Effective Strategies for Research Integrity Training—a Meta-analysis

Katsarov, Johannes; Andorno, Roberto; Krom, André; van den Hoven, Mariëtte

Published in:
Educational Psychology Review

DOI:
10.1007/s10648-021-09630-9

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Educational Psychology Review (2022) 34:935–955
https://doi.org/10.1007/s10648-021-09630-9

META-ANALYSIS

Effective Strategies for Research Integrity Training—a Meta-analysis

Johannes Katsarov[1,2] · Roberto Andorno[1] · André Krom[3] · Mariëtte van den Hoven[3]

Accepted: 2 July 2021 / Published online: 27 August 2021

Abstract

This article reviews educational efforts to promote a responsible conduct of research (RCR) that were reported in scientific publications between 1990 and early 2020. Unlike previous reviews that were exploratory in nature, this review aimed to test eleven hypotheses on effective training strategies. The achievement of different learning outcomes was analyzed independently using moderator analysis and meta-regression, whereby 75 effect sizes from 30 studies were considered. The analysis shows that the achievement of different learning outcomes ought to be investigated separately. The attainment of knowledge strongly benefited from individualized learning, as well as from the discussion and practical application of ethical standards. Contrarily, not covering ethical standards tended to be a feature of successful courses, when looking at other learning outcomes. Overall, experiential learning approaches where learners were emotionally involved in thinking about how to deal with problems were most effective. Primarily intellectual deliberation about ethical problems, often considered the "gold standard" of ethics education, was significantly less effective. Several findings from previous reviews, e.g., the preferability of mono-disciplinary groups, could not be replicated with multivariate analysis. Several avenues for future research efforts are suggested to advance knowledge on the effectiveness of research integrity training.

Keywords Meta-analysis · Meta-regression · Research integrity · RCR · Ethics education

Highlights
- Meta-regression was used to test the robustness of conclusions from prior reviews
- Different goals of research integrity training benefit from different approaches
- Experiential training approaches are more effective than classical case discussions
- The familiarization with ethical codes only supported the development of knowledge
- Higher-order learning outcomes did not benefit from a coverage of ethical codes

✉ Johannes Katsarov
johannes_katsarov@hotmail.de

Extended author information available on the last page of the article

Introduction

For several decades, education on responsible conduct of research (RCR) has been offered to promote ethically accountable research by means of raising people's awareness of relevant issues and developing their abilities to address them. In the USA, serious cases of misconduct led the Office of Research Integrity, the National Academy of Sciences, and the National Institutes of Health to require obligatory training for researchers upon receiving a grant in the early 1990s. This has boosted training and education on RCR in the curriculum in the USA, leading to a bulk of literature on the aims, methods, and effects of such trainings (Steneck, 2007). Important topics include the involvement of humans and animals in research, issues related to authorship, intellectual property, mentor-mentee relationships, and an accountable use of research data (Macrina, 2014). In Europe, where researchers face the same challenges, recent EU-funded projects in the Horizon2020 framework program stimulate RCR education via innovative educational methods and tools.[1] Broadly conceived, RCR trainings try to promote ethical (i.e., reliable, honest, respectful, and accountable; ALLEA, 2017) behavior related to research, also known as scientific/research integrity. In practice, a broad variety of courses focus on increasing knowledge, attitudes, and competences of students and researchers on integrity issues, e.g., the ability to recognize harmful research practices, the knowledge of codes of conduct, or attitudes that promote a culture of accountability.

Thus far, two meta-analyses have been published inquiring factors that moderate the effectiveness of RCR education (Antes et al., 2009; Watts et al., 2017). These meta-analyses have been accompanied by further systematic reviews, which explored the effectiveness of RCR courses qualitatively (e.g., Marušic et al., 2016; Todd et al., 2017a) or with advanced statistical methods (e.g., Mulhearn et al., 2017). One way in which this review is innovative lies in its goal to seek further validation of claims about the effectiveness of RCR education through the use of meta-regression. Unlike the previous reviews that were exploratory in nature, we test eleven hypotheses on successful training strategies to promote research integrity. Working with multivariate meta-regression allows us to test the impact of different factors while controlling for other possible influences simultaneously. To some degree, our hypotheses are based on the findings from previous reviews, and we only want to test their robustness. However, some of our hypotheses are also new insofar as we posit that the effectiveness of RCR training is relative to the attainment of specific *types* of learning outcomes.

Thus far, systematic reviews on the effectiveness of RCR education (e.g., Todd et al., 2017a; Watts et al., 2017) have not analyzed whether teaching methods that promote *one* kind of learning outcome (e.g., the memorization of ethical codes of conduct) also promote *all other* types of learning (e.g., the ability to reason about ethical problems). In their analyses, all effect sizes are pooled, independent of *what* has been learned. This approach does not adequately reflect the empirical fact that cognitive, skill-based, and affective learning outcomes are structurally different and benefit from different types of learning (Kraiger et al., 1993). It is also surprising insofar as studies have demonstrated that different components of moral/ethical functioning tend to be independent from each other: Correlations between measures of different competences in the moral/ethical domain tend to be low (e.g., You and Bebeau, 2013).

[1] For example, PRINTEGER (https://printeger.eu/), INTEGRITY (www.h2020integrity.eu), Path2Integrity (https://www.path2integrity.eu/), and Embassy of Good Science (https://embassy.science/wiki/Main_Page).

Our approach is innovative insofar as we stipulate that approaches that support the achievement of one type of learning outcome of RCR may not be as helpful in achieving other learning outcomes and vice versa. Drawing on a recent taxonomy of learning outcomes in the moral/ethical domain (Maesschalck and De Schrijver, 2015), a synthesis of the abilities underlying moral agency (Tanner and Christen, 2014), and a taxonomy of learning objectives for RCR (Antes and DuBois, 2014), we distinguish five types of learning outcomes:

- *Knowledge*: Ability to understand, remember, and recall concepts, facts, and procedures related to RCR.
- *Attitude*: Endorsement and expression of beliefs, motivations, and attitudes that reflect research integrity and a willingness to exercise research in a responsible manner, including the endorsement of specific norms and standards.
- *Sensitivity*: Ability to notice, recognize, and identify ethical problems related to RCR.
- *Judgment*: Capacity to engage in professional ethical decision-making, drawing on experience and meta-cognitive strategies like the anticipation of consequences and the testing of assumptions.
- *Behavior*: Actual or planned ethical behaviors of individuals, including measures of moral courage and self-efficacy, because these mirror people's readiness to behave in line with their judgments.[2]

"Knowledge" is clearly the learning outcome for which the largest "average standard effect sizes" (*Md*) were found in the meta-analysis by Watts and colleagues in 2017 ($Md = 0.78$ for $k = 27$ effect sizes). Relatively small average effects were found for "moral judgment" ($Md = 0.25$, $k = 13$) and "moral reasoning" ($Md = 0.39$, $k = 47$)—both of which we would summarize under the category "judgment".[3] Training people's ability to understand, remember, and recall ethical concepts ("knowledge") can be considered relatively easy in comparison to, say, promoting people's general maturity of moral reasoning as measured with a test like the *Defining Issues Test* (Rest et al., 1999). This is exemplified by a knowledge test question used by Melcer et al. (2020), which asks respondents "Which of the following is NOT considered a contribution to a paper?", followed by multiple choice items. In contrast, the *Defining Issues Test* assesses a person's ability and motivation to coherently prioritize principled ("post-conventional") reasons over egoistic and "conventional" (e.g., law-abiding) reasons in deliberating about several moral dilemmas. People's ability to exercise moral judgment can be considered a higher-order cognitive ability (in contrast to understanding, remembering, and recalling theoretical concepts), and it is well known that the development of this ability takes considerable time (Rest et al., 1999). This may help to explain the finding of Watts et al. (2017) that educational interventions of less than 8 h duration ($Md = 0.61$, $k = 47$) tended to be significantly *more* effective than interventions of 16 h and more ($Md = 0.39$, $k = 65$): Since

[2] Originally, we also intended to code two additional types of learning outcomes: an increase in "basic skills," and an increase of "bias awareness." We found no instance of the prior and only one instance of the latter, which is why we concentrate on the reduced set of five learning outcomes.

