics on instructional effectiveness. First, a 5-point Likert scale was used to code for emphasis on case-based instruction, with five representing great emphasis and one representing little to no emphasis. In addition, we coded for several characteristics that varied depending upon the specific cases used, such as length (i.e., 1 paragraph, 2–6 paragraphs, 6+ paragraphs), complexity (i.e., low, moderate, high), emotional content (i.e., low, moderate, high), and realism (i.e., low, moderate, high).

Practice characteristics.

Finally, characteristics of practice opportunities were coded to assess the influence of these characteristics on instructional effectiveness. A 5-point Likert scale was used to assess frequency of practice opportunities, where a five represented many opportunities and a one represented little to no opportunities. The extent to which the content of practice opportunities was specific to a particular field was also coded (i.e., field-specific, balanced, field-general). Further, we coded for length of practice opportunities (i.e., up to 30 minutes, 30+ minutes), realism (i.e., low, moderate, high), and whether practice occurred in an individual, mixed, or group-based format. We also coded for size of practice groups (i.e., 2–5, 5+).

Analysis Overview

All analyses, including the calculation of effect sizes, were conducted using Comprehensive Meta-Analysis (CMA) software version 3. Based on the procedures developed by Hedges and Olkins (1985), the random-effects model was used to estimate inverse-variance weighted effect sizes. Beyond this weighting procedure which helps to control for sampling error (Hedges & Vevea, 1998), no further statistical corrections were applied. Estimates of sample heterogeneity were calculated using the I^2 statistic, the inverse of which (i.e., 100 - I^2) provides an estimate of the variance accounted for by sampling error or other artifacts (Kepes, McDaniel, Brannick, & Banks, 2013). Thus, when the variance due to sampling error is small (< 75%), this was interpreted to indicate that moderators are likely present (Hunter & Schmidt, 2004). Moderator analyses were included if at least two effect sizes were available (Arthur, Bennett, & Huffcutt, 2001).

In order to assess the potential influence of outliers on the stability of the overall effect size, a one-sample-removed analysis was conducted (Borenstein, Hedges, Higgins, & Rothstein, 2009). The one-sample-removed analysis provides an estimate of the overall effect size given that each sample is the only sample removed, one sample at a time. However, no individual sample demonstrated a significant impact on the overall effect size, as demonstrated by the fact that the overall Cohen's *d* shifted by no greater than .01 due to the removal of any sample. Thus, outliers were not considered a serious issue.

The magnitude of Cohen's *d* effect sizes was interpreted with respect to conventional standards as described by Cohen (1992). Specifically, .20 was interpreted as a small effect, .50 was interpreted as a medium-sized effect, and .80 or greater was interpreted as a large effect. In addition, 95% confidence intervals were estimated around the inverse variance-weighted mean effects. These confidence intervals were interpreted with respect to guidelines provided by Hunter and Schmidt (2004). Thus, if the confidence interval included zero, sampling error was not ruled out as an explanation of the size of the effect. In addition, confidence intervals were used to compare the effectiveness of intervention characteristics across several moderators. When confidence intervals for competing moderator categories overlapped, sampling error, again, was not ruled out as a potential explanation for the observed differences in mean effect sizes. On the other hand, when confidence intervals failed to overlap, this was interpreted as a statistically significant result.

Results

Overall Effectiveness and Time of Publication

Table 1 presents the overall meta-analysis results and a breakdown of effect sizes by time and general criterion type. Overall, ethics training programs demonstrated medium-sized effects on ethics outcomes (d = .48, SD = .04). Because the estimated variance due to sampling error was small (11%), the presence of moderators was investigated.The three general criterion types demonstrating the largest effect sizes were knowledge (d = .78, SD = .14), perceptions of self (d = .66, SD = .19), and ethical decision making (d = .51, SD = .12). On the other hand, the poorest performing, general criterion types included perceptions

of others (d = −.01, SD = .11), conceptual development (d = .24, SD = .18), and moral judgment (d = .25, SD = .06).

Effects of Moderating Variables

Year of publication.—Training programs described in publications occurring before 2007 demonstrated smaller effects (d = .36, SD = .07) than training programs in publications occurring between 2007 and 2015 (d = .56, SD = .06).

Criteria characteristics.

As can be observed in Table 2, the Defining Issues Test (DIT), in its various versions, was the most popular criterion used to evaluate ethics training effectiveness. The strongest effects were observed for the field-specific DIT (d = 1.14, SD = .35), whereas the weakest effects were observed for the DIT-2 (d = .16, SD = .15) and the moral judgment test (MJT; d = .16, SD = .06). In-house measures demonstrated significantly larger effects (d = .77, SD = .11) than off-the-shelf measures (d = .32, SD = .05). Further, field-specific measures, or those that focused on ethics applying in a particular field (e.g., engineering), showed significantly larger effects (d = .70, SD = .11) than field-general measures, or those that attempted to target multiple fields (d = .29, SD = .06).

Study design characteristics.

Table 3 presents moderator analysis results for a number of study design characteristics. No significant differences were observed based on the type of study design employed. Overall decay in knowledge or skill gained from instruction was also non-significant (d = −.02, SD = .06), indicating that the effects of ethics training appear to hold over time. No significant differences were observed between studies based on sample size, funding status, or the impact factor of the journal of publication. However, publication type proved to be a statistically significant moderator, with peer-reviewed studies (d = .59, SD = .06) demonstrating larger effect sizes than unpublished studies (d = .25, SD = .05).

Participant characteristics.

Table 4 shows that few statistically significant results were observed based on participant characteristics. For example, no significant differences were observed between participants based on their career stage (e.g., undergraduates, graduates, professionals). However, greater benefits were observed for participants with no prior exposure to ethics instruction (d = 1.13, SD = .46), when compared to those with prior ethics training experience (d = .46, SD = .08). Although no statistically significant differences were observed based on gender or age, the data exhibited clear patterns showing that males (d = .66, SD = .15) appeared to benefit more from ethics training than females (d = .36, SD = .11). Additionally, middle-aged participants (d = .63, SD = .15) appeared to benefit more than their younger counterparts (d = .38, SD = .05).Regarding participant field, the largest effects were observed for engineering participants (d = .66, SD = .12), whereas the weakest effects were observed for instructional settings that included participants from multiple fields (d = .20, SD = .06). Finally, location of instruction showed statistically significant differences, with international training programs (d = .79, SD

= .12) showing larger effect sizes than domestic (i.e., United States) training programs (d = .42, SD = .05).

Quality ratings.

As presented in Table 5, no statistically significant differences were observed in ethics programs based on ratings of instructional quality or the quality of study design. However, a significant difference was observed in ratings of criterion quality. Specifically, below average quality criteria (d = .57, SD = .07) demonstrated larger effects than average quality criteria (d = .28, SD = .07).

General instructional parameters.

Table 6 presents the results of a number of moderators related to general instructional characteristics. Although few statistically significant differences may be observed based on these moderators, a number of potentially noteworthy patterns emerged. For example, effect sizes slightly decreased as the total hours of training time increased. Stand-alone programs (d = .51, SD = .05) demonstrated similar benefits to integrated, or embedded, instructional programs (d = .44, SD = .08). No significant differences were observed based on delivery format or class size. However, voluntary ethics training programs (d = .52, SD = .09) demonstrated significantly larger effects than mandatory programs (d = .26, SD = .03).

Trainer characteristics.

As demonstrated in Table 7,clear trends were observed regarding the influence of trainer characteristics. First, program effectiveness appeared to increase as the number of trainers increased. Programs employing more than two trainers (d = 1.07, SD = .24), for example, showed significantly larger effect sizes than programs employing a single trainer (d = .36, SD = .09). Trainer expertise showed a similar trend, with trainers of above average expertise (d = .87, SD = .32) showing larger effects than trainers of average (d = .66, SD = .14) or below average expertise (d = .35, SD = .06). Trainer rotation appeared to have no significant impact on training effectiveness.

Instructional development.

Table 8 presents the results examining instructional development characteristics, such as the source of course development and instructional objectives. Although programs developed in-house were the most frequent and showed the largest effect sizes (d = .54, SD = .07), the benefits observed were not significantly greater than those programs that employed a mixed development (d = .40, SD = .05) or off-the-shelf (d = .40, SD = .21) instructional approach. Courses employing a small number of instructional objectives showed slightly larger effect sizes (d = .66, SD = .08) than courses employing no objectives (d = .43, SD = .07) or a large number of objectives (d = .44, SD = .08). Finally, a trend was observed with regard to the specificity of instructional objectives. Highly specific instructional objectives appeared to inhibit effective instruction (d = .46, SD = .08), compared with programs employing average specificity (d = .64, SD = .13) or broadly defined (d = .62, SD = .24) objectives.

Instructional content.

