INTEGRITY

D3.1 Manuscript with literature review

Deliverable Code	D3.1
Distribution Level	Public
Responsible Partner	UZH
Authors	Johannes Katsarov (UZH), Roberto Andorno (UZH), , André Krom (UU), Mariëtte vd Hoven (UU)
Checked by	Orsolya Varga (UD) Date: 29-12-2020
Approved by Project Coordinator	Mariëtte van den Hoven (UU) Date: 30-12-2020

This project received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 824586

Contents

Introduction

The Horizon 2020 Integrity project aims to empower students and early career researchers through education for Responsible Coduct of Research (RCR). It does so in an evidence-based way, and by using a scaffolded approach. This means that education for RCR will be tailored 1) to their educational level and discipline, and 2) to the specific needs that student groups may have to enable them to responsibly navigate issues of research integrity – current and new (i.e. in a way that is 'future proof').

To these aims this report (D3.1) presents the results of a literature view (meta analysis) on the impact of RCR education. The UZH partner has, together with the UU partner conducted a systematic review of the literature on the effectiveness of RCR education and tested 10 hypotheses for their robustness. We deemed that this was necessary, as several of the recent meta reviews that have been published on the effectiveness of RCR education do not show unilateral results, nor are not always based on the same type of studies. The result of the literature review is submitted to an international peer review journal on Dec 28, 2020. Below, the proof reading of the submission is to be found.

Summary of the deliverable

The deliverable 3.1 consists in the submission of a manuscript to a peer-reviewed journal presenting the results of a systematic review of the literature on the teaching of research integrity. It is described by the Annex 1 to the Grant Agreement as follows:

> "This review will map literature on teaching research integrity for three different groups. The literature will be reviewed in the period of 1990-2018 and will be categorized according to relevant criteria, including target groups, disciplines, aim and focus of the teaching tools, working method, evaluations and assessment of effects. The review will be presented as a manuscript for an international peer reviewed journal".

Accordingly, between months 1 and 24 we reviewed scientific studies published between 1990 and early 2020 that investigated the effectiveness of education in responsible conduct of research (RCR). On 28 December 2020, the result of this work was submitted as a manuscript entitled "Education for a Responsible

Conduct of Research. A Meta-Analytical Review" to *Educational Psychology Review,* a peer-reviewed journal published by Springer (Impact Factor in 2019: 5.167).[1] The full article in available in the Annex.

The paper's abstract summarizes the study and its main findings as follows:

> This article reviews educational efforts to promote a responsible conduct of research (RCR) that were reported in scientific journals between 1990 and early 2020. Unlike previous reviews that were exploratory in nature, this review aimed to test eleven hypotheses on successful 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 confirms 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. Differentiated analyses on the attainment of attitudinal, behavioral, and sensitivity-related learning outcomes were not possible due to the small number of relevant studies. Several avenues for future research efforts are suggested to advance knowledge on the effectiveness of research integrity training.

In the following, we present succinctly the work done during these two years and the main findings, which can be found in detail in the full article in the Annex.

The first important finding we made while starting our research, at the beginning of 2019, was that a group of US researchers had recently conducted a very comprehensive meta-analysis on the effectiveness of RCR training,[2] alongside two qualitative studies on related questions. Taking especially this comprehensive work into account, our research began by scrutinizing the outcomes of these meta-studies and others in the field of ethics training. This was also an impetus to focus on these meta-reviews in task 3.3.

[1] See: https://www.springer.com/journal/10648

[2] 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.

Based on that initial research, in January 2020 we presented a preliminary list of "Ten Features of Good Research Integrity Courses" at the INTEGRITY meeting in Vilnius (see Annex). These ten features were deduced from the available review studies, ensuring that they were corroborated by means of different methods and that the underlying data basis was strong (e.g., only considering factors that were found in 10 studies and more). At that time, we warned that a shortcoming of the available meta-reviews was that the effects of methods and contents were not distinguished in terms of whether they support the attainment of different types of learning outcomes. This is why, unlike prior meta-analyses on this issue,[3] our investigation would explicitly consider that *different teaching approaches* may be needed to achieve *different kinds of learning outcomes* related to RCR.