[3] What all measures of "judgment" have in common—whether they were quantitative or qualitative, focused on research or more general—is that respondents' ability to justify moral evaluations with ethically sound arguments is assessed. Building on the work of Jean Piaget and Lawrence Kohlberg, people who refer to ethical principles or standards (e.g., respecting people's autonomy, justice, responsibility, prevention of harm) when justifying decisions/preferences receive better scores than people whose reasoning tends to be based on egoistic motives or conventions (e.g., law).

effect sizes for higher- and lower-order cognitive abilities were pooled, a large number of relatively short interventions that assessed the development of "knowledge" could have led to the statistical observation that (relatively long) interventions that tested the development of "judgment" were less effective. In our study, we perform distinct and comparative analyses for different dependent variables, i.e., types of learning outcomes, to rule out problems like this.

Eleven Hypotheses on the Effectiveness of RCR Courses

Building on the previous reflection and the findings from earlier reviews, our meta-analysis aims to test eleven hypotheses on successful RCR training strategies.[4] The first two hypotheses are grounded on the critique of previous reviews and do not require further explanation:

H1: Courses' effectiveness increases with their duration when single types of learning outcomes are analyzed.

H2: Different teaching approaches will prove significantly more helpful in promoting different types of learning outcomes, as articulated in hypotheses H3–H5.

As prior reviews (Todd et al., 2017a; Watts et al., 2017) suggest, an active engagement of learners was a common characteristic of the most successful courses. Courses tend to be more effective if they engage people in individual learning. Moreover, effective courses did not only rely on group interaction, which can deteriorate individual engagement. Passive learning and a total reliance on group activities were common characteristics of weak courses. The most effective courses combined individual and group-based activities. We formulate the following hypothesis to replicate these findings:

H3: Courses that combine individual and group-based learning activities are more effective in achieving attitudinal learning and higher-order learning outcomes than courses that only draw on either individual or group-based learning activities.

Courses that strongly focus on ethical decision-making and which challenge learners to develop relevant competences have been found to be among the most effective (Torrence et al., 2017; Mulhearn et al., 2017). Ineffective RCR courses frequently place a priority on abstract moral principles instead of providing clear guidance to learners, as to how to deal with ethical challenges and dilemmas (Torrence et al., 2017). Process-based contents yielded good results overall, with the strongest effects found for courses that dealt with emotional analysis, forecasting, and the analysis of consequences (Watts et al., 2017). Coverage of possible reasoning errors has also been found to be an important content (Antes et al., 2009), as well as considering one's own motives, values, and emotions (Torrence et al., 2017). Relatedly, different reviews found that courses that applied ethical decision-making competences to real-world cases tended to be more effective than those who did not (Antes et al., 2009; Todd et al., 2017a; Watts et al., 2017). This suggests that good RCR courses engage learners in imagining how they would deal with cases relevant to their prospective research practice. From a theoretical perspective, we expect that experiential learning is especially stimulating, because it challenges learners to consider how they would deal with actual, practical challenges. Learning that looks at the motivational and situational factors that drive unethical behavior should be even more effective than courses that discuss ethical cases in view of what is right or

[4] Originally, we registered twelve hypotheses. However, when coding the studies, we recognized that two of our hypotheses were identical, so we merged them into what is now H4. Moreover, while we maintained all of the registered hypotheses, we now present them in a different order, and partially with a slightly modified language.

good, but which refrain from looking at psychological and social influences. Based on these considerations, we formulate a novel hypothesis concerning the general strategy of RCR education:

H4: Courses that challenge learners to imagine how they would personally deal with ethically problematic situations in their area of research (and thereby emphasize the importance of competences for ethical decision-making) are more effective at promoting attitudinal learning, ethical sensitivity, moral judgment, and behavioral learning than courses that do not include relevant activities or which deliberate about cases in an impersonal fashion. However, no advantage of experiential learning is expected for the promotion of knowledge.

Prior reviews indicate that highly effective research-integrity courses introduce and explain rules, standards, and guidelines for RCR and stress their importance for practice. This was a common feature found for highly effective courses (Torrence et al., 2017). Relatedly, courses that cover a wider range of RCR topics tend to achieve larger effects (Watts et al., 2017). The need for sufficient consideration of diverse RCR topics is probably also the reason why embedded efforts to foster research integrity tended to yield lower effects than discrete, stand-alone courses (Antes et al., 2009; Watts et al., 2017). From a theoretical perspective, we only expect that the coverage, discussion, and—ideally—application of rules, guidelines, and standards for RCR will promote (related) knowledge, attitudes, sensitivity, and behavior, e.g., the ability of learners to recognize ethical problems relevant in the given domain. On the other hand, judgment may not improve: While learners' reasoning competences may expand through the discussion (and application) of standards, negative effects or stagnation may follow from the availability of clear rules that people can defer to, and which "allow" them to refrain from autonomous judgment. This leads us to the following hypothesis:

H5: Courses that introduce learners to rules, standards, or guidelines for a responsible conduct of research are more effective at promoting knowledge, sensitivity, and attitudinal learning than courses that do not include these kinds of contents. Yet, judgment-related competence is not promoted by appealing to this strategy.

Beyond these hypotheses that are sensitive to the type of learning outcome, we expect the remainder of our hypotheses to be generic, i.e., independent of the type of learning outcome. The main purpose here is to replicate findings from previous studies with more sophisticated methods.

First, various reviews suggest that the effectiveness of RCR courses is undermined, if they are offered to highly diverse groups of learners, e.g., across different faculties like engineering and social science (Watts et al., 2017). Effective courses are either aimed at a relatively wide group of learners, focusing on general contents, or at a specific group of learners, looking at specific issues (Mulhearn et al., 2017; Todd et al., 2017a; Watts et al., 2017). From a learner-centered perspective, this makes sense, because cases that are relevant for one group of learners may be irrelevant for other groups, and the common ground found across domains may be so abstract that learners fail to see its practical relevance. Based on these considerations, our hypothesis reads:

H6: Courses that are offered to mono-disciplinary groups of learners focusing on one domain are more effective than courses offered to learners from diverse domains.

Moreover, two reviews found that RCR courses with a relatively high degree of practice, i.e., the repeated application of learned abilities and knowledge, typically yielded larger learning effects, especially when the exercises were performed individually (Torrence et al., 2017; Watts et al., 2017). Learning theories support the need for practice, as people take time to organize their knowledge structures, automatize analytical and judgment-related processes,

and adopt new attitudes and behaviors to a mature degree (Kraiger et al., 1993). This leads us to the following hypothesis:

H7: Courses that challenge learners to practice their abilities for a responsible conduct of research repeatedly are more effective than courses with little or no repetition.

Three reviews suggest that effective RCR courses benefit from *blended learning* (Todd et al., 2017a; Todd et al., 2017b; Watts et al., 2017). In other words, courses that combined online learning activities with interactive units where learners were physically present tended to yield higher effect sizes (on average) than courses that either relied on pure online activities or face-to-face activities only. Based on these considerations, our hypothesis reads:

H8: Courses that make use of blended learning are more effective than pure online or pure face-to-face courses when controlling for other expected influences.

One prior review also indicates that RCR courses were more effective if the teachers had relatively good expertise (Watts et al., 2017). In the meta-analysis by Antes et al. (2009), courses tended to be more effective when the authors of the respective articles served as instructors, which could also be indicative of a high level of expertise. A recent survey involving 99 RCR teachers across Europe suggests that teachers with special training, e.g., as educators, perceived themselves as more effective than teachers who lacked this background (Andorno et al., 2019).

H9: Courses offered by experienced and/or trained teachers are more effective than courses offered by novice teachers.

Another insight of the previously mentioned survey (Andorno et al., 2019) was that many lecturers reported that their courses were undermined by a weak appreciation of research ethics at the respective institutions. Similarly, a systematic review on ethics training for physicians noted that a lack of institutional or departmental support posed a significant barrier (Martakis et al., 2016). One aspect where RCR courses differ relates to their institutional recognition, particularly whether they form part of the (mandatory) core curriculum, or whether they are offered by the side, voluntarily, without credit, etc. Previous meta-analyses have not found support for the assumption that courses are more successful when they are mandatory and/or advocated by an organization (Antes et al., 2009). However, we want to test whether these findings are upheld when multivariate regression analysis is used, which allows for the control of multiple variables.