Characteristics of instructional content were examined as moderators and presented in Table 9, 10, and 11. No significant differences were observed between programs that provided content details versus programs that failed to present such details, providing some justification for the inclusion of such studies in the meta-analysis. Training programs that focused on more comprehensive coverage of the nine ORI research guidelines (d = .56, SD = .10) trended towards slightly larger effect sizes than those that focused on only a small number of these guidelines (d = .41, SD = .06). A similar trend was identified regarding breadth of topic areas. That is, programs that delivered a large number (i.e., 12+) of topics (d = .57, SD = .09) trended towards slightly larger effect sizes than programs focusing on fewer topics (d = .46, SD = .08). Both field-specific programs (d = .61, SD = .09) and field-general programs (d = .82, SD = .11) significantly outperformed programs that attempted to take a balanced instructional approach (d = .22, SD = .05). No statistically significant differences were found regarding an emphasis on decision processes or number of stakeholders.

Regarding specific instructional content, the largest effects were observed for programs that included the following topic areas: sexual harassment (d = 1.60, SD = .37), the Nuremberg code (d = 1.58, SD = .52), personal integrity (d = .96, SD = .23), financial compliance (d = .88, SD = .35), group biases (d = .84, SD = .16), data integrity (d = .82, SD = .20), and field differences (d = .80, SD = .19). On the other hand, the lowest effect sizes were observed among programs including the following topics: appropriate statistical analysis (d = .17, SD = .17), power differentials (d = .18, SD = .04), diversity (d = .19, SD = .09), organizational values (d = .19, SD = .17), peer review (d = .19, SD = .05), and lab safety (d = .19, SD = .15). All of the process-based content areas showed some benefits to trainees. However, the three processes that showed the largest benefits included emotional analysis (d = .76, SD = .17), forecasting (d = .71, SD = .12), and analysis of consequences (d = .68, SD = .11).

Delivery methods and activities.

Tables 12shows that studies failing to provide information regarding delivery methods and activities showed no differences in mean effect size compared with studies providing delivery details. Also, no statistically significant differences were observed based on moderators of delivery methods and activities. However, a number of potentially noteworthy trends emerged. For example, programs employing a small number of delivery activities (d = .58, SD = .10) showed slightly larger effect sizes than programs employing a moderate (d = .42, SD = .06) or large number (d = .46, SD = .10) of activities. Part-learning methods (d = .60, SD = .09) trended towards larger effects than whole-learning (d = .46, SD = .07) or mixed (d = .43, SD = .10) methods. Self-directed methods (d = .63, SD = .10) also showed some benefit compared with guided instructional approaches (d = .40, SD = .09). Further, individual-based programs (d = .53, SD = .09), or programs employing a mixed individual and group approach (d = .52, SD = .07), trended towards larger effect sizes than programs employing a purely group-based, or team-based, approach (d = .27, SD = .07). Finally, programs emphasizing active trainee participation (d = .52, SD = .12) appeared to demonstrate slightly greater benefits to participants compared with programs where a more passive learning approach was employed (d = .41, SD = .08).

Regarding specific delivery methods and activities, Table 13 demonstrates that the largest effect sizes were observed when courses incorporated the following elements: humor (d = .83, SD = .17), note-taking (d = .85, SD = .14), workbooks (d = .68, SD = .15), debates (d = .63, SD = .28), and current events (d = .60, SD = .42). The smallest effects were observed for courses that used games (d = .18, SD = .04), mentoring (d = .19, SD = .17), and service learning (d = .25, SD = .14).

Case-based instruction.

Table 14 presents results bearing on the impact of case-based instruction and case characteristics on training program effectiveness. Although few statistically significant moderators were observed, several trends were noted. For example, programs with an above average emphasis on case instruction appeared especially beneficial to participants (d = .59, SD = .17). Case length also showed a clear trend with longer cases (d = .73, SD = .26) demonstrating larger effects than cases of moderate (d = .55, SD = .14) or short length (d = .43, SD = .11). Cases of moderate complexity (d = .66, SD = .26) showed larger effects than cases of low (d = .52, SD = .14) or high (d = .33, SD = .08) complexity. Clear trends were also observed with regard to emotional content and realism. That is, in both cases, high emotional content (d = .20, SD = .04), orhigh case realism (d = .29, SD = .07), appeared to inhibit instructional effectiveness, when compared with cases of low emotional content (d = 1.03, SD = .38) and low realism (d = .46, SD = .22).

Practice characteristics.

Table 15displays results for the final category of moderators examined in the present study—characteristics of practice opportunities. Programs offering more practice opportunities (d = .76, SD = .16) showed significantly larger effects than programs offering few practice opportunities (d = .32, SD = .06). Additionally, several other characteristics of these practice opportunities showed noteworthy trends. For example, field-general practice (d = .63, SD = .14) showed slightly larger effects than field-specific practice (d = .40, SD = .05). Once again, realism appeared to inhibit training effectiveness, such that practice exhibiting low realism (d = .67, SD = .14) showed larger effect sizes than moderate (d = .32, SD = .07) or high realism (d = .33, SD = .07) practice. Programs with individual-based (d = .56, SD = .15) practice opportunities also showed slightly larger effect sizes than those offering only group-based (d = .28, SD = .07) practice activities. Smaller practice groups (d = .35, SD = .10) showed slightly larger effect sizes than larger groups (d = .27, SD = .07).

Discussion

Before turning to the conclusions emerging from this meta-analytic effort, several limitations should be noted. First, the issue of missing data must be addressed.Few studies provided complete information that could be coded and included in all the moderator analyses examined here. As a result, several moderator variables that were initially coded were excluded from the final analyses simply because they appeared too scarcely to generate stable estimates. On a related note, we urge additional caution when interpreting results based on only a few (e.g., < 10) effect sizes (Hunter & Schmidt, 2004). Although missing data is nearly always a concern when conducting meta-analytic procedures, Rosenthal

(2001) noted that collecting a large number of studies that provide adequate information, and collecting study data from a large number of sources,representsone of the best remedies for overcoming this limitation—a key strategy employed here. Sensitivity analyses were also conducted to determine the extent to which studies offering fewer details concerning their curriculum and delivery characteristics might differ systematically from studies offering more course details. However, no statistically significant bias was found, as indicated by the overlapping confidence intervals presented in the first two rows of Tables 9 and 12.

Additionally, file drawer biasis a potential concern that must be addressed. File drawer bias refers to the issue that only significant results tend to be published in peer-reviewed journal outlets, whereas non-significant results tend to remain in “file drawers” (Rosenthal, 1979). One concern is that observed effects may be biased upwards because studies showing null effects are inaccessible to meta-analysts. Once again, a comprehensive literature search that includes targeting nontraditional sources of data (e.g. theses, dissertations, technical reports, conference presentations, etc.) represents an important strategy for mitigating file drawer bias. In the present study, approximately one-third (49 of 150) of effect sizes were drawn from unpublished studies. Further, the present effort represents a three-fold increase in total articles and a five-fold increase in total effect sizes over the most comprehensive meta-analysis of ethics training programs in the sciences to date (i.e., Antes et al., 2009). Although the influence of file drawer bias in the present study was estimated to be small, it is possible that some relevant studies were overlooked.Considering the significant differences observed between courses frompeer-reviewed (d = .59) versus unpublished studies (d = .25), including unpublished sources in this meta-analysis appears to have provided a more conservative, and perhaps, less biased estimate of average RCR instructional effectiveness than might have been found if only published studies were included.

Further, the method used to combine effect sizes across studies presents a limitation. Specifically, because of the relatively few effect sizes available for each criterion type (e.g., moral judgment, knowledge), moderators of instructional effectiveness were examined based on the overall effect size, or an aggregate of all available effect sizes available for the moderator of interest. As a result, we are limited in our ability to assess differences among criteria with respect to the moderators examined.

Finally, only ethics evaluation programs in the sciences were examinedin the present study. Along these lines, an important question comes to the fore: might ethics programs in other fields show different results? The present study showed some differences in effect sizes among the domains examined here (e.g., medicine/health, engineering, social sciences), but investigating field differences regardinginstructional content, methods, and evaluation characteristics was beyond the scope of the present study. For example, in a meta-analysis of ethics programs in the business professions, Medeiros et al. (2016) found that programs employing group-based activities were more effective than programs that emphasized individual-based activities. A reverse trend was found in the present study, such that courses emphasizing individual-based activities were more effective. In addition, whereas purely online programs constituted 10% of the courses examined here, Weber’s (2015) survey of corporate ethics programs found that 81% of organizations relied on computer-based

methods for delivering ethics training content. In other words, key differences appear to exist between fields regarding not only the guidelines and standards of concern to the profession, but also how instructional content is delivered. Thus, future research islikely needed to identify best practices of ethics instruction programs on a field-specific basis.