For our literature review, we screened 1.548 abstracts of papers related to the topic, and after careful analysis, we selected 84 articles for full review. Using a pre-configured Excel table, we extracted information from each of the selected studies regarding twelve criteria, such as for instance target groups, course duration, course emphasis (theoretical, deliberative, or experiential), use of cases, use of online tools, etc. The studies were coded independently by the four members of our team. Where codings diverged, criteria and interpretations were discussed until consensus was found. In the end, 30 studies met our criteria for inclusion, as they investigated the extent to which learners' knowledge, attitudes, or competences related to RCR improved due to an educational intervention, whereby pretests and/or a control group were used to measure the course outcome. Overall, we were able to test 10 hypotheses about effective RCR courses. To do so, we drew on 75 effect sizes that we calculated for five different types of learning outcomes (knowledge of RCR, moral judgment, moral sensitivity, moral attitudes, and moral behavior).

[3] 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*, 379–402; Watts, L.L. et al., op. cit.

It should be mentioned that meta-analyses on ethics training have so far mainly reported the results of so-called *moderator analyses.*[4] In our study, we went beyond the use of moderator analyses and worked with so-called *meta-regression analyses.*[5] Unlike moderator analyses, which only look at the impact of a single variable, meta-regressions take multiple variables into account simultaneously. In this way, we investigated which course characteristics had the greatest impact on the development of moral judgment and the development of RCR-related knowledge, as well as on the achievement of all five learning outcomes overall. Distinct analyses on the development of moral attitudes, sensitivity, and behavior were not possible due to the small number of studies that had assessed these learning outcomes.

In line with our expectations, *different goals of RCR education benefited from different teaching approaches.* This was obvious when looking at the coverage of codes of conduct as part of RCR education. Applying codes of conduct to ethical cases had a significant positive impact on learners' knowledge acquisition. However, *not* covering codes of tended to be a feature of the most effective courses overall. An explanation of this seemingly paradoxical finding is that attitudinal and behavioral learning is hampered through *reactance* when people are expected to adopt evaluations that they have not concluded themselves.[6]

Our second main finding concerns the classification of courses as 'theoretical', 'deliberative' or 'experiential'. We had expected that deliberative courses would yield, on average, larger effects than theoretical courses, and that experiential courses would be even more effective – at least for attitudinal learning, the development of moral judgment, sensitivity, and behavior (but not for the development of knowledge). We did not, however, expect that this 'course emphasis' variable would be the strongest predictor of courses' effectiveness.

[4] Moderator analyses look at the average effect sizes of all studies that reported a certain characteristic, e.g., which used blended learning (the combination of face-to-face and remote learning activities).

[5] Meta-regressions calculate the degree to which different variables (e.g., course characteristics) tend to lead to larger or smaller effect sizes.

[6] Worchel, S., & Brehm, J.W. (1971). Direct and implied social restoration of freedom. *Journal of Personality and Social Psychology, 18*(3), 294–304.

Experiential learning was by far more effective in promoting all types of learning outcomes than the theoretical and deliberative approaches were.

Based on these findings, a few recommendations can be made:

1. RCR education should make use of *experiential learning*, and engage learners in an active reflection of how they would personally deal with challenging situations pertaining to research integrity. Useful approaches could include the discussion of engaging stories,[7] the use of role play and simulations,[8] or techniques like coping-modeling/problem-solving, where learners watch a video of a difficult situation and discuss different courses of action.[9]
2. Teachers should be careful about the way in which they introduce *codes of conduct* for RCR in their courses. Fruitful approaches could be to first win learners' support for rules and regulations pertaining to RCR (e.g., through experiential learning) before concrete codes are presented. Moreover, engaging learners in a critical discussion of codes of conduct may help them to make sense of them and learn to appreciate their value.
3. Teachers should not rely on the belief that learners' acquisition of knowledge will automatically equip them with higher-order competences like the *abilities to notice or solve ethical problems.* Our findings support the contention that we are dealing with distinct, independent learning outcomes. This brings us to our final recommendation, namely, to choose few methods of instruction, and to do so wisely.