H10: Courses that benefit from a strong institutional recognition (e.g., systematic integration in curricula, strong commitment to research integrity) are more effective than courses that cannot build on institutional endorsement.

Finally, our study aims at investigating the effectiveness of RCR courses for three groups of learners (high school students; university students and professionals below doctoral level; researchers and doctoral candidates). Despite findings from prior meta-analyses, which suggest that some groups learn more than others (Antes et al., 2009; Watts et al., 2017), we do not expect to find significant differences between these groups when applying multivariate analysis, because degrees of learning generally tend to be relative to desired outcomes:

H11: Effects found for different groups of learners do not differ systematically when controlling for other expected influences (e.g., the use of blended learning).

Method

Following the PRISMA standard for systematic reviews (Moher et al., 2009), we took diverse measures to ensure a comprehensive selection of studies and to safeguard the robustness of our analysis.

Search Strategies

To identify relevant studies, we drew on an existing database of 531 articles that had been cited in one of 21 reviews related to ethics education (Appendix A). In addition to this database, we searched Web of Science, ERIC, Google Scholar, and all ProQuest databases for relevant articles using combinations of the following terms in their titles: (research ethics OR responsible conduct of research OR research integrity OR scientific integrity) AND (teaching OR learning OR training OR course OR trial). Google Scholar and ProQuest were selected to identify unpublished articles and dissertations. Other search methods, through which further articles were identified for possible inclusion, included checking the references of included articles and using SCOPUS to identify articles that had cited relevant studies.

Inclusion and Exclusion Criteria

Using the inclusion and exclusion criteria listed in Table 1, abstracts were collected for 1.548 records and screened by the first author. When an immediate decision about inclusion was not possible based on the abstracts, or if no abstract were available, full texts were consulted. During the process of screening, two fundamental decisions were made by the research team to define the scope of the review more rigorously. First, unlike two prior meta-analyses on training for a responsible conduct of research (Antes et al., 2009; Watts et al., 2017), which took a broader view in looking at "ethics instruction in the sciences," we chose to apply a narrower definition of RCR in our review. By the wider definition employed in previous reviews, it is not clear why courses related to business ethics were excluded, for instance, while professional ethics education for nurses (that did not deal with research issues) was included. An explicit focus on courses that deal with research ethics and integrity permits a clearer picture of the effectiveness of actual RCR training.

Second, contrary to Marušic et al. (2016) who reviewed interventions to prevent misconduct and promote integrity in research and publication, we also decided to exclude studies that were purely focused on preventing plagiarism. Although plagiarism is an issue related to RCR, relevant studies tend to have a pure focus on transmitting knowledge related to correct citation and imbuing learners with attitudes against plagiarism. We find this focus too narrow to

Table 1 Inclusion and exclusion criteria and numbers of studies excluded

Criterion	Inclusion	Exclusion
Language	English	Non-English reported studies
Time period	January 1990 to June 2020	Studies outside the time period
Dependent variable	Studies investigating learning outcomes of RCR courses through relevant tests (e.g., of RCR-related knowledge)	Studies only investigating perceived learning outcomes, student satisfaction, or dependent variables of no direct relevance to RCR training
Cognitive consequences approach	Studies that investigated learning through an intervention in contrast to prior knowledge or an untreated control group	Media and method comparison studies that did not assess the effectiveness of an intervention with regards to prior knowledge of an untreated control group
Availability	The full study must be available to consult via a journal or the internet	Studies of which the full text was not available to consult
Statistical information	Studies reporting sufficient information to calculate an effect size	Studies reporting insufficient information to calculate an effect size

characterize an RCR course and have therefore excluded relevant studies. This also prevents a strong bias in our findings due to the large volume of relevant studies.

Data Extraction and Analysis

Once a preliminary sample of studies had been selected for full-text analysis, the authors pre-registered the approach for the data extraction and analysis (Registration DOI: 10.17605/OSF.IO/W9J3U). This registration included the hypotheses stated in "Eleven Hypotheses on the Effectiveness of RCR Courses," the preliminary operationalization of the moderator and control variables, and analytical procedures to test the different hypotheses.

Effect Sizes

Effect sizes were calculated separately for five types of learning outcomes as specified in "Eleven Hypotheses on the Effectiveness of RCR Courses." Effect sizes for each outcome were calculated with *Comprehensive Meta-Analysis* (Version 3.3.070). To calculate the "standardized mean difference" (*Cohen's d*), we used one of five formulas (Appendix B), depending on the reported statistics and whether we were dealing with a pre-/post-test comparison for a single group (paired comparison), a comparison of an intervention group's post-test results with the post-test results of an untreated control group (control-group comparison), or a combination of both (paired + control). Based on *Cohen's d*, we then calculated the "standardized mean difference corrected for bias" (*Hedges' g*) by multiplying d with a correction factor J. The smaller the sample size, the more the correction factor (J) reduces the final effect size (g). Due to this correction of potential bias from small sample studies, g is considered a more robust effect measure than d (Lakens, 2013).

In cases where both a t- and a p-statistic were available for paired effects, preference was given to the t-statistic because p-values were often less accurate (e.g., when they were only expressed as $p < 0.001$, which would cap the calculated effect size below its real value). In cases where we had enough data to calculate effect sizes autonomously, we ignored effect sizes calculated by the authors themselves. If one-tailed t-tests were not mentioned explicitly, we assumed that two-tailed t-tests had been performed. Some t-statistics were computed from the F-scores of ANOVAs (analysis of variance) using the formula $F = t^2$. When several effect sizes of one outcome type existed, e.g., four measures of judgment, we calculated a mean effect size per intervention to reduce the risk of multiplicity.

Moderator and Control Variables

Using a pre-configured *MS Excel* table, information was extracted on each educational intervention regarding (1) type of education (high school + general citizens; higher education students + graduates below PhD; researchers including PhD candidates), (2) target group(s), (3) mono-disciplinary or multi-disciplinary course, (4) type of instruction (pure individual, pure group, or individual and group learning), (5) course emphasis (*theoretical* = no engagement of learners with practical ethical problems; *deliberative* = active engagement of learners with concrete ethical problems but without addressing psychological and emotional dimensions of ethical problem-solving; *practical* = engagement of learners with concrete ethical problems addressing both cognitive and affective dimensions), (6) introduction and application of ethical guidelines (no; superficial; applied), (7) quality and quantity of engagement with

cases, (8) use of e-learning (no; pure e-learning; blended learning), (9) course duration, (10) competence of teachers, (11) institutional recognition of the course, (12) whether the study had undergone peer review, and (13) the gender mix of the learners (Appendix C provides an overview of the codings per study). Additionally, we coded whether a course had used one of 16 educational methods identified across all studies and counted the number of combined methods.

The coding criteria for these moderator variables were specified at the time of pre-registration, i.e., before any of the studies had been coded. In the first phase of coding, five randomly selected studies that none of the authors had read before were coded by several authors independently. Where codings diverged, criteria and interpretations were discussed until consensus was found. Based on this consensus, all studies were coded by the first author. For a final quality check, six randomly selected papers were coded by other members of the team. Inter-rater reliability was estimated at .983 based on only one deviation.

After coding, the two variables *course emphasis* and *case engagement* appeared to be redundant, which is why we merged two of our original hypotheses into one (now H4). For the variable *institutional recognition*, we merged the number of categories from five to three because two of the categories were coded very rarely. For the same reason, we merged courses with a duration of "<2 h" and "2–5 h" in one category, and we merged the volumes of 1–2 and 3–4 treated cases into one category.