With these limitations in mind, a number of noteworthy conclusions may be observed here. First, ethics training programs in the sciences appear to be improving. For example, before correcting for unreliability in the criterion, Antes et al. (2009) reported an average Cohen's *d* of .37, whereas in the present effort, an uncorrected Cohen's *d* of .48 was observed. Compared with studies published prior to 2007 ($d = .36$), studies published since 2007 ($d = .56$) have exhibited effect sizes that are, on average, .20 larger—a positive shift of practical significance. It is noteworthy that the uncorrected *d* estimate drawn from studies before 2007 demonstrates a near-exact replication of the uncorrected *d* observed by Antes et al. (2009). Further, nearly twice as many effect sizes were identified for studies published since 2007, compared with studies published before 2007. In sum, the number of evaluation efforts with regard to ethics education in the sciences appears to have increased substantially within the last decade, and thisincreased attentionappears to have paid off, such that ethics instruction in the sciences has improved in overall effectiveness.

Another key finding emerging from the present effort is that the benefits of ethics instruction appear to hold over time. The retention interval—or time between the first post-test and a second, follow-up post-test—of the studies included in this meta-analysis ranged between one month and two years, with an average time period between assessments of 6 months. It is not uncommon for traditional training interventions to show substantial declines in trainee knowledge and skill following the intervention. Indeed, in a meta-analysis examining the magnitude of skill decay over time, Arthur, Bennett, Stanush, and McNelly (1998) found that at around 6 months following training, the average size of skill decay observed was large ($d = -1.04$). The results of the present meta-analysis, in comparison, show virtually no decay ($d = -.02$). These conclusions must be tempered, however, by the fact that few ethics evaluation studies have collected follow-up measurements. Indeed, this estimate is based on a relatively small number of effect sizes ($k = 8$) by meta-analytic standards. The percentage of variance accounted for by sampling error (66%) also indicated some moderators may be operating. Thus, future efforts investigating how the characteristics of ethics instructionmight support maintenance of knowledge and skills over time may be of some value.

Although the ethics training programs analyzed here, on average, demonstrated sizable benefits to participants, it is important to consider what specific instructional features were associated with the largest benefits. For example, employing multiple trainers that bring sufficient expertise into the instructional environment appears critical. Due to the nature of ethics instruction—that is, training individuals to respond effectively to complex, ill-defined, ambiguous situations—considerable knowledge and skill may be required to effectively facilitate these complex processes (Mumford et al., 2008). Information is lacking regarding who is typically selected to lead the delivery of ethics training programs in the sciences. However, in a survey of corporate ethics training programs, Weber and Wasieleski (2013) found that 91% of the 61 organizations surveyed relied on organizational members as ethics

trainers in comparison with 9% that relied on external ethics training experts. If these results are indicative of ethics training in the sciences, program directors might benefit from giving stronger consideration to identifying ethics instructors that offer the combination of expertise in both training delivery as well as the nuances of ethics as applied within the profession.

Features of instructional content also appeared to make a difference regarding training benefits (Mulhearn et al., 2016). For example, in-house development of instructional content, guided by a small number of broadly defined instructional objectives, supportedthe development of viable ethics training programs.Further, two paths emerged with regard to the field-specificity of training content that demonstrated the most effective results. That is, both field-specific and field-general programs showed sizable effects. Programs that attempted to achieve a mix, or balance, of field-specific and field-general content, however, evidenced less value to participants. Examination of the specific areas of instructional content showed a similar pattern. For example, courses that focused on professional guidelines as they apply within a particular field—such as, authorship and publication practices, research design, data management, data integrity, intellectual property—showed sizable effects. On the other hand, courses that focused on the application of field-general guidelines—such as the Common Rule, institutional compliance, and field differences—proved of similar value. Courses that focused on content areas that were less relevant to the application of professional guidelines were of limited value (e.g., statistical analysis, power differentials, diversity, organizational values, civil maturity, lab safety, and community issues). It is important to keep in mind, however, that many of the topics showing the weakest effects were observed in few courses. Finally, courses that emphasized process-based content showed moderate to large effects. Specifically, asking trainees to practice forecasting downstream consequences and the impact of emotions on their decisions proved of particular value, whereas analysis of personal values, stakeholders, and constraints proved of less value.

A number of findings also emerged with regard to more and less effective delivery methods and activities. Students, or trainees, in RCR education courses appear to benefit most when training emphasizes individual-based, as opposed to group-based, activities that encourage at least a moderate degree of active participation. Regarding specific delivery activities, it is noteworthy that virtually no, single activity, on its own, showed sizable effects. Thus, courses that incorporate a variety of focused activities that encourage active participation appear especially effective. Nevertheless, some activities proved more beneficial than others. For example, activities that encouraged active processing of training content, such as class debates, note-taking, and individual workbooks, all showed moderate to large gains. On the other hand, activities focused on social interaction, such as mentoring, service learning, and games, proved of less value. Finally, courses that emphasized instruction vis-à-vis cases, particularly longer cases of moderate complexity and low to moderate realism and emotional content, also showed sizable benefits to participants.

In addition to content and delivery methods, characteristics of the trainees were also associated with training effectiveness. Training proved most beneficial for students reporting no prior exposure to formal ethics instruction. Thus, programs that require re-training on a

continuing basis appear of some, albeit limited, value. Further, older students (30+ years old) appeared to benefit most from ethics training. This finding may be of some concern considering universities are often expected to play a primary role in educating scientists to learn the ethical norms and guidelines of their professions during the undergraduate and graduate school years. Sub-disciplines within the sciences also showed some, albeit minor, differences with regard to instructional effectiveness. A more substantive difference was observed between training programs that attempted to cater to trainees from multiple fields (d = .20) versus those that focused on trainees within a particular field (d = .45 to .66). In this case, it would appear some training efficacy is being sacrificed in the name of efficiency —a point that should be of interest to ethics educators that are responsible for identifyingeffective and efficient means of educating scientists across multiple disciplines. Finally, trainee gender also showed a clear pattern—men benefited more from training than women. This finding may perhaps be attributed to women's higher initial ethical decision making(Borkowski & Ugras, 1998; O'Fallon & Butterfield, 2005).

Another noteworthy finding emerged with regard to program evaluation. That is, identifying and selecting appropriate criteria for evaluating ethics instruction programs may be as important as training content decisions for demonstrating program effectiveness (Steele et al., 2016).Programs that attempted to improve trainees' perceptions of others, conceptual development, or moral judgment showed some of the weakest effects, whereas programs focusing on improving knowledge, perceptions of self, and ethical decision making showed the largest effects.More broadly, custom-developed criteria(i.e., in-house)targeting specific fields demonstrated the largest effects. Alternatively, off-the-shelf, general measures (e.g., DIT-2, MJT, PCM) demonstrated some of the smallest effects. Along these lines, one finding emerged that was of some surprise, given the considerable resources that have been devoted to RCR education in the United States over the last three decades. That is, international courses showed larger effects than U.S. courses. It is noteworthy, however, that little variance was accounted for by sampling error (i.e., 8 to 12%). Thus, the presence of additional moderators was suggested. We examined if these differences in domestic and international programs might be accounted for by differences in criteria employed. Indeed, it was found that international programs tended to emphasize the use of knowledge measures (d = .78), whereas domestic programs more commonly employed assessments of moral reasoning (e.g., DIT; d = .39). In other words, because these two general criterion types showed differences in the magnitude of effect sizes observed and also varied proportionally in the frequency with which they were used by international and U.S. ethics programs, differences in criteria may account for some of the observed differenceshere.

In conclusion,ethics education programs in the sciences, on the whole, appear to be improving. Indeed, we found that evaluation studies of programs and courses published in the last decade have demonstrated larger gains for participants in the way of knowledge, skills, and attitudes, compared with the gains observed amongolder studies examined in prior meta-analytic work (Antes et al., 2009; Waples et al., 2009). However, the programs examined also varied considerably with regard to the magnitude of their effectiveness. For example, programs that delivered instructional content using multiple, expert trainers, integrated cases, and incorporatedpractice activities that encouraged active trainee participation demonstrated particularly large effects. Indeed, the variability in effect sizes

observed leads to an important conclusionregarding the current state of RCR education: there is still ample room for improvement. Finally, the present study serves to demonstrate the value of conducting broad, systematic evaluation efforts in regular intervals to allow for benchmarking the effects of ethics education programs over time (Mumford et al., 2015). We hope the present study serves as a source of data-based recommendations for improving RCR education and as a model for future evaluation research along these lines.

Acknowledgments

We would like to thank Allison Antes, Jason Borenstein, Jeffrey Engler, Michael Kalichman, Brian Martinson, and Michael Verderame for their contributions to the present effort. The project described was supported by Grant Number ORIIR140010–01−00 from the National Institutes of Health and the Office of Research Integrity. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the Department of Health and Human Services or the Office of Research Integrity.