[7] For instance, the interactive novel *The Brewsters*. See Rozmus, C.L., Carlin, N., Polczynski, A., Spike, J., & Buday, R. (2015). The Brewsters: A new resource for interprofessional ethics education. *Nursing Ethics, 22*(7), 815–26, DOI:10.1177/0969733014547974.

[8] For instance, the digital game Academical by Melcer et al. See Melcer, E.F., Grasse, K.M., Ryan, J., Junius, N., Kreminski, M., Squinkifer, D., 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, DOI:10.1145/3402942.3403005.

[9] Simola, S.K. (2010). Use of a ''coping-modeling, problem-solving'' program in business ethics education. *Journal of Business Ethics, 96*, 383–401, DOI: 10.1007/s10551-010-0473-6.

ANNEX

1. "Ten features of good research integrity courses" (summary of preliminary findings), January 2020.
2. Literature review manuscript submitted to *Educational Psychology Review,* 28 December 2020.

ANNEX 1: TEN FEATURES OF GOOD RESEARCH INTEGRITY COURSES

Target Group & Interactivity

1. Good research integrity courses are offered to distinct groups of learners. They 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 (1,2,3). A mix of general and specific issues tends to be detrimental to learning (2,3). Ethics courses for several fields, e.g., across the domains of biomedical ethics and engineering, tended to yield negligible effect sizes (3).

2. Good research integrity courses combine approaches of individual and group-based learning to actively engage learners. Active engagement of learners was a common characteristic of the most successful courses (4). Courses tend to be more effective, if they engage people in individual learning (1,4). Moreover, effective courses did not only rely on group interaction, which can deteriorate individual engagement (1,4). Passive learning and a total reliance on group activities were common characteristics of weak courses (4).

Aims & Scope

3. Good research integrity courses aim to equip learners with competences for dealing with common ethical issues of their practice. Courses that place a strong focus on processes of ethical decision making and which challenge learners to develop relevant competences have been found to be among the most effective (2,3). Ineffective research integrity 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 (2). 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 (1). Coverage of possible reasoning errors has also been found to be an important content (7), as well as considering one's own motives, values and emotions (2).

4. Good research integrity courses introduce and explain rules, standards and guidelines for a responsible conduct of research and stress their importance for practice. This was a common feature found for highly effective courses (2). Courses that cover a wider range of RCR topics tend to achieve larger effects (1). 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 (1,7).

Method

5. Good research integrity courses engage learners in imagining how they would deal with cases relevant to their prospective research practice. Courses that applied ethical decision-making competences to real-world cases tended to be more effective than those who did not (1,4,7). Relatively long cases (8+ paragraphs) yielded better results than shorter cases, if they weren't too complex, weren't too emotional, and weren't too realistic (1).

6. Good research integrity courses actively engage learners in practice. Courses with a relatively high degree of practice, i.e., the repeated application of learned abilities and knowledge, typically yielded larger learning effects (1,2), especially when the exercises were performed individually (1).

Resources

7. Good research integrity courses make use of blended learning. Courses that combine face-to-face and online learning activities tend to yield larger effects than courses that only rely on either face-to-face or online learning activities (1,4,8).

8. Good research integrity courses dedicate enough time to the achievement of the relevant learning outcomes. Evidence suggests that some learning outcomes may be achieved in less time than others (1,6). Where the goal is to sensitize learners to a wide range of ethical issues in a domain of research and equip them with various competences, full-semester courses tend to be more effective than shorter courses (6).

9. Good research integrity courses are offered by competent teachers. Experienced trainers tended to achieve better outcomes than inexperienced trainers (1).

10. Good research integrity courses build on institutional recognition. The practice and teaching of research integrity suffer, if relevant courses are viewed as inferior compared to the core curriculum of degree programs, and if researchers in the organization lack a culture of research integrity (5).

References

(1) 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.

(2) 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.

(3) 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.

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

(5) Andorno, R., Katsarov, J., & Rossi, S. (*in preparation*). Integrity Survey 2019.