Risk of Bias Assessment

To assess the risk of bias in studies, we adapted the 10-item *Medical Education Research Study Quality Instrument* (MERSQI) by Reed et al. (2007), which has demonstrated a high interrater and intra-rater reliability and validity in terms of citation rate and impact factor. The MERSQI assesses possible bias in individual studies based on a separate quality assessment for each outcome. Thus, if a study assessed two types of learning, e.g., of knowledge and judgment, we calculated separate quality scores for each outcome measure. Our adapted quality scale assesses (1) the study design, (2) number of included institutions, (3) response rate, (4) quality of assessment, (5) reliability, content validity, and convergent/divergent validity of the measure, (6) sophistication and adequacy of data analysis, and (7) risk of social desirability bias. In line with the PRISMA standard, we analyzed the risk of bias separately for each dimension using moderator and regression analysis.

Moderator Analysis and Meta-regression

To test our hypotheses, we primarily performed meta-regression analyses, as planned upon registration. Expecting a high degree of heterogeneity in effect sizes, we performed all analyses with random-effect models. We did not expect that effect sizes would be normally distributed, so we applied the Method of Moments (a.k.a. DerSimonian and Laird method). First, we conducted moderator analyses for all covariates to gain a first overview of possible reasons for heterogeneity of effect sizes. Then, we performed meta-regressions to identify the best explanations for different effect sizes (per outcome category) using multiple covariates simultaneously. The goal was to identify an optimal model using the available covariates. For our purposes, an optimal model bears the following characteristics:

- It maximizes the chance that its covariates explain any of the variance, i.e., the *F*-value.

- It minimizes unexplained variance between groups, i.e., Tau^2 and I^2.
- It uses as few covariates as possible to perform these functions (parsimony criterion).

Using the three principles explained above, we "distilled" the best models from hundreds of tested models, whereby we took all moderator and risk of bias variables into consideration. Through this approach, we also intended to address the *multiple comparisons problem*, i.e., the risk that hypotheses are accepted or rejected naively when several variables are tested simultaneously: Statistical significance ($p < .05$) may arise due to sampling error in such cases. Meta-regression reduces multiplicity and significance testing in meta-analyses and therefore provides more robust results (Pigott and Polanin, 2015). Moreover, we consider strategies to optimize model-fit (e.g., the F, Tau^2, and I^2 values) to be more robust than simply looking for significant p-statistics. Finally, to reduce the risk of overestimating the significance of predictors, we used the Knapp-Hartung adjustment to obtain more reliable estimates (cf. Higgins et al., 2002).

Results

Study Selection

Only 30 of the 84 studies selected for full-text analysis were considered eligible for inclusion. Thirteen of the 66 studies included in the prior meta-analysis by Watts et al. (2017) fulfilled our inclusion criteria, with the majority of these studies ($n = 36$) being excluded because they did not refer to RCR education. Seventeen studies included in our meta-analysis were not considered in the meta-analysis by Watts et al. (2017), although only four of them were published after 2015. Overall, our sample of studies is more restrictive in its focus on RCR courses while also including ten studies that were not included in previous reviews (Fig. 1).

Description of Included Studies

As Table 2 shows, the 30 included studies yielded 75 effect sizes for the five outcomes of interest. A substantial heterogeneity of effect sizes was only confirmed for attitudinal, judgment-related, and knowledge-related learning with $I^2 > 50\%$ in these cases (Deeks et al., 2019). This suggests that a moderator analysis is justified for these three outcomes, while a moderator analysis for behavior and sensitivity is questionable with $I^2 < 25\%$. When pooled, a satisfactory heterogeneity of effect sizes was found to warrant moderator analyses for combined effect sizes of attitudinal, behavioral, and sensitivity-related learning. For practical purposes, we label this pool of related learning outcomes "orientational learning outcomes."[5] Obviously, pooling these three distinct learning outcomes bears the aforementioned risk of overgeneralizing the impact of factors that only influence one outcome. However, when studies measured several of these three outcome

[5] What all the measures of "orientation" have in common is a strong attitudinal component, which predisposes people to evaluate ethical problems differently. For example, *careless response behavior* (a behavioral measure; DuBois et al., 2018) indicates that people do not value science. Noticing issues of ethical importance (a measure of moral sensitivity; Clarkeburn et al., 2002) indicates that people find that relevant aspects and problems merit attention.

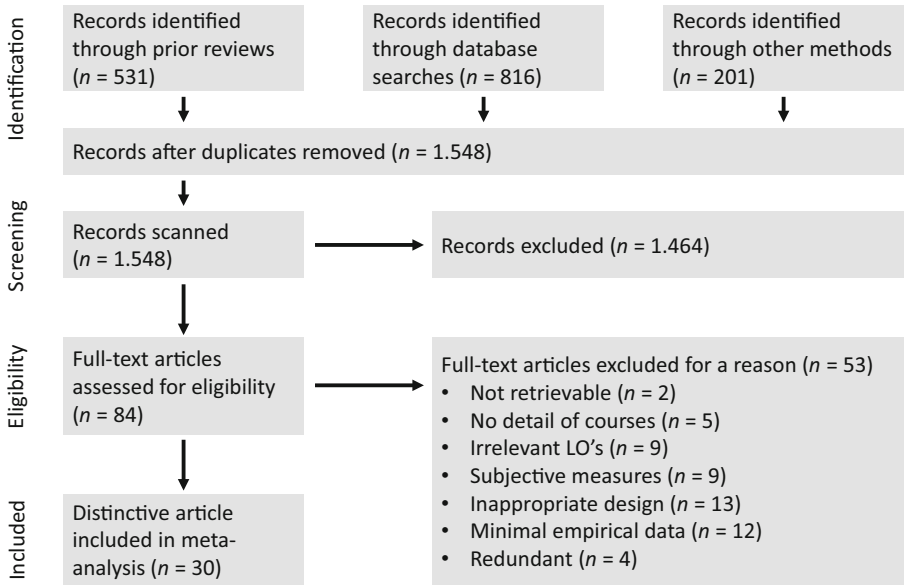

Figure 1 PRISMA flow diagram of article selection.

types simultaneously, effect sizes tended to be similar (large deviations were found between knowledge, judgment, and orientation outcomes). Therefore, we concluded that the risk of overgeneralization was relatively small. Nonetheless, the orientation dimension does not merit a differentiated analysis in our view because its three components should be analyzed separately. However, due to the common denominator (strong attitudinal component), we think that this dimension adds value in Table 3 and for the discussion, because it shows how effects may vary strongly for knowledge, judgment, and orientation outcomes.

Sufficient heterogeneity is only one criterion to justify the use of meta-analytical methods like moderator analysis and meta-regression. Another important criterion is sufficient statistical power to rule out the possibility that insignificant results are due to a lack of studies (Hedges and Pigott, 2004). Following Valentine et al. (2010), we estimated the necessary number of effect sizes per learning outcome with an Excel-based calculator provided at https://osf.io/juzfg/. The relevant criteria to assess the statistical power are (a) the minimum effect size that is estimated to be relevant, (b) the average number of people per group, and (c) the number of effect sizes, as well as the question, how high heterogeneity is (in the case of random-effects models). Criterion (a) should be set a priori, in view of what kind of an effect size one would generally find relevant. In view of the meta-regression analyses that we perform, we find it prudent to expect effect size differences (regression coefficients) of a least 0.3 (at least for binary independent variables). Anything lower than this number would be ambiguous, considering the relevant low number of included studies and typical standard errors. Based on the high heterogeneity and the average sample size of 48.31, we need 15 effect sizes to have a minimum power of .8 (Cohen, 1977). Therefore, we have sufficient statistical power to analyze the results separately for knowledge and judgment outcomes.