References

*Al-Jalahma M & Fakhroo E (2004). Teaching medical ethics: Implementation and evaluation of a new course during residency training in Bahrain. Education for Health, 17, 62–72. [PubMed: 15203475]

*Alfred M (2010). Design and development of an interactive multimedia training simulator for engineering ethics education (SEEE). Unpublished dissertation University of Houston.

Antes AL (2014). A systematic approach to instruction in research ethics. Accountability in Research, 21, 50–67. [PubMed: 24073607]

Antes AL, Murphy ST, Waples EP, Mumford MD, Brown RP, Connelly S, & Devenport LD (2009). A meta-analysis of ethics instruction effectiveness in the sciences. Ethics & Behavior, 19, 379–402. [PubMed: 19838311]

Antes AL, Wang X, Mumford MD, Brown RP, Connelly S, & Devenport LD (2010). Evaluating the effects that existing instruction on responsible conduct of research has on ethical decision making. Academic Medicine, 85, 519–526. [PubMed: 20182131]

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

*Austin KA Gorsuch GJ Lawson WD & Newberry BP (2011). Developing and designing online engineering ethics instruction for international graduate students. Instructional Science, 39, 975–997.

*Avila SA (2006). Effects of intervention on early childhood educator ethical sensitivity and judgment Unpublished dissertation. California State University, Fresno.

Bagdasarov Z, Thiel CE, Johnson JF, Connelly S, Harkrider LN, Devenport LD, & Mumford MD (2013). Case-based ethics instruction: The influence of contextual and individual factors in case content on ethical decision-making. Science and Engineering Ethics, 19, 1305–1322. [PubMed: 23143838]

*Barchi FH Kasimatis-Singleton M Kasule M Khulumani P & Merz JF (2013). Building research capacity in Botswana: A randomized trial comparing training methodologies in the Botswana ethics training initiative. BMC Medical Education, 13, 1–7. [PubMed: 23286697]

*Baykara ZG Demir SG & Yaman S (2015). The effect of ethics training on students recognizing ethical violations and developing moral sensitivity. Nursing Ethics, 22, 661–675. [PubMed: 25096245]

*Bebeau MJ & Thoma SJ (1994). The impact of a dental ethics curriculum on moral reasoning. Journal of Dental Education, 58, 684–692. [PubMed: 7962920]

Bernardin HJ, & Buckley MR (1981). Strategies in rater training. Academy of Management Review, 6, 205–212.

*Bernstein D De George R Douglas M Rosenbloom JL Starrett S Anderegg A & Luth M (2010). The University of Kansas initiative in ethics education in science and engineering: Final report. Unpublished manuscript.

*Borenstein, J Drake M Kirkman R & Swann J (2010). The engineering and science issues test (ESIT): A discipline-specific approach to assessing moral judgment. Science and Engineering Ethics, 16, 387–407. [PubMed: 19597969]

Borenstein M, Hedges LV, Higgins JP, & Rothstein HR (2009). Introduction to meta-analysis. West Sussex: Wiley.

Borkowski SC, & Ugras YJ (1998). Business students and ethics: A meta-analysis. Journal of Business Ethics, 17, 1117–1127.

*Borovečki A ten Have H & Orešković S (2006). Education of ethics committee members: Experiences from Croatia. *Journal of Medical Ethics*, 32, 138–142. [PubMed: 16507656]

*Brkic S Bogdanovic G Vuckovic-Dekic L Gavrilovic D & Kezic I (2012). Science ethics education: Effects of a short lecture on plagiarism on the knowledge of young medical researchers. Journal of the Balkan Union of Oncology, 17, 570–574.

*Brock ME Vert A Kligyte V Waples EP Sevier ST & Mumford MD. (2008). Mental models: An alternative evaluation of a sensemaking approach to ethics instruction. Science and Engineering Ethics, 14, 449–472. [PubMed: 18568427]

*Cain JJ (2007). An instructional design treatment of online, asynchronous threaded discussions to increase moral reasoning skills Unpublished dissertation. University of Kentucky.

*Canary HE (2007). Teaching ethics in communication courses: An investigation of instructional methods, course foci, and student outcomes. Communication Education, 56, 193–208.

*Canary HE & Ellison K (2013). Macroethics modules for the CITI responsible conduct of research courses.

*Canary HE Herkert JR Ellison K & Wetmore JM (2012, 6). Microethics and macroethics in graduate education for scientists and engineers: Developing and assessing instructional models Paper presented at the 119th Annual Conference & Exposition for the American Society for Engineering Education, San Antonio, TX.

*Chase NM (1998). A cognitive development approach to professional ethics training for counselor education students*Unpublished dissertation*. The College of William and Mary.

*Cho KC & Shin G (2014). Operational effectiveness of blended e-learning program for nursing research ethics. Nursing Ethics, 21, 484–495. [PubMed: 24258252]

*Chung C (2015). Comparison of cross culture engineering ethics training using the simulator for engineering ethics education. Science and Engineering Ethics, 21, 471–478. [PubMed: 24718714]

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

*Clarkeburn HM Downie J Gray C & Matthew RS (2003). Measuring ethical development in life sciences students: A study using Perry's developmental model. Studies in Higher Education, 28, 443–456.

Cohen J (1992). A power primer. Psychological Bulletin, 112, 155–159. [PubMed: 19565683]

Craft JL (2013). A review of the empirical ethical decision-making literature: 2004–2011. Journal of Business Ethics, 117, 221–259.

*Cummings R Maddux CD Cladianos A &Richmond A (2010). Moral reasoning of education students: The effects of direct instruction in moral development theory and participation in moral dilemma discussion.Teachers College Record, 112, 621–644.

Dane E, & Sonenshein S (2015). On the role of experience in ethical decision making at work: An ethical expertise perspective. Organizational Psychology Review, 5, 74–96.

*Drake MJ Griffin PM Kirkman R & Swann JL (2005). Engineering ethical curricula: Assessment and comparison of two approaches. Journal of Engineering Education, 94, 223–231.

*DuBois JM Dueker JM Anderson EE & Campbell J (2008). The development and assessment of an NIH-funded research ethics training program. Academic Medicine, 83, 596–603. [PubMed: 18520469]

Ericsson KA, Krampe RT, & Tesch-Römer C (1993). The role of deliberate practice in the acquisition of expert performance. Psychological Review, 100, 363–406.

*Evans B & Bendel R (2004). Cognitive and ethical maturity in baccalaureate nursing students: Did a class using narrative pedagogy make a difference? Nursing Education Perspectives, 25, 188–195. [PubMed: 15387514]

*Frisch NC (1987). Value analysis: A method for teaching nursing ethics and promoting the moral development of students. The Journal of Nursing Education, 26, 328–332. [PubMed: 2824724]

*Gadbury-Amyot CC Simmer-Beck M McCunnif M & Williams KB (2006). Using a multifaceted approach including community-based service-learning to enrich formal ethics instruction in a dental school setting. Journal of Dental Education, 70, 652–666. [PubMed: 16741133]

*Gaul AL (1987). The effect of a course in nursing ethics on the relationship between ethical choice and ethical action in baccalaureate nursing students. The Journal of Nursing Education, 26, 113–117. [PubMed: 3035118]

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

*Goldie J Schwartz L McConnachie A & Morrison J (2002). The impact of three years' ethics teaching, in an integrated medical curriculum, on students′ proposed behaviour on meeting ethical dilemmas. Medical Education, 36, 489–497. [PubMed: 12028400]

*Goldman SA & Arbuthnot J (1979). Teaching medical ethics: The cognitive-developmental approach.Journal of Medical Ethics, 5, 170–181. [PubMed: 541818]

*Grant TA (2012). Comparing the principle-based SBH maieutic method to traditional case study methods of teaching media ethics Unpublished dissertation. University of Idaho.

Haws DR (2001). Ethics instruction in engineering education: A (mini) meta-analysis. Journal of Engineering Education, 90, 223–229.

Hedges LV, & Olkin I (1985). Statistical methods for meta-analysis. New York: Academic Press.

Hedges LV, & Vevea JL (1998). Fixed- and random-effects models in meta-analysis. Psychological Methods, 3, 486–504.

Heitman E, Olsen CH, Anestidou L, & Bulger RE (2007). New graduate students' baseline knowledge of the responsible conduct of research. Academic Medicine, 82, 838–845. [PubMed: 17726387]

Hunt DE (1977). Assessing conceptual level by the paragraph completion method. Ontario Institute for Studies in Education.

Hunter JE, & Schmidt FL (2004). Methods of meta-analysis: Correcting error and bias in research findings. Sage.

Johnson JF, Bagdasarov Z, Connelly S, Harkrider L, Devenport LD, Mumford MD, & Thiel CE (2012). Case-based ethics education: The impact of cause complexity and outcome favorability on ethicality. Journal of Empirical Research on Human Research Ethics, 7, 63–77.