(6) Katsarov, J., Schmocker, D., Tanner, C., & Christen, M. (*submitted*). Moral Sensitivity Training – A Systematic Review. Submitted to *Educational Psychologist*.

(7) 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.

(8) Todd, E. M., Watts, L. L., Mulhearn, R. J., Torrence, B. S., Turner, M. R., Connelly, S., & Mumford, M. D. (2017). 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.

Educational Psychology Review

Education for a Responsible Conduct of Research – A Meta-Analytical Review

--Manuscript Draft--

Manuscript Number:	
Full Title:	Education for a Responsible Conduct of Research – A Meta-Analytical Review
Article Type:	Meta-Analysis
Keywords:	meta-analysis; meta-regression; integrity; RCR; ethics education
Corresponding Author:	Johannes Katsarov University of Zurich Zurich, SWITZERLAND
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Zurich
Corresponding Author's Secondary Institution:	
First Author:	Johannes Katsarov
First Author Secondary Information:	
Order of Authors:	Johannes Katsarov
	Roberto Andorno, PD Dr. iur.
	André Krom, PhD
	Mariëtte van den Hoven, PhD
Order of Authors Secondary Information:	
Funding Information:	Horizon 2020 Framework Programme (824586) Dr Mariëtte van den Hoven
Abstract:	This article reviews educational efforts to promote a responsible conduct of research (RCR) that were reported in scientific journals between 1990 and early 2020. Unlike previous reviews that were exploratory in nature, this review aimed to test eleven hypotheses on successful 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 confirms 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. Differentiated analyses on the attainment of attitudinal, behavioral, and sensitivity-related learning outcomes were not possible due to the small number of relevant studies. Several avenues for future research efforts are suggested to advance knowledge on the effectiveness of research integrity training.

Education for a Responsible Conduct of Research – A Meta-Analytical Review

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

University of Zurich, Switzerland *Utrecht University, the Netherlands*

* Corresponding author. University of Zurich, Center for Ethics, Zollikerstrasse 117, CH-8008 Zurich, Switzerland. E-mail address: johannes_katsarov@hotmail.de

Funding: This project received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 824586.

Declaration of interest: none

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

Abstract (up to 200 words)

This article reviews educational efforts to promote a responsible conduct of research (RCR) that were reported in scientific journals between 1990 and early 2020. Unlike previous reviews that were exploratory in nature, this review aimed to test eleven hypotheses on successful 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 confirms 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. Differentiated analyses on the attainment of attitudinal, behavioral, and sensitivity-related learning outcomes were not possible due to the small number of relevant studies. Several avenues for future research efforts are suggested to advance knowledge on the effectiveness of research integrity training.

Keywords: meta-analysis, meta-regression, integrity, RCR, ethics education

Click here to view linked References

1. 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 building 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 1990's. 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., honest, fair, and careful) 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, Murphy, Waples, Mumford, Brown, Connelly, & Devenport, 2009; Watts, Medeiros, Mulhearn, Steele, Connelly, & Mumford, 2017). These meta-analyses have been accompanied by further systematic reviews, which explored the effectiveness of RCR courses qualitatively (e.g., Marušic, Wager, Utrobicic, Rothstein, and Sambunjak, 2016; Todd, Torrence, Watts, Mulhearn, Connelly, & Mumford, 2017) or with advanced statistical methods (e.g., Mulhearn, Steele, Watts, Medeiros, Mumford, & Connelly, 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, Torrence et al., 2017, 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 assumption does not reflect the empirical fact that cognitive, skill-based, and affective learning outcomes are structurally different and benefit from different types of learning (Kraiger, Ford, & Salas, 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 & 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 & de Schrijver, 2015), a synthesis of the abilities underlying moral agency (Tanner & Christen, 2014), and a taxonomy of learning objectives for RCR (Antes & 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'. 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, Narvaez, Thoma, & Bebeau, 1999). This is exemplified by a knowledge test question used by Melcer, Grasse, Ryan, and colleagues (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 and colleagues (2017) that educational interventions of less than eight hours' duration (*Md* = 0.61, *k* = 47) tended to be significantly more effective than interventions of 16 hours and more (*Md* = 0.39, *k* = 65): Since 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.