Table 2 Main effects (random effects analysis)

Outcome	*Mg*	*SE*	*K*	*N*	*95% CI*	I^2
Knowledge	0.64	0.07	28	1,086	[0.50, 0.79]	75.54***
Judgment	0.41	0.08	23	803	[0.25, 0.56]	85.06***
Orientation	0.52	0.52	25	712	[0.39, 0.65]	73.03***
Attitude	0.46	0.16	6	184	[0.15, 0.77]	65.50*
Behavior	0.69	0.17	5	121	[0.36, 1.03]	0.00
Sensitivity	0.42	0.11	13	407	[0.20, 0.64]	3.09
Overall	0.52	0.05	75	2,601	[0.43, 0.61]	79.15***

Note. Mg weighted mean effect size; *SE* standard error; *k* number of effect sizes; *CI* confidence interval; Significance test levels: *$p < .05$, **$p < .01$, ***$p < .001$

Risk of Bias

Publication Bias

Since we are dealing with different learning outcomes, with different mean effect sizes (*Mg*) and different sample sizes (*k*), we estimated the risk of publication bias separately for different types of learning outcomes. Using Duval and Tweedie's *trim and fill* procedure to estimate the number of missing studies that would lead to a symmetric funnel plot, we found that no studies were missing for knowledge, attitudes, behavior, and sensitivity. For judgment, the analysis suggests three missing studies, leading to an imputed point estimate, i.e., a corrected mean effect size of $Mg = 0.33$ (*95% CI* = [0.16, 0.50]). On the other hand, we found that the mean effect sizes of peer reviewed ($Mg = 0.55$, $SE = 0.05$, $k = 64$) and unpublished studies ($Mg = 0.50$, $SE = 0.13$, $k = 13$) did not differ significantly ($Q = 0.10$, $p = 0.75$).

Study-Internal Risk of Bias

We performed moderator analyses with ten factors that could result in biased results from studies (Appendix D). No significant heterogeneity of effect sizes was found for study design, number of institutions, response rate, assessment type, disclosure of reliability, content validity, and correlations, quality of data analysis, or social desirability bias. Some minor differences were observed when performing this analysis distinctly for different types of learning outcomes (Appendix E), which is why we also performed a meta-regression to estimate risks of bias. This analysis suggests the following risks (Appendix F):

- Studies that lacked a control group may overestimate effects (especially knowledge).
- Studies with a low response rate may overestimate effects (knowledge only).
- Studies who evaluated learning qualitatively may *underestimate* effects when coding is not performed in a blinded fashion (knowledge only).
- Mistakes in data analysis may have led to *underestimated* effects (knowledge).
- Studies that controlled for social desirability bias (e.g., through advances statistical methods like ANCOVAs) tended to report larger effects, which implies that uncontrolled social desirability bias may lead to *underestimated* effects.

Table 3 Moderator analysis.

	Knowledge			Judgment			Orientation		
Variable	*Mg*	*k*	*95% CI for g*	*Mg*	*k*	*95% CI for g*	*Mg*	*k*	*95% CI for g*
Target group									
School/citizens	–			–			0.68	3	[0.34, 1.01]
HE/professionals	0.68	13	[0.46, 0.90]	0.38	11	[0.15, 0.62]	0.40	12	[0.24, 0.56]
Researchers	0.55	10	[0.29, 0.81]	0.54	9	[0.26, 0.81]	0.54	9	[0.37, 0.72]
Mixed	0.74	5	[0.36, 1.12]	–			–		
Domain									
Single	0.65	21	[0.48, 0.83]	0.45	17	[0.24, 0.65]	0.44	19	[0.32, 0.56]
Mixed	0.61	7	[0.30, 0.93]	0.33	6	[−0.04, 0.69]	0.62	5	[0.42, 0.82]
Type of Instruction									
Pure individual	0.69	11	[0.45, 0.94]	–			0.36	9	[0.19, 0.53]
Pure group	0.64	7	[0.31, 0.97]	0.38	4	[−0.04, 0.81]	0.40	6	[0.23, 0.56]
Mixed	0.59	10	[0.32, 0.96]	0.39	18	[0.20, 0.58]	0.68	9	[0.53, 0.82]
Course emphasis									
Theoretical	0.53	8	[0.32, 0.74]	–			0.30	8	[0.13, 0.48]
Deliberative	0.48	12	[0.29, 0.67]	0.13	12	[−0.04, 0.30]	0.49	11	[0.35, 0.63]
0.81]Practical	1.01	8	[0.78, 1.24]	0.63	11	[0.46, 0.79]	0.69	5	[0.51, 0.84]
Ethical guidelines									
No	–			1.08	3	[0.52, 1.63]	0.47	9	[0.31, 0.63]
Superficial	0.43	10	[0.21, 0.65]	–			0.42	9	[0.23, 0.61]
Applied	0.81	15	[0.63, 0.99]	0.37	16	[0.18, 0.57]	0.64	5	[0.43, 0.85]
Quantity of cases									
None	0.46	5	[0.17, 0.76]	–			0.30	8	[0.13, 0.48]
1–	0.52	6	[0.24, 0.80]	0.85	3	[0.30, 1.41]	0.43	5	[0.26, 0.71]
5–7	0.61	9	[0.38, 0.84]	0.49	5	[0.10, 0.88]	0.68	7	[0.52, 0.84]
8+	0.91	8	[0.66, 1.17]	0.33	15	[0.12, 0.54]	0.56	4	[0.30, 0.82]
E-Learning									
On-site only	0.54	12	[0.30, 0.78]	0.45	19	[0.26, 0.64]	0.47	18	[0.33, 0.61]
Online/self-directed	0.69	13	[0.47, 0.90]	–			0.57	5	[0.26, 0.88]
Blended learning	0.87	3	[0.38, 1.37]	–			–		
Course duration (h)									
<5	0.45	6	[0.16, 0.75]	0.85	3	[0.35, 1.35]	0.34	11	[0.21, 0.48]
5–10	0.94	6	[0.65, 1.23]	0.47	7	[0.21, 0.73]	–		
10–20	0.60	6	[0.31, 0.89]	0.45	4	[0.13, 0.78]	0.36	3	[0.15, 0.57]
20+	0.60	10	[0.35, 0.84]	0.20	9	[−0.05, 0.45]	0.72	9	[0.59, 0.85]
Inst. recognition									
Voluntary, external	0.67	12	[0.43, 0.91]	0.02	3	[−0.50, 0.53]	–		
Voluntary, internal	0.71	9	[0.44, 0.99]	0.63	7	[0.33, 0.93]	0.39	15	[0.26, 0.53]
Mandatory	0.53	7	[0.22, 0.83]	0.37	13	[0.16, 0.57]	0.59	8	[0.43, 0.75]
Methods									
Lecture	0.63	16	[0.43, 0.90]	0.37	18	[0.17, 0.57]	0.49	15	[0.35, 0.64]
Reading materials	0.65	17	[0.46, 0.85]	0.28	17	[0.11, 0.46]	0.33	12	[0.20, 0.46]
Seminar	–			0.51	9	[0.28, 0.73]	–		
Writing papers/pres.	–			0.50	9	[0.23, 0.78]	0.58	4	[0.26, 0.89]
Case discussion	0.61	17	[0.41, 0.81]	0.33	14	[0.11, 0.55]	0.60	13	[0.47, 0.73]
Role play	1.01	8	[0.79, 1.24]	0.62	10	[0.44, 0.81]	0.57	3	[0.22, 0.92]
Reflection	–			0.61	10	[0.43, 0.78]	0.60	5	[0.40, 0.81]
Stakeholder exposure	0.93	5	[0.61, 1.25]	0.38	6	[0.04, 0.72]	0.32	4	[0.04, 0.60]
Small-group exercise	0.66	11	[0.41, 0.92]	0.33	10	[0.06, 0.60]	0.57	10	[0.42, 0.72]
Individual exercises	0.59	11	[0.34, 0.83]	0.47	14	[0.25, 0.67]	0.73	5	[0.56, 0.91]
Feedback (exercise)	0.69	9	[0.43, 0.95]	0.41	5	[0.03, 0.78]	0.64	6	[0.46, 0.82]
1–3 methods	0.52	12	[0.28, 0.76]	0.74	4	[0.37, 1.11]	0.43	10	[0.24, 0.61]
4–6 methods	0.70	14	[0.49, 0.92]	0.07	10	[−0.10, 0.23]	0.46	9	[0.28, 0.65]
7–9 methods	–			0.60	9	[0.44, 0.76]	0.60	5	[0.39, 0.81]

Note. If no mean effect size (*Mg*) is indicated, less than three studies (*k*) were found applicable

Moderator Analysis

We performed separate moderator analyses for knowledge, judgment, and orientation, looking at the mean effect sizes for the hypothesized influence factors and sixteen teaching methods (Table 3). The moderator analysis includes all methods except for those used less than four times overall, including problem-based learning, individual coaching, stress management practice, and watching movies or documentaries: due to the small numbers, results would have been highly unreliable.