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

*Jurkiewicz CL. (2002). The influence of pedagogical style on students' level of ethical reasoning. Journal of Public Affairs Education, 8, 263–274.

*Kavathatzopoulos I (1994). Training professional managers in decision-making about real life business ethics problems: The acquisition of the autonomous problem-solving skill. Journal of Business Ethics, 13, 379–386.

*Keefer M Wilson S Dankowicz H & Loui M (2014). The importance of formative assessment in science and engineering ethics education: Some evidence and practical advice. *Science &* Engineering Ethics, 20, 249–260.

Kepes S, McDaniel MA, Brannick MT, & Banks GC (2013). Meta-analytic reviews in the organizational sciences: Two meta-analytic schools on the way to MARS (the Meta-analytic Reporting Standards). Journal of Business and Psychology, 28, 123–143.

*Kligyte V Marcy RT Waples EP Sevier ST Godfrey ES Mumford MD & Hougen DF (2008). Application of a sensemaking approach to ethics training in the physical sciences and engineering. Science and Engineering Ethics, 14, 251–278. [PubMed: 18074243]

*Lin C Lu M Chung C & Yang C (2010). A comparison of problem-based learning and conventional teaching in nursing ethics education. Nursing Ethics, 17, 373–382. [PubMed: 20444778]

Lind G (1998). An introduction to the Moral Judgment Test (MJT) Unpublished manuscript. Konstanz: University of Konstanz.

*Litzky BE & Oz E (2008). Ethical issues in information technology: Does education make a difference? International Journal of Information and Communication Technology Education, 4, 67–83.

*Loui MC (2009). Ethics assessment: Information trust institute summer undergraduate internship program. Unpublished manuscript.

*Major-Kincade TL Tyson JE & Kennedy KA (2000). Training pediatric house staff in evidence-based ethics: An exploratory controlled trial. Journal of Perinatology, 21, 161–166.

*Martin T Rayne K Kemp NJ Hart J & Diller KR (2005). Teaching for adaptive expertise in biomedical engineering ethics. Science and Engineering Ethics, 11, 257–276. [PubMed: 15915863]

*McCormack WT Allen WL Connelly S Coolen LM Crites J Engler J Garvan CW Haidet P Hockensmith J McElroy W Volpe R & Verderame MF (2015).Team-based learning for RCR training supports ethical decision-making. Unpublished conference poster.

*McCormack WT& Garvan CW (2014). Team-based learning instruction for responsible conduct of research positively impacts ethical decision-making. Accountability in Research: Policies & Quality Assurance, 21, 34–49.

*McKellar KA (1998). Ethical decision-making: Does practice make a difference? *Unpublished dissertation*. St. Mary's University of San Antonio.

Medeiros KE, Watts LL, Mulhearn TJ, Steele LM, Mumford MD, & Connelly S (2016).A meta-analyticreview of business ethics education from 1988 to 2015.Manuscript under review.

*Morgan B Morgan F Foster V & Kolbert J (2000). Promoting the moral and conceptual development of law enforcement trainees: A deliberate psychological educational approach. Journal of Moral Education, 29, 203–218.

Mulhearn TJ, Steele LM, Watts LL, Medeiros KE, Mumford MD, & Connelly S (2016). Review of instructional approaches in ethics education. Manuscript under review.

*Mumford (2015).*Professional ethics training: Mean comparisons of ethics training effectiveness in 2014–2015*. Retrieved from The University of Oklahoma.

*Mumford MD Connelly S Brown RP Murphy ST Hill JH Antes AL Waples EP & Devenport LD (2008). A sensemaking approach to ethics training for scientists: Preliminary evidence of training effectiveness. Ethics & Behavior, 18, 315–339. [PubMed: 19578559]

Mumford MD, Devenport LD, Brown RP, Connelly S, Murphy ST, Hill JH, & Antes AL (2006). Validation of ethical decision making measures: Evidence for a new set of measures. Ethics & Behavior, 16, 319–345.

Mumford, M. D., Watts, L. L., Medeiros, K. E., Mulhearn, T. J., Steele, L. M., & Connelly, S. (in review). Biomedical ethics education may benefit from integrating compliance and analysis approaches.

*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 MJ, & Butterfield KD (2005). A review of the empirical ethical decision-making literature: 1996–2003. Journal of Business Ethics, 59, 375–413.

*Ogunrin OA Ogundiran TO & Adebamowo C (2013). Development and pilot testing of an online module for ethics education based on the Nigerian national code for health research ethics. BMC Medical Ethics, 14, 1–17. [PubMed: 23281968]

*Powell ST Allison MA & Kalichman MW (2007). Effectiveness of a responsible conduct of research course: a preliminary study. Science and Engineering Ethics, 13, 249–264. [PubMed: 17717736]

*Ramalingam S Bhuvaneswari S & Sankaran R (2014). Ethics workshops: Are they effective in improving competencies of faculty and postgraduates? Journal of Clinical and Diagnostic Research, 8, 1–3.

Rest JR (1979). Development in judging moral issues. Minneapolis, MN: University of Minnesota Press.

Rest JR (1986).Moral development: Advances in research and theory. New York: Praeger.

Rest JR, Narvaez D, Thoma SJ, & Bebeau MJ (1999). DIT2: Devising and testing a revised instrument of moral judgment. Journal of Educational Psychology, 91, 644–659.

*Roberts LW Warner TD Dunn LB Brody JL Hammond KAG & Roberts BB (2007). Shaping medical students' attitudes towards ethically important aspects of clinical research: Results of a randomized, controlled educational intervention. Ethics & Behavior, 17, 19–50.

*Roberts LW Warner TD Hammond KAG Brody JL Kaminsky A & Roberts BB (2005). Teaching medical students to discern ethical problems in human clinical research studies. Academic Medicine, 80, 925–930. [PubMed: 16186612]

*Rozmus CL Carlin N Polczynski A Spike J & Buday R (2015). The Brewsters: A new resource for interprofessional ethics education. Nursing Ethics, 22, 815–826. [PubMed: 25252587]

*Ryden MB & Duckett L (1991). Ethics education for baccalaureate nursing students. *Unpublished research report* University of Minnesota, Minneapolis.

*Rzymska I Rzymski R Wilczak M Wloszcak-Szubda A Jarosz MJ & Musielak, M (2014). The influence of passive and active moral training on medical university on changes of students' moral competence index: Results from randomized single blinded trial. Annals of Agricultural and Environmental Medicine, 21, 161–166. [PubMed: 24738517]

*Sanders S & Hoffman K (2010). Ethics education in social work: Comparing outcomes of graduate social work students. Journal of Social Work Education, 46, 7–22.

*Schmaling KB & Blume AW (2009). Ethics instruction increases graduate students' responsible conduct of research knowledge but not moral reasoning.Accountability in Research, 16, 268–283. [PubMed: 19757232]

*Schuh L & Burdette D (2004). Initiation of an effective neurology resident ethics curriculum. Neurology, 62, 1897–1898. [PubMed: 15159507]

*Self DJ & Wolinsky FD (1992). Evaluation of teaching medical ethics by an assessment of moral reasoning. Medical Education, 26, 178–184. [PubMed: 1614342]

Sitzmann T, Kraiger K, Stewart D, & Wisher R (2006). The comparative effectiveness of web-based and classroom instruction: A meta-analysis. Personnel Psychology, 59, 623–664.

Steele LM, Mulhearn TJ, Medeiros KE, Watts LL, Mumford MD, & Connelly S (2016). How do we know what works? A review and critique of current practices in ethics training evaluation. Manuscript under review

*Swisher LL van Kessel G Jones M Beckstead J & Edwards I (2012). Evaluating moral reasoning outcomes in physical therapy ethics education: Stage, schema, phase, and type. Physical Therapy Reviews, 17, 167–175.

Thiel CE, Connelly S, Harkrider L, Devenport LD, Bagdasarov Z, Johnson JF, & Mumford MD (2013). Case-based knowledge and ethics education: Improving learning and transfer through emotionally rich cases. Science and Engineering Ethics, 19, 265–286. [PubMed: 22038062]

*Thirunavukarasu P Brewster L Pecora S & Hall D (2010). Educational intervention is effective in improving knowledge and confidence in surgical ethics: A prospective study. American Journal of Surgery, 200, 665–669. [PubMed: 21056150]

*Uthe-Burow C (2002). An exploratory study of ethical training as a factor of moral development Unpublished dissertation. University of South Dakota.

*Vartiainen T Siponen M & Myyry L (2011). The effects of teaching the universality thesis on students' integrative complexity of thought. Journal of Information Systems Education, 22, 261–270.

Waples EP, Antes AL, Murphy ST, Connelly S, & Mumford MD (2009). A meta-analytic investigation of business ethics instruction. Journal of Business Ethics, 87, 133–151.

Weber J (2015). Investigating and assessing the quality of employee ethics training programs among US-based global organizations. Journal of Business Ethics, 129, 27–42.