[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.

2. Eleven hypotheses on the effectiveness of RCR courses

Building on the previous critique and the findings from earlier reviews, our meta-analysis aims to test eleven hypothesis on successful RCR training strategies.[3] 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, Torrence et al., 2017, 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. These findings lead us to the following hypothesis:

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, Watts, Mulhearn, Turner, Todd, Mumford, & Connelly, 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, Torrence et al., 2017; 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 ought to be even more effective than courses that discuss ethical cases in view of what is right or good, but which refrain from looking at psychological and social influences. Based on these considerations, we expect that:

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

[3] 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.

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 increased 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.

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, Torrence, et al., 2017; 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, Torrence et al., 2017; Todd, Watts et al., 2017; 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 and colleagues (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 (*blinded for peer review*).

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 (*blinded*) 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, Czabanowska, Schröder-Bäck, 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:

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).

3. Method

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

3.1. 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.

3.2. 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 and colleagues (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 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.

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

3.3 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*; temporarily hidden for double-blind review; see Appendix G*). This registration included the hypotheses stated in Section 2, 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 Section 2. 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, (3) type of instruction (pure individual, pure group, or individual and group learning), (4) 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), (5) introduction and application of ethical guidelines (no; superficial; applied), (6) quality and quantity of engagement with cases, (7) use of e-learning (no; pure e-learning; blended learning), (8) course duration, (9) competence of teachers, (10) institutional recognition of the course, (11) whether the study had undergone peer review, and (12) 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 hours" and "2-5 hours" 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, Cook, Beckman, Levine, Kern, & Wright (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 & 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, Thompson, Deeks, & Altman, 2002).

4. Results

4.1. Study selection

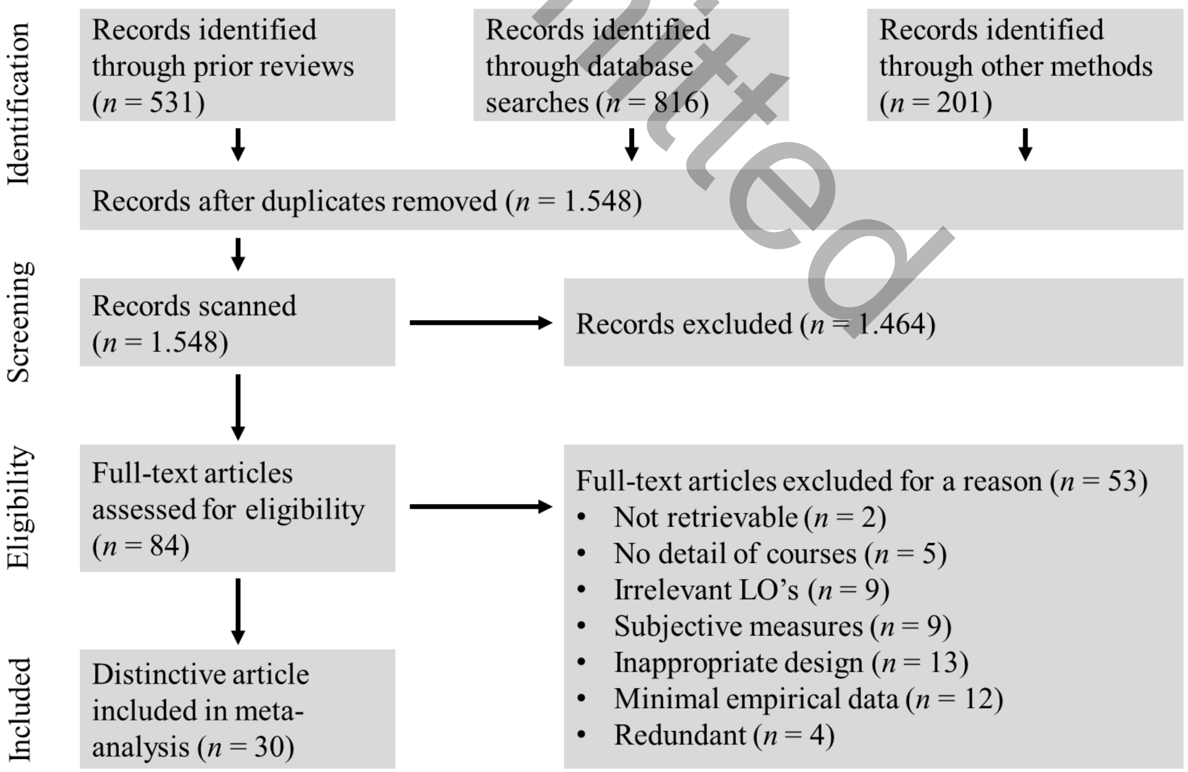
Identification
Screening
Eligibility
Included
Records identified through prior reviews (n = 531)
Records identified through database searches (n = 816)
Records identified through other methods (n = 201)
Records after duplicates removed (n = 1.548)
Records scanned (n = 1.548)
Records excluded (n = 1.464)
Full-text articles assessed for eligibility (n = 84)
Full-text articles excluded for a reason (n = 53)
• Not retrievable (n = 2)
• No detail of courses (n = 5)
• Irrelevant LO’s (n = 9)
• Subjective measures (n = 9)
• Inappropriate design (n = 13)
• Minimal empirical data (n = 12)
• Redundant (n = 4)
Distinctive article included in meta-analysis (n = 30)