Effects are only reported when they account for a minimum of three studies (k). Following Hunter and Schmidt (2004), effect sizes for fewer than 10 studies should be interpreted with caution. Interpretation should be made carefully in any case though since differences in mean effect sizes (Mg) may be due to other relationships. The meta-regressions in "Meta-regression Analyses" account for the possibility of multiple, interconnected influences, which is why we present the moderator analysis without further comment.

Meta-regression Analyses

Meta-regressions to explain variance in *knowledge*-related learning led to a model that explained all between-groups variance ($F = 10.76$, $p = 0.000$). Significant positive effects were found for courses that emphasized individual learning, experiential learning (practice orientation), and an application of ethical guidelines. Significant negative effects were found for courses that were mandatory, avoided concrete cases (theoretical), or did not teach ethical guidelines. No significant effects were found for specific teaching methods, the target group and mono- or multidisciplinarity of the course, the quantity of cases, course duration, the use of e-learning, gender mix, and risk of bias variables.

The best model to explain the effectiveness of moral *judgment* training explains 98% of between-groups variance ($F = 20.70$, $p = 0.000$). A significant positive effect was found for a practical, experiential engagement with concrete ethical problems related to RCR. A significant negative effect was found for the number of teaching methods that were employed. Students of higher education and non-research professionals tended to learn less than other groups, as well as mixed groups of participants, while high-school students and general citizens appear to have benefited more strongly from interventions. No significant effects were found for other variables when these variables were included in models, including all risk of bias variables, group variables, and distinct teaching methods.

We did not perform a meta-regression for the orientation outcomes (attitude, behavior, and sensitivity) due to a lack of heterogeneity. Sufficient heterogeneity would have been available to conduct a meta-regression for attitudinal learning. However, we deemed the number of six relevant studies too low.

To test our generic hypotheses, we pooled all learning outcomes. The best "all combined" model ($F = 8.62$, $p = 0.000$) explains 95% of the between-groups variance, with a heterogeneity of $I^2 = 12.10\%$ remaining unexplained. Significant positive effects were found for a practical engagement with concrete cases, for courses that did not teach ethical guidelines, and for courses of more than 5 h (in comparison to shorter courses). A significant negative effect was found for the number of employed teaching methods. Predominantly male groups tended to benefit more strongly from interventions. A significant bias was found for studies with relatively small or undisclosed response rates, which tended to find larger effects. A comparison of learning outcomes confirmed that

behavioral and knowledge-related learning tended to be greater than attitudinal, judgment-, and sensitivity-related learning (Table 4).

Discussion

After testing our eleven hypotheses against the studies that were reviewed, we come to the following conclusions. With regard to the hypotheses 1–5 that are related to distinct learning outcomes, we found the following:

H1: In contrast to our expectation, courses' effectiveness did not increase with their duration when single types of learning outcomes were analyzed (for knowledge and judgment). However, overall, short courses of less than 5 h appear to have been less effective than longer courses. Caution is required in interpreting this information though: A key finding is that even short interventions can yield large effects in terms of learning, e.g., the digital game *Academical* used by Melcer et al. (2020). At the same time, the employed measures may assess very narrow learning outcomes, so that large effect sizes may not be generalizable.

H2: In line with our expectations, different teaching approaches appeared to be more/less helpful in achieving distinct learning outcomes. This becomes most apparent in the different factors that supported the acquisition of knowledge in contrast to judgment-related competence (see discussion of H3-5).

H3: Supporting our hypothesis, a combination of individual and group-based learning appeared to promote learning in terms of orientation outcomes (attitudes, behaviors, and sensitivity). This finding is only tentative, as no regression could be performed, and the different effects found in the moderator analysis could be based on the influence of other variables. No support was found for effects on judgment-related learning. For knowledge acquisition, a concentration on individual learning appeared to be more fruitful than any group-based activities.

H4: As expected, practically oriented courses that emphasized experiential learning in dealing with concrete cases were more effective in promoting judgment. Unexpectedly, this effect also occurred for the acquisition of knowledge. Overall, a practical emphasis appears to be the best predictor of high-impact RCR courses—independently of the concrete method that is used (e.g., role play, personal assessment plus coaching, reading an interactive novel). This finding probably marks the most important outcome of our review. It is noteworthy because deliberative case discussion continues to be considered the "gold standard" of ethics education.[6] Moreover, this finding is novel because previous reviews did not consider general teaching approaches, and mainly focused on specific contents and methods of instruction (e.g., Torrence et al., 2017; Mulhearn et al., 2017; Watts et al., 2017).

H5: As expected, courses that introduced and applied ethical rules, standards, or guidelines for RCR were more effective in promoting knowledge. As expected, no positive effect was found for judgment-related learning. Overall, courses that did *not* have students apply ethical guidelines tended to be most effective though. An explanation for this paradoxical finding could be that attitudinal and behavioral learning is hampered through *reactance* when people are expected to adopt evaluations that they have not concluded autonomously (Worchel and Brehm, 1971). Additionally, presenting learners with the solution (guidelines) before engaging

[6] For instance, in a European survey of 99 research-integrity teachers, half suggested that the discussion of cases was the best approach (Andorno et al., 2019).

Table 4 Meta-regression analysis: best models

	Knowledge		**Judgment**		**All combined**	
Covariate (Ref.)	*Coef*	*95% CI*	*Coef*	*95% CI*	*Coef*	*95% CI*
Intercept	0.51***	[0.30, 0.72]	0.97***	[0.49, 1.44]	0.73***	[0.39, 1.07]
Instruction (mixed)						
Pure individual	**0.55***	[0.08, 1.01]	*c*		*c*	
Pure group	−0.23	[−0.50, 0.03]	*c*		*c*	
Emphasis (deliberative)						
Theoretical	**−0.53***	[−0.99, −0.07]	–	–	0.14	[−0.11, 0.40]
Practical	**0.33***	[0.02, 0.64]	**1.10*****	[0.79, 1.41]	**0.74*****	[0.54, 0.95]
Ethical guidelines (Sup.)						
No	**−1.03****	[−1.72, −0.33]	*c*		**0.35***	[0.09, 0.61]
Applied	**0.26****	[0.07, 0.45]	*c*		0.13	[−0.10, 0.37]
Course duration (<5h)						
5–10h	*c*		*c*		**0.36***	[0.08, 0.63]
10–20h	*c*		*c*		**0.42***	[0.10, 0.74]
>20h	*c*		*c*		**0.32***	[0.01, 0.63]
Inst. Recognition (V/ext.)						
Voluntary, internal	−0.09	[−0.32, 0.14]	*c*		*c*	
Mandatory	**-0.31****	[−0.52, −0.10]	*c*		*c*	
No. Methods (Cont., 1–9)	*c*		**−0.17*****	[−0.25, −0.08]	**−0.13*****	−0.18, −0.08]
Feedback on Exercises	*c*		*c*		−0.13	−0.27, 0.02]
Response Rate (<50%)						
50–74%	*c*		*c*		**−0.26****	[-0.45, −0.08]
75%+	*c*		*c*		**−0.29****	[−0.47, −0.12]
Content validity						
Reported	*c*		*c*		0.19	[−0.01, 0.38]
Target group (Res.)						
School/citizens	*c*		**1.02***	[0.24, 1.79]	0.35	[−0.09, 0.80]
HE/professionals	*c*		**−0.30****	[-0.50, −0.11]	−0.17	[−0.36, 0.01]
Mixed	*c*		**−0.53***	[−0.98, −0.07]	−0.33	[−0.72, 0.06]
Gender Mix (Mixed)						
<30% female	*c*		*c*		**0.27****	[0.07, 0.47]
>70% female	*c*		*c*		−0.33	[−0.81, 0.14]
LO type (knowledge)						
Attitude					**−0.30***	[−0.55, −0.05]
Behavior					−0.06	[−0.30, 0.19]
Judgment					**−0.31*****	[−0.48, −0.14]
Sensitivity					**−0.29***	[−0.52, −0.06]
Test of model						
F	10.76***		20.70***		8.62***	
Goodness of fit						
Tau^2	$0.0000^{n.s.}$		$0.0029^{n.s.}$		$0.0058^{n.s.}$	
I^2 *(unexplained)*	0.00%		10.40%		12.10%	
Explained variance						
R^2 *analog*	1.00		0.98		0.95	
From Null-I^2	75.54%		85.06%		79.15%	
No. of Studies	28		23		75	