Weber J, & Wasieleski DM (2013). Corporate ethics and compliance programs: A report, analysis and critique. Journal of Business Ethics, 112, 609–626.

*Wilson WR (2013). Using the Chernobyl incident to teach engineering ethics. Science Engineering Ethics, 19, 625–640. [PubMed: 22170503]

Wilson RJGlaser JW Rasinski-Gregory D Gibson MJ & Bayley C (1993). Education for ethics committees: What to learn and how to teach. *Health care ethics committees—The next generation* (pp. 45–68).San Francisco: Jossey-Bass.

*Workman M & Gathegi (2007). Punishment and ethics deterrents: A study of insider security contravention. Journal of the American Society for Information Science and Technology, 58, 212–222.

Table 1.

Overall Meta-Analysis and Criterion Type

			Weighted		Var. (%) due to sampling error	95% CI	
	K	*N*	*Md*	*SD*		Lower	Upper
Instruction Effectiveness							
Overall	150	10,069	.48	.04	11	.40	.57
Time of Publication							
Before 2007	58	3,632	.36	.07	12	.23	.49
2007 – 2015	92	6,437	.56	.06	10	.45	.67
General Criterion Type							
Moral Reasoning	47	3,469	.39	.09	7	.21	.57
Knowledge	27	1,457	.78	.14	9	.51	1.04
Ethical Awareness	16	1,347	.44	.08	48	.28	.59
Ethical Decision Making	15	803	.51	.12	17	.27	.74
Perceptions of Self	14	1,022	.66	.19	9	.30	1.03
Moral Judgment	13	663	.25	.06	43	.13	.37
Meta-cognitive Strategies	8	811	.51	.13	22	.26	.76
Conceptual Development	6	317	.24	.18	36	-.11	.58
Perceptions of Others	4	180	-.01	.11	100	-.22	.21

Note. *k* = number of effect sizes; *N* = total sample size; *Md* = Inverse variance-weighted (Hedges & Olkin, 1985), uncorrected mean effect size (*d*); *SD* = Standard deviation of mean effect size; Variance (%) due to sampling error = 100 - I^2; CI = Confidence interval.

Table 2.

Characteristics of Criteria

		Weighted			Var. (%) due to sampling error	95% CI	
	K	***N***	***Md***	***SD***		**Lower**	**Upper**
Specific Criterion Type							
DIT	28	1,806	.33	.09	11	.15	.52
DIT2	13	821	.16	.15	14	-.14	.47
Field-specific DIT	7	854	1.14	.35	6	.46	1.82
EDM	5	555	.37	.13	19	.12	.62
PCM	4	113	.19	.12	100	-.04	.42
MJT	4	242	.16	.06	100	.03	.28
Reported Reliability							
No	94	7,215	.46	.06	9	.36	.57
Yes	56	2,854	.52	.07	17	.38	.66
Measure Development							
Off-the-shelf	67	4,216	.32	.05	15	.22	.43
In-house	42	2,769	.77	.11	9	.54	.99
Measure Field-specificity							
Field-specific	31	2,153	.70	.11	8	.48	.92
Combination	20	1,316	.46	.14	11	.18	.73
General	73	4,374	.29	.06	14	.17	.40

Note. *k* = number of effect sizes; *N* = total sample size; *Md* = Inverse variance-weighted (Hedges & Olkin, 1985), uncorrected mean effect size (*d*); *SD* = Standard deviation of mean effect size; Variance (%) due to sampling error = 100 - I^2; CI = Confidence interval; DIT = Defining Issues Test (Rest, 1979); DIT2 = Defining Issues Test 2 (Rest et al., 1999); Field-specific DIT = Field-specific Defining Issues Test; EDM = Ethical Decision Making (Mumford et al., 2006); PCM = Paragraph Completion Method (Hunt et al., 1978); MJT = Moral Judgment Test (Lind, 2002).

Table 3.

Study Design Characteristics

		Weighted			Var. (%) due to sampling error	95% CI	
	K	*N*	*Md*	*SD*		Lower	Upper
Design Type							
Pre-Post w/ Control	41	3,659	.47	.12	10	.23	.70
Pre-Post	83	4,308	.52	.05	10	.42	.63
Post w/ Control	26	2,102	.31	.08	44	.17	.46
Skill Decay (Post 2 - Post 1)	8	230	-.02	.06	66	-.14	.10
Sample Size							
Less than 50	84	2,608	.48	.05	31	.39	.57
50 – 100	44	2,997	.43	.07	13	.29	.56
100+	22	4,464	.56	.16	3	.25	.87
Study Funded							
No	106	6,921	.47	.05	13	.37	.56
Yes	44	3,148	.52	.10	8	.32	.71
Publication Type							
Unpublished	49	3,487	.25	.05	32	.15	.36
Peer Reviewed	101	6,582	.59	.06	9	.48	.70
Impact Factor							
Below Average (< 0.963)	10	658	.51	.07	100	.37	.64
Average (0.964 – 1.516)	26	1,646	.70	.10	11	.52	.89
Above Average (> 1.517)	15	585	.39	.11	22	.19	.60

Note. *k* = number of effect sizes; *N* = total sample size; *Md* = Inverse variance-weighted (Hedges & Olkin, 1985), uncorrected mean effect size (*d*); *SD* = Standard deviation of mean effect size; Variance (%) due to sampling error = 100 - I^2; CI = Confidence interval.

Table 4.

Characteristics of Participants

		Weighted			Var. (%) due to sampling error	95% CI	
	K	***N***	***Md***	***SD***		**Lower**	**Upper**
Career Stage							
Undergraduate Students	39	3,183	.53	.12	8	.31	.76
Graduate Students	87	5,945	.44	.05	13	.34	.54
Professionals/Residents	17	667	.51	.11	25	.29	.73
Mixed	6	237	.63	.28	6	.09	1.17
Prior Instruction							
No	7	313	1.13	.28	17	.58	1.69
Yes	30	952	.46	.08	26	.30	.62
Gender							
More than 70% Males	22	2,003	.66	.15	5	.36	.95
Mixed Gender	51	3,817	.43	.08	14	.28	.58
More than 70% Females	25	1,218	.36	.11	19	.14	.57
Average Age							
18 – 25	40	2,305	.38	.05	27	.27	.48
26 – 30	19	991	.37	.09	48	.18	.55
30+	16	771	.63	.15	14	.32	.93
Field							
Health/Medicine	56	3,302	.50	.07	11	.37	.63
Engineering	43	2,724	.66	.12	10	.43	.89
Social Science	26	1,216	.45	.12	11	.21	.69
Mixed	20	2,131	.20	.06	25	.08	.33
Location of Instruction							
Domestic (U.S.)	125	8,536	.42	.05	12	.33	.52
International	25	1,533	.79	.12	8	.56	1.02

Note. *k* = number of effect sizes; *N* = total sample size; *Md* = Inverse variance-weighted (Hedges & Olkin, 1985), uncorrected mean effect size (*d*); *SD* = Standard deviation of mean effect size; Variance (%) due to sampling error = 100 - I^2; CI = Confidence interval.

Table 5.

Quality Ratings

			Weighted		Var. (%) due to sampling error	95% CI	
	K	***N***	***Md***	***SD***		**Lower**	**Upper**
Rating of Instruction							
Below Average	62	4,469	.38	.07	12	.25	.52
Average	65	3,705	.62	.08	9	.47	.77
Above Average	23	1,895	.39	.07	24	.25	.52
Rating of Study Design							
Below Average	50	2,620	.53	.08	8	.38	.68
Average	34	2,921	.59	.10	7	.39	.79
Above Average	36	2,198	.38	.08	25	.23	.53
Rating of Criterion							
Below Average	76	4,266	.57	.07	13	.44	.70
Average	47	3,279	.28	.07	12	.15	.41
Above Average	18	1,517	.50	.08	25	.36	.65

Note. *k* = number of effect sizes; *N* = total sample size; *Md* = Inverse variance-weighted (Hedges & Olkin, 1985), uncorrected mean effect size (*d*); *SD* = Standard deviation of mean effect size; Variance (%) due to sampling error = 100 - I^2; CI = Confidence interval.

Table 6.