Figure 1: PRISMA flow diagram of article 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 and colleagues (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 and colleagues (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.

4.2. 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-, and knowledge-related learning with $I^2 > 50\%$ in these cases (Deeks, Higgins, & Altman, 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 into the outcome type "orientation", a satisfactory heterogeneity of effect sizes was found to warrant moderator analyses for combined effect sizes of attitudinal, behavioral, and sensitivity-related learning. 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 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.

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

4.3. 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.

4.4. Moderator Analysis

We performed separate moderator analyses for knowledge, judgment, and orientation, looking at the mean effect sizes for they 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.

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 Section 4.5 account for the possibility of multiple, interconnected influences, which is why we present the moderator analysis without further comment.

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]
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]
1-4	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.

4.5. 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 ignored 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 hours (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
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						
Tau²	0.0000$^{n.s.}$		0.0029$^{n.s.}$		0.0058$^{n.s.}$	
I² (unexplained)	0.00%		10.40%		12.10%	
Explained Variance						
R² analog	1.00		0.98		0.95	
From Null-I²	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).

5. Discussion

After testing our eleven hypotheses against the studies that were reviewed, we come to the following conclusions. With regards 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 five hours 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 and colleagues (2020). At the same time, the employed measures may be focused on assessing very narrow learning outcomes, which means 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 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. A practical emphasis appears to be the best predictor of high-impact RCR courses.

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 applying 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 & Brehm, 1971). Additionally, presenting learners with the solution (guidelines) before engaging them in deliberation and problem solving could lead to reduced learning efforts and insights. It appears that the training of moral judgment ought to be separated from learning how to apply codes of conduct. Alternatively, 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.

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 and warrants additional investigation.

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.

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 explanation is the small size 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 positive 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 develop more knowledge than participants of mandatory courses.

H11: Unexpectedly, effects found for different groups of learners did differ systematically 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 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, Maeda, & Bebeau, 2011), or a lower tendency of men to identify themselves as moral/ethical (Yang, Ming, Wang, & Adams, 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 and colleagues (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 and colleagues (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 failed to share important information about the courses, e.g., the exact course duration (which we then estimated roughly). Due to this lack of information, we were not able to investigate the effect of teachers' competence. Future studies ought to 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 which may contradict 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 should bear in mind that a relatively robust finding of our review is 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.

Literature

References marked with a * were included in the meta-analysis.

*Aalborg, A., Sullivan, S., Cortes, J., Basagoitia, A., Illanes, D., & Green, M. (2016). Research ethics training of trainers: developing capacity of Bolivian health science and civil society leaders. *Acta Bioethica*, *22*(2), 281–91, DOI:10.1016/j.aogh.2015.02.541.