Note. Coef, regression coefficient based on weighted mean effect size. *Ref.*, reference group in comparison to which the coefficient describes a relative difference. *Sup.*, superficial. *V/ext.* voluntary/external. *Cont.*, continuous variable: coefficient describes average distance between two levels. *Res.*, researchers. *c* = statistically non-significant effect removed from the best model to reduce collinearity. Significance test levels: *$p < .05$, **$p < .01$, ***$p < .001$ (*n.s.* = not significant)

them in deliberation and problem solving could lead to reduced learning efforts and insights. It appears that the training of higher-order learning outcomes ought to be separated from learning how to apply codes of conduct. This contradicts the widespread idea that "learners may need to comprehend guidelines before acquiring ethical decision-making skills" (Antes, 2014, p. 55). Instead, it might be necessary to identify approaches that allow for a constructive introduction of ethical guidelines, e.g., by having learners apply them to complex cases, which require individual judgment. This finding is highly relevant, because it is opposite to the prior finding of Torrence et al. (2017) by which effective RCR courses introduced and explained rules, standards, and guidelines for RCR, irrespective of the learning outcome. Finally, it is particularly important because higher-order learning outcomes tend to predict actual behavior better than knowledge (Reed et al., 2007), and "the ultimate objective of RCR education is to foster ethical behavior" (Antes and DuBois, 2014, p. 109).

H6: Unexpectedly, courses that were offered to single-domain groups of learners were not more effective than courses offered to groups of learners from multiple domains. The moderator analysis suggests that mixed groups may even be beneficial for orientational learning outcomes. This stands in contradiction to previous reviews, one of which stressed that mixed-domain groups were a major deterrent of effective RCR education and should be avoided as a "golden rule" (Mulhearn et al., 2017). This finding merits further investigation. We assume that mixed-group courses may have had other characteristics that led to their lack of effectiveness.

H7: In contrast to our hypothesis, courses that challenged learners to practice their abilities repeatedly did not appear to be more effective than courses with little or no repetition. The number of treated cases had no significant impact on either learning outcome. One possible explanation is that courses that treated very many cases may have done so superficially. We recommend that future studies examine the power of exercise experimentally, comparing the outcomes of courses with more/less practice.

H8: Against our predictions, courses that employed blended learning did not tend to be more effective than pure online or pure face-to-face courses. One possible explanation is the small number of studies that employed blended learning. However, knowledge acquisition appears to have benefited most strongly from pure individual learning, which could be performed online, for instance. Therefore, the added value of blended learning merits additional scrutiny in future studies.

H9: Estimating the impact of teacher competence was impossible due to lack of data. Only very few studies shared relevant information.

H10: Contrary to our expectations, making courses mandatory had a negative impact on one learning outcome, knowledge acquisition, and no impact on the other learning outcomes. A greater motivation of voluntary (self-selected) participants may explain this effect. However, the moderator analysis suggests that orientational learning outcomes may have benefited from courses' mandatory nature. Overall, we wish that we could have operationalized institutional recognition better than we did, which basically boiled down to the question whether courses were mandatory or not. Due to the lack of data from prior studies, we suggest that future studies investigate the relevance of institutional recognition experimentally and operationalize institutional recognition with more variables. Our data only suggests that learners participating in courses voluntarily tended to acquire more knowledge than participants of mandatory courses.

H11: Unexpectedly, effects found for different groups of learners did differ significantly when controlling for other expected influences (at least in view of moral judgment): Overall, it appears that high-school students and researchers tended to learn more than students of higher education. One explanation could be a lack of motivation of non-research professionals (e.g.,

psychology students) to deliberate about RCR. The larger effects found for high-school students may also be explained through the fact that moral judgment is known to advance more strongly among adolescents than adults (Rest et al., 1999). Predominantly, male groups also tended to make greater advances. This may be due to a greater average maturity of women in terms of moral sensitivity, for instance (You et al., 2011), or a lower tendency of men to identify themselves as moral/ethical (Yang et al., 2017). Both factors might lead to women learning "less" because they cannot achieve the same pre-/post-differences that men do.

Limitations and Directions for Future Research

By employing multivariate meta-regression to test a series of hypotheses, our review delivers findings that are arguably more robust than those of previous reviews. A practical course orientation with an emphasis on experiential learning and an emotional engagement with ethical decision-making appears to be the best predictor of effective RCR education: relevant effects were found for each learning outcome, and when excluding diverse single studies from the analysis. In contrast, our other findings are less robust. For instance, if more studies had employed blended learning, we might have seen a positive effect here.

Several limitations are worth mentioning here. First, we did not control systematically whether studies reported their findings selectively. What we do know is that some findings from the included studies were under-reported so that we could not calculate effect sizes on their basis. For instance, Canary et al. (2012) used the Defining Issues Test (DIT) to measure judgment (in addition to the ESIT), but no statistics were reported because no significant effect was found. If we had combined both scores (DIT + ESIT), the judgment effect sizes would have been smaller for this study, which would have had an impact on our results. However, considering how the DIT did not display significant results in any included study, even when large effects were found for other learning outcomes (Bernstein et al., 2010), future studies may want to investigate effects for general judgment and RCR-focused judgment separately. Watts et al. (2017) already found that "off-the-shelf measures" like the DIT tended to yield smaller effect sizes than custom measures for RCR education.

In general, there were diverse limitations to the data that we worked with. A couple of studies only reported p-values of <.001, based on which we computed effect sizes: If they had reported the T-statistics, the effect sizes would have been larger for these studies, because we had to calculate with a p-value of .001 instead of what may have been a p-value of .00004. Overall, many studies did not provide important information about the courses, e.g., the exact course duration (which we then estimated). Due to this lack of information, we were not able to investigate the effect of teachers' competence. Future studies could report this kind of information more systematically. We suggest that authors and reviewers check whether the information is available, which we employed in this review, including diverse risks of bias.

Due to the small number of studies, no meta-regressions were possible for attitudinal, behavioral, and sensitivity-related learning. The moderator analysis for these orientation outcomes suggests that these types of learning behaved differently than the development of knowledge and judgment. However, without more studies, there is no way to tell. One option could be to work with studies across all domains of ethics education, e.g., including business and medical ethics studies, to investigate these outcomes. Cross-disciplinary reviews (e.g., Mulhearn et al., 2017) indicate that differences between disciplines of ethics training may be negligible.

As these considerations show, the inclusion of further studies with better information in future meta-analyses may lead to clearer results, some of could end up contradicting some of our findings. To build the knowledge basis of what works in RCR education, we would like to articulate the following recommendations: First, we need replication studies, which test the effectiveness of well-elaborated teaching approaches with diverse groups of learners. Second, our field would strongly benefit from added-value research: In randomized control trials with two of more groups, learners participate in the exact same course with only one difference, e.g., whether exercises are conducted in groups or individually. Finally, authors could benefit from a relatively robust finding of our review that methods that promote one type of learning (e.g., the development of judgment) may not be helpful in promoting other types of learning (e.g., sensitivity to ethical problems). We suggest that colleagues select several measures of good quality for their studies and contrast the results per learning outcome.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10648-021-09630-9.

Author Contribution Johannes Katsarov: conceptualization, methodology, formal analysis, data curation, writing—original draft, project administration, project administration. Roberto Andorno: conceptualization, methodology, validation, writing—review and editing, supervision. André Krom: conceptualization, methodology, validation, writing—review and editing, supervision. Mariëtte van den Hoven: conceptualization, methodology, validation, writing—review and editing, supervision, funding acquisition.

Funding Open Access funding provided by Universität Zürich. This research was supported by the European Union's Horizon 2020 Research and Innovation Programme [grant no. 824586].

Availability of Data and Material (Data Transparency) On request.

Code Availability Not applicable.

Declarations

Competing Interests The authors declare no competing interests.

References

ALLEA. (2017). *The European code of conduct for research integrity* (rev. ed.). All European Academies.