General Instructional Parameters

			Weighted		Var. (%) due to sampling error	95% CI	
	K	***N***	***Md***	***SD***		**Lower**	**Upper**
Total Hours of Instruction							
Less than 8 Hours	47	3,528	.61	.09	11	.43	.79
Between 8 and 16 Hours	36	2,439	.50	.07	17	.37	.64
More than 16 Hours	65	4,040	.39	.07	9	.25	.53
Instructional Format							
Integrated	46	3,150	.44	.08	11	.29	.59
Stand-alone	104	6,919	.51	.05	11	.40	.61
Delivery Format							
Face-to-face	101	6,346	.44	.05	12	.33	.54
Hybrid	11	572	.77	.16	19	.46	1.09
Online	15	1,276	.38	.10	27	.20	.57
Average Class Size							
Up to 20	46	2,832	.46	.07	28	.33	.59
21 – 40	47	3,021	.39	.08	18	.24	.54
40+	51	3,688	.55	.08	8	.40	.70
Instruction Mandatory							
No	59	3,353	.52	.09	13	.35	.69
Yes	46	3,846	.26	.03	40	.19	.33

Note. *k* = number of effect sizes; *N* = total sample size; *Md* = Inverse variance-weighted (Hedges & Olkin, 1985), uncorrected mean effect size (*d*); *SD* = Standard deviation of mean effect size; Variance (%) due to sampling error = 100 - I^2; CI = Confidence interval.

Table 7.

Characteristics of Trainers

		Weighted			Var. (%) due to sampling error	95% CI	
	k	***N***	***Md***	***SD***		**Lower**	**Upper**
Number of Trainers							
1	40	2,163	.36	.09	10	.19	.54
2	21	1,452	.46	.07	40	.32	.60
2+	7	547	1.07	.24	9	.59	1.54
Trainer Expertise							
Below Average	15	1,241	.35	.06	50	.23	.47
Average	17	1,433	.66	.14	11	.38	.95
Above Average	6	465	.87	.32	7	.23	1.50
Trainer Rotation							
No	33	2,094	.60	.11	8	.38	.81
Yes	26	1,825	.44	.07	37	.31	.57

Note. *k* = number of effect sizes; *N* = total sample size; *Md* = Inverse variance-weighted (Hedges & Olkin, 1985), uncorrected mean effect size (*d*); *SD* = Standard deviation of mean effect size; Variance (%) due to sampling error = 100 - I^2; CI = Confidence interval.

Table 8.

Characteristics of Instructional Development

			Weighted		Var. (%) due to sampling error	95% CI	
	K	***N***	***Md***	***SD***		**Lower**	**Upper**
Development							
Mostly Off-the-shelf	9	358	.40	.21	9	-.01	.80
Mixed	34	2,004	.40	.05	29	.30	.50
Mostly In-house	70	4,548	.54	.07	10	.39	.68
Number of Objectives							
1 – 3	33	2,266	.66	.08	12	.50	.83
3+	29	1,710	.44	.08	19	.29	.60
Specificity of Objectives							
Below Average	9	798	.62	.24	10	.15	1.09
Average	29	1,347	.64	.13	14	.38	.89
Above Average	15	930	.46	.08	22	.31	.61

Note. *k* = number of effect sizes; *N* = total sample size; *Md* = Inverse variance-weighted (Hedges & Olkin, 1985), uncorrected mean effect size (*d*); *SD* = Standard deviation of mean effect size; Variance (%) due to sampling error = 100 - I^2; CI = Confidence interval.

Table 9.

Characteristics of Instructional Content

			Weighted		Var. (%) due to sampling error	95% CI	
	k	***N***	***Md***	***SD***		**Lower**	**Upper**
Presented Content Details							
No	36	3,144	.36	.08	10	.20	.51
Yes	114	6,925	.53	.05	11	.42	.63
Coverage of 9 ORI Guidelines							
1 − 3 Guidelines	60	3,979	.41	.06	18	.30	.53
3+ Guidelines	28	2,054	.56	.10	8	.36	.77
Breadth of Content Covered							
1 − 6 Topics	57	4,100	.46	.08	8	.29	.62
7 − 12 Topics	52	2,871	.48	.07	16	.34	.62
12+ Topics	30	2,371	.57	.09	11	.39	.74
Field-specificity of Content							
Field-specific	42	2,299	.61	.09	10	.43	.78
Balanced	55	3,483	.22	.05	25	.12	.32
General	32	3,193	.82	.11	6	.60	1.05
Emphasis on Processes							
Below Average	42	3,246	.45	.09	9	.28	.63
Average	28	1,651	.63	.13	6	.38	.89
Above Average	18	1,150	.55	.11	19	.35	.76
Number of Stakeholders							
Below Average	22	1,299	.48	.08	41	.32	.63
Average	19	810	.36	.09	35	.18	.54
Above Average	10	1,162	.37	.10	16	.17	.57

Note. *k* = number of effect sizes; *N* = total sample size; *Md* = Inverse variance-weighted (Hedges & Olkin, 1985), uncorrected mean effect size (*d*); *SD* = Standard deviation of mean effect size; Variance (%) due to sampling error = 100 - I^2; CI = Confidence interval.

Table 10.

Specific Instructional Content

		Weighted			Var. (%) due to sampling error	95% CI	
	k	*N*	*Md*	*SD*		Lower	Upper
Guidelines	85	5,698	.57	.06	9	.45	.70
Codes of conduct	70	3,970	.53	.06	11	.40	.65
Common rule	10	1,437	.78	.23	6	.34	1.22
Belmont report	9	1,646	.50	.16	5	.20	.81
Nuremberg code	2	131	1.58	.52	9	.56	2.59
Protect. of human subjects	65	4,315	.44	.06	14	.32	.55
Protect. of animal subjects	25	2,975	.47	.09	12	.30	.64
Privacy and confidentiality	37	2,009	.57	.09	12	.40	.75
FFP	31	2,763	.58	.10	8	.38	.78
Authorship & publication	28	1,953	.60	.11	7	.38	.82
Peer review	21	1,496	.19	.05	63	.10	.29
General compliance	33	1,930	.46	.09	17	.29	.63
Institutional compliance	22	1,488	.60	.12	7	.37	.84
External reg. compliance	12	390	.55	.18	14	.19	.90
International reg. compliance	3	84	.37	.11	95	.16	.59
Legality	37	2,401	.37	.06	30	.25	.49
Financial compliance	2	60	.88	.35	19	.19	1.58
Sensemaking	14	1,275	.46	.10	25	.28	.65
Constraints	6	818	.46	.13	17	.21	.71
Personal biases	13	1,111	.56	.11	25	.36	.77
Group biases	3	452	.84	.16	42	.53	1.15
Personal integrity	10	561	.96	.23	16	.51	1.40
Professionalism	24	1,271	.43	.13	10	.18	.69
Maintaining objectivity	7	874	.48	.13	19	.24	.73
Strategies	11	1,163	.55	.10	21	.35	.76
Scientific misconduct	27	2,391	.50	.10	10	.32	.69
Research design	16	753	.61	.17	10	.28	.93
Data sharing	13	1,160	.53	.12	15	.29	.76
Data management	27	2,041	.60	.11	8	.39	.81
Data integrity	11	903	.82	.20	9	.42	1.22
Intellectual property	11	1,205	.73	.15	7	.44	1.02
Statistical analysis	2	34	.17	.17	100	-.17	.51
Internet use & computing	10	682	.61	.12	49	.37	.86
Mentor mentee relationships	35	1,964	.33	.06	27	.21	.45
Power differentials	8	642	.18	.04	100	.11	.26
Leadership	2	192	.69	.35	16	.01	1.37
Collaboration	20	1,691	.40	.08	15	.23	.56
Field differences	10	1,007	.80	.19	6	.44	1.17

		Weighted			Var. (%) due to sampling error	95% CI	
	k	*N*	*Md*	*SD*		Lower	Upper
Cross-cultural differences	19	941	.37	.15	10	.08	.66
Diversity	7	331	.19	.09	56	.01	.36
Whistleblowing	11	1,214	.64	.22	4	.21	1.06
Stakeholders	51	3,271	.41	.05	28	.31	.51
Stakeholder culture	3	242	.44	.34	11	-.22	1.10
Conflicts of interest	48	2,308	.57	.08	12	.41	.73
Virtues or character	13	725	.57	.11	39	.36	.77
Values	17	627	.22	.06	87	.10	.34
Personal responsibility Accountability	18	741	.51	.10	35	.31	.72
Organizational values	6	147	.19	.17	100	−.14	.52
Moral maturity	10	380	.30	.11	45	.09	.51
Moral philosophy	49	3,043	.58	.08	9	.42	.73
Social responsibility	46	3,113	.45	.07	13	.32	.59
Civil maturity	6	331	.21	.13	48	−.05	.47
Scientists as a responsible member of society	17	1,598	.44	.08	26	.28	.60
Human rights	20	1,004	.50	.10	17	.30	.70
Environmental impacts	13	1,028	.47	.11	40	.25	.70
Safety (General)	10	1,202	.45	.12	19	.21	.69
Bioethics (General)	18	1,232	.48	.11	17	.28	.69
Lab safety	4	998	.19	.15	17	−.11	.49
Sexual harassment	3	143	1.60	.37	18	.88	2.33
Genetic engineering	4	153	.59	.28	34	.04	1.13
Stem cell research	7	343	.55	.12	45	.32	.77
Nature of ethical dilemmas	41	2,535	.43	.07	20	.30	.53
Historical development	33	2,799	.40	.09	12	.23	.58
Contemporary ethical issues	39	3,113	.67	.14	6	.39	.95
Community issues	17	722	.23	.06	80	.11	.35
Difference between ethics and other decisions	8	876	.55	.12	17	.31	.79

Note. *k* = number of effect sizes; *N* = total sample size; *Md* = Inverse variance-weighted (Hedges & Olkin, 1985), uncorrected mean effect size (*d*); *SD* = Standard deviation of mean effect size; Variance (%) due to sampling error = 100 - I^2; CI = Confidence interval.

Table 11.

Specific Process-Based Content

	Weighted				Var. (%) due to sampling error	95% CI	
	k	***N***	***Md***	***SD***		**Lower**	**Upper**
Ethical awareness	83	5,708	.54	.06	9	.42	.66
Consequences	26	1,521	.68	.11	10	.46	.90
Constraints	21	1,566	.40	.08	20	.25	.54
Forecasting	15	1,152	.71	.12	16	.48	.94
Motives	20	1,486	.45	.08	21	.29	.62
Strategies	31	2,348	.60	.11	7	.38	.82
Emotions	11	1,033	.76	.17	6	.42	1.09
Cognitive	20	1,136	.59	.10	18	.39	.79
Errors	7	674	.48	.12	25	.25	.71
Stakeholders	23	1,392	.41	.09	18	.22	.59
Meta-ethical	17	1,155	.46	.18	7	.11	.82
Values	12	559	.39	.15	32	.09	.69

Note. *k* = number of effect sizes; *N* = total sample size; *Md* = Inverse variance-weighted (Hedges & Olkin, 1985), uncorrected mean effect size (*d*); *SD* = Standard deviation of mean effect size; Variance (%) due to sampling error = 100 - I^2; CI = Confidence interval.

Table 12.

Characteristics of Delivery Methods and Activities

		Weighted			Var. (%) due to sampling error	95% CI	
	k	***N***	***Md***	***SD***		**Lower**	**Upper**
Presented Delivery Details							
No	43	2,964	.48	.09	10	.30	.67
Yes	107	7,105	.48	.05	12	.39	.58
Breadth of Activities							
1 − 3 Activities	51	3,569	.58	.10	8	.40	.77
4 − 6 Activities	44	3,149	.42	.06	17	.30	.53
6+ Activities	31	2,017	.46	.10	10	.25	.66
Whole vs. Part-learning							
Whole-learning	43	3,463	.46	.07	12	.33	.59
Mixed	35	2,255	.43	.10	10	.22	.63
Part-learning	40	2,144	.60	.09	12	.43	.77
Guided vs. Self-directed							
Guided	54	2,758	.40	.09	11	.23	.57
Mixed	39	3,118	.47	.06	15	.35	.59
Self-directed	32	2,259	.63	.10	11	.43	.82
Individual vs. Group-based							
Individual-based	53	3,413	.53	.09	11	.36	.70
Mixed	45	2,928	.52	.07	14	.39	.65
Group-based	23	1,574	.27	.07	28	.13	.40
Level of Active Participation							
Low	45	3,131	.41	.08	13	.25	.57
Moderate	50	3,038	.52	.07	13	.37	.66
High	23	1,323	.52	.12	12	.28	.76

Note. *k* = number of effect sizes; *N* = total sample size; *Md* = Inverse variance-weighted (Hedges & Olkin, 1985), uncorrected mean effect size (*d*); *SD* = Standard deviation of mean effect size; Variance (%) due to sampling error = 100 - I^2; CI = Confidence interval.

Table 13.

Specific Delivery Methods and Activities

		Weighted			Var. (%) due to sampling error	95% CI	
	k	***N***	***Md***	***SD***		**Lower**	**Upper**
Lecture	105	6,201	.43	.05	13	.33	.52
Repeated exposure	27	1,651	.58	.08	14	.43	.74
Moral method	10	640	.30	.12	28	.07	.53
Problem-based	28	2,078	.41	.06	26	.30	.52
Team-based	52	2,954	.45	.06	20	.34	.56
Humor	2	74	.83	.17	100	.50	1.16
Case-based	114	7,031	.50	.05	11	.40	.60
Book review	3	99	.29	.23	30	-.17	.74
Essays	44	2,579	.39	.07	14	.25	.54
Workbooks	3	319	.68	.15	41	.39	.98
Worksheets	12	971	.55	.09	19	.36	.73
Discussions	86	5,387	.43	.06	13	.32	.53
Discussions (Large group)	64	4,328	.44	.06	13	.32	.55
Discussions (Small group)	62	4,252	.40	.05	15	.30	.50
Web-based discussion	18	671	.40	.10	26	.20	.59
Role plays	33	2,114	.44	.05	32	.33	.54
Debates	9	581	.63	.28	7	.08	1.17
Computer-based	41	2,790	.52	.07	17	.38	.67
Self-reflection	47	3,224	.43	.07	13	.29	.56
Review	21	1,324	.59	.10	15	.39	.79
Note-taking	7	136	.85	.14	84	.58	1.13
Games	2	495	.18	.04	100	.10	.26
Current events	6	595	.60	.42	3	-.22	1.42
Mentoring	6	147	.19	.17	100	-.14	.52
Service learning	5	207	.25	.14	61	-.03	.53
Readings	65	5,044	.45	.08	8	.30	.60
Presentations	19	962	.37	.12	25	.15	.60

Note. *k* = number of effect sizes; *N* = total sample size; *Md* = Inverse variance-weighted (Hedges & Olkin, 1985), uncorrected mean effect size (*d*); *SD* = Standard deviation of mean effect size; Variance (%) due to sampling error = 100 - I^2; CI = Confidence interval.

Table 14.

Characteristics of Case-Based Instruction

			Weighted		Var. (%) due to sampling error	95% CI	
	k	***N***	***Md***	***SD***		**Lower**	**Upper**
Emphasis on Case Instruction							
Below Average	47	2,836	.47	.09	9	.29	.66
Average	43	2,656	.47	.05	28	.37	.57
Above Average	20	1,429	.59	.17	5	.26	.91
Case Length							
1 Paragraph	14	677	.43	.11	33	.22	.65
2 – 6 Paragraphs	8	588	.55	.14	23	.28	.81
6+ Paragraphs	5	654	.73	.26	3	.22	1.23
Case Complexity							
Low	8	240	.52	.14	36	.24	.80
Moderate	10	574	.66	.26	7	.15	1.18
High	12	1,142	.33	.08	26	.17	.49
Case Emotional Content							
Low	4	117	1.03	.38	12	.29	1.78
Moderate	7	626	.80	.25	7	.31	1.28
High	8	686	.20	.04	100	.13	.27
Case Realism							
Low	11	593	.46	.22	9	.03	.88
Moderate	22	1,684	.40	.07	19	.25	.54
High	17	1,306	.29	.07	38	.16	.42

Note. *k* = number of effect sizes; *N* = total sample size; *Md* = Inverse variance-weighted (Hedges & Olkin, 1985), uncorrected mean effect size (*d*); *SD* = Standard deviation of mean effect size; Variance (%) due to sampling error = 100 - I^2; CI = Confidence interval.

Table 15.

Characteristics of Practice Opportunities

		Weighted			Var. (%) due to sampling error	95% CI	
	k	***N***	***Md***	***SD***		**Lower**	**Upper**
Frequency of Practice							
Below Average	34	2,233	.32	.06	21	.20	.45
Average	15	551	.47	.10	42	.28	.67
Above Average	10	554	.76	.16	18	.46	1.06
Length of Practice Sessions							
Up to 30 Minutes	19	1,114	.61	.09	20	.43	.78
30+ Minutes	20	1,082	.53	.12	13	.30	.77
Field-specificity of Practice							
Field-specific	33	1,720	.40	.05	37	.30	.50
Balanced	20	1,522	.45	.13	7	.20	.70
General	8	430	.63	.14	47	.35	.91
Realism of Practice							
Low	19	727	.67	.14	14	.39	.95
Moderate	22	1,277	.32	.07	33	.19	.45
High	11	857	.33	.07	30	.20	.47
Individual vs. Group Practice							
Individual-based	24	1,319	.56	.15	10	.27	.85
Mixed	27	1,354	.37	.06	35	.25	.49
Group-based	19	1,287	.28	.07	30	.15	.41
Size of Practice Group							
2 – 5 Trainees	17	970	.35	.10	30	.16	.53
5+ Trainees	9	652	.27	.07	32	.12	.41

Note. *k* = number of effect sizes; *N* = total sample size; *Md* = Inverse variance-weighted (Hedges & Olkin, 1985), uncorrected mean effect size (*d*); *SD* = Standard deviation of mean effect size; Variance (%) due to sampling error = 100 - I^2; CI = Confidence interval.