Andorno, R., Katsarov, J., & Rossi, S. (2019). *Results of mapping of current practice.* Public report of the H2020 project INTEGRITY. Retrieved in May 2021 from https://h2020integrity.eu/wp-content/uploads/2019/12/D3.2-Results-of-mapping-current-practice.pdf.

Antes, A. L. (2014). A systematic approach to instruction in research ethics. *Accountability in Research, 21*(1), 50–67. https://doi.org/10.1080/08989621.2013.822269.

Antes, A. L., & DuBois, J. M. (2014). Aligning objectives and assessment in responsible conduct of research instruction. *Journal of Microbiology and Biology Education, 15*(2), 108–116. https://doi.org/10.1128/jmbe.v15i2.852.

Antes, A. L., Murphy, S. T., Waples, E. P., Mumford, M. D., Brown, R. P., Connelly, S., & Devenport, L. D. (2009). A meta-analysis of ethics instruction effectiveness in the sciences. *Ethics & Behavior, 19*(5), 379–402. https://doi.org/10.1080/10508420903035380.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (2nd ed.). Academic Press.

Cook, D. A., & Reed, D. A. (2015). Appraising the quality of medical education research methods: The medical education research study quality instrument and the Newcastle-Ottawa scale-education. *Academic Medicine, 90*(8), 1067–1076. https://doi.org/10.1097/ACM.0000000000000786.

Deeks, J. J., Higgins, J. P. T., & Altman, D. G. (2019). Chapter 10: Analysing data and undertaking meta-analyses. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.), *Cochrane handbook for systematic reviews of interventions version 6.0* Available from www.training.cochrane.org/handbook.

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods, 9*(4), 426–445. https://doi.org/10.1037/1082-989X.9.4.426.

Higgins, J. P. T., Thompson, S., Deeks, J. J., & Altman, D. G. (2002). Statistical heterogeneity in systematic reviews of clinical trials: A critical appraisal of guidelines and practice. *Journal of Health Services Research & Policy, 7*(1), 51–61. https://doi.org/10.1258/1355819021927674.

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

Kalichman, M. K. (2013). A brief history of RCR education. *Accountability in Research, 20*(5), 380–394. https://doi.org/10.1080/08989621.2013.822260.

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology, 78*(2), 311–328. https://doi.org/10.1037/0021-9010.78.2.311.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 1–12. https://doi.org/10.3389/fpsyg.2013.00863.

Macrina, F. L. (2014). *Scientific integrity. Text and cases in responsible conduct of research* (4th ed.). ASM Press.

Maesschalck, J., & De Schrijver, A. (2015). Researching and improving the effectiveness of ethics training. In A. Lawton, Z. van der Wal, & L. Huberts (Eds.), *Ethics in public policy and management: A global research companion* (pp. 198–212). Routledge.

Martakis, K., Czabanowska, K., & Schröder-Bäck, P. (2016). Teaching ethics to pediatric residents. A literature analysis and synthesis. *Klinische Pädiatrie, 228*(5), 263–268. https://doi.org/10.1055/s-0042-109709.

Marušic, A., Wager, E., Utrobicic, A., Rothstein, H. R., & Sambunjak, D. (2016). Interventions to prevent misconduct and promote integrity in research and publication. *Cochrane Database of Systematic Reviews, 4*(4), MR000038. https://doi.org/10.1002/14651858.MR000038.pub2.

*Melcer, E. F., Grasse, K. M., Ryan, J., Junius, N., Kreminski, M., Squinkifer, D., Hill, B., & Wardrip-Fruin, N. (2020). Getting academical: a choice-based interactive storytelling game for teaching responsible conduct of research. *Proceedings of FDG '20, Sept. 15-18, 2020, Bugibba, Malta, Article No. 78, 1-12*. https://doi.org/10.1145/3402942.3403005.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097.

Mulhearn, T. J., Steele, L. M., Watts, L. L., Medeiros, K. E., Mumford, M. D., & Connelly, S. (2017). Review of instructional approaches in ethics education. *Science and Engineering Ethics, 23*(3), 883–912. https://doi.org/10.1007/s11948-016-9803-0.

Pigott, T. D., & Polanin, J. R. (2015). The use of meta-analytic statistical significance testing. *Research Synthesis Methods, 6*(1), 63–73. https://doi.org/10.1002/jrsm.1124.

Reed, D. A., Cook, D. A., Beckman, T. J., Levine, R. B., Kern, D. E., & Wright, S. M. (2007). Association between funding and quality of published medical education research. *JAMA, 298*(9), 1002–1009. https://doi.org/10.1001/jama.298.9.1002.

Rest, J. R., Narvaez, D., Thoma, S. J., & Bebeau, M. J. (1999). DIT2: Devising and testing a revised instrument of moral judgment. *Journal of Educational Psychology, 91*(4), 644–659. https://doi.org/10.1037/0022-0663.91.4.644.

Steneck, N. H. (2007). *ORI introduction to the responsible conduct of research*. Office of Research Integrity.

Tanner, C., & Christen, M. (2014). Moral intelligence – A framework for understanding moral competences. In M. Christen, J. Fischer, M. Huppenbauer, C. Tanner, & C. van Schaik (Eds.), *Empirically informed ethics* (pp. 119–136). Springer.

Todd, E. M., Torrence, B. S., Watts, L. L., Mulhearn, T. J., Connelly, S., & Mumford, M. D. (2017a). Effective practices in the delivery of research ethics education: A qualitative review of instructional methods. *Accountability in Research, 24*(5), 297–321. https://doi.org/10.1080/08989621.2017.1301210.

Todd, E. M., Watts, L. L., Mulhearn, R. J., Torrence, B. S., Turner, M. R., Connelly, S., & Mumford, M. D. (2017b). A meta-analytic comparison of face-to-face and online delivery in ethics instruction: The case for a hybrid approach. *Science and Engineering Ethics, 23*(6), 1719–1754. https://doi.org/10.1007/s11948-017-9869-3.

Torrence, B. S., Watts, L. L., Mulhearn, T. J., Turner, M. R., Todd, E. M., Mumford, M. D., & Connelly, S. (2017). Curricular approaches in research ethics education: Reflecting on more and less effective practices in instructional content. *Accountability in Research, 24*(5), 269–296. https://doi.org/10.1080/08989621.2016.1276452.

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics, 35*(2), 215–247. https://doi.org/10.3102/1076998609346961.

Watts, L. L., Medeiros, K. E., Mulhearn, T. J., Steele, L. M., Connelly, S., & Mumford, M. D. (2017). Are ethics training programs improving? A meta-analytic review of past and present ethics instruction in the sciences. *Ethics & Behavior, 27*(5), 351–384. https://doi.org/10.1080/10508422.2016.1182025.

Worchel, S., & Brehm, J. W. (1971). Direct and implied social restoration of freedom. *Journal of Personality and Social Psychology, 18*(3), 294–304. https://doi.org/10.1037/h0031000.

Yang, J., Ming, X., Wang, Z., & Adams, S. M. (2017). Are sex effects on ethical decision-making fake or real? A meta-analysis on the contaminating role of social desirability response bias. *Psychological Reports, 120*(1), 25–38. https://doi.org/10.1177/0033294116682945.

You, D., & Bebeau, M. J. (2013). The independence of James Rest's components of morality: Evidence from a professional ethics curriculum study. *Ethics and Education, 8*(3), 202–216. https://doi.org/10.1080/17449642.2013.846059.

You, D., Maeda, Y., & Bebeau, M. J. (2011). Gender differences in moral sensitivity: A meta-analysis. *Ethics & Behavior, 21*(4), 263–282. https://doi.org/10.1080/10508422.2011.585591.

Affiliations

Johannes Katsarov[1,2] · Roberto Andorno[1] · André Krom[3] · Mariëtte van den Hoven[3]

[1] Center for Ethics, University of Zurich, Zollikerstrasse 117, CH-8008 Zurich, Switzerland

[2] NICE Foundation, Burg Bas Backerhof 20, 2111 TDAerdenhout, The Netherlands

[3] Ethics Institute, Utrecht University, Janskerkhof 13, 3512 BLUtrecht, The Netherlands