*Aggarwal, R., Gupte, N., Kass, N., Taylor, H., Ali, J., Bhan, A., … Bolliger, R.C. (2011). A Comparison of Online versus On-Site Training in Health Research Methodology: A Randomized Study. *BMC Medical Education, 11*(37), DOI:10.1186/1472-6920-11-37.

*Ajuwon, A.J., & Kass, N. (2008). Outcome of a research ethics training workshop among clinicians and scientists in a Nigerian university. *BMC Medical Ethics*, *9*(1), DOI:10.1186/1472-6939-9-1.

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), 108–16, DOI: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*, 379–402. DOI:10.1080/10508420903035380.

*Arnott, E., Hastings, P., & Allbritton, D. (2008). Research Methods Tutor: evaluation of a dialogue-based tutoring system in the classroom. *Behavior Research Methods, 40*(3), 694–98, DOI:10.3758/BRM.40.3.694.

*Austin, K.A., Gorsuch, G.J., Lawson, W.D., & Newberry, B.P. (2011). Developing and designing online engineering ethics instruction for international graduate students. *Instructional Science, 39*(6), 975–97, DOI:10.1007/s11251-010-9162-1.

*Barber, L.K., Bailey, S.F., & Bagsby, P.G. (2015). Improving Research Participant Ethics: The Utility of an Online Educational Module. *Teaching of Psychology*, *42*(2), 143–48, DOI:10.1177/0098628315573137.

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

*Canary, H.E., Herkert, J.R., Ellison, K., & Wetmore, J.M. (2012). 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, Texas.

*Clarkeburn, H., Downie, J.R., & Matthew, B. (2002). Impact of an Ethics Programme in a Life Sciences Curriculum. *Teaching in Higher Education*, *7*(1), 65–79, DOI:10.1080/13562510120100391.

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), 1067–76, DOI: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

*Dols, J.D., Hoke, M.M., & Rauschhuber, M.L. (2017). Mock Institutional Review Board: Promoting Analytical and Reasoning Skills in Research Ethics. *Nurse Educator*, *42*(6), E40–E8, DOI:10.1097/NNE.0000000000000377.

*DuBois, J.M., Dueker, J.M., Anderson, E.E., & Campbell, J. (2008). The development and assessment of an NIH-funded research ethics training program. *Academic Medicine*, *83*(6), 596–603, DOI:10.1097/ACM.0b013e3181723095.

*DuBois, J.M., Chibnall, J.T., Tait, R., & Vander Wal, J.S. (2018). The professionalism and integrity in research program: Description and preliminary outcomes. *Academic Medicine 93*(4), 586–92, DOI:10.1097/ACM.0000000000001804.

*Fisher, C.B., & Kuther, T.L. (1997). Integrating research ethics into the introductory psychology course curriculum. *Teaching of Psychology*, *24*(3), 172–75, DOI:10.1207/s15328023top2403_4.

*Folayan, M.O., Adaranijo, A., Durueke, F., Ajuwon, A., Adejumo, A., Ezechi, O., ... & Akanni, O. (2014). Impact of Three Years Training on Operations Capacities of Research Ethics Committees in Nigeria. *Developing World Bioethics*, *14*(1), 1–14, DOI:10.1111/j.1471-8847.2012.00340.x

*Fowler, S.R., Zeidler, D.L., & Sadler, T.D. (2009). Moral Sensitivity in the Context of Socioscientific Issues in High School Science Students. *International Journal of Science Education, 31*(2), 279–96, DOI:10.1080/09500690701787909.

*Han, H., & Jeong, C. (2014). Improving epistemological beliefs and moral judgment through an STS-based science ethics education program. *Science and Engineering Ethics, 20*(1), 197–220, DOI:10.1007/s11948-013-9429-4.

*Heitman, E., Salis, P.J., & Bulger, R.E. (2002). Teaching Ethics in Biomedical Science: Effects on Moral Reasoning Skills. In N.H. Steneck & M.D. Scheetz (Eds.), *Investigating Research Integrity. Proceedings of the First ORI Research Conference on Research Integrity* (pp. 195–202). Rockville: Office of Research Integrity.

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, DOI:10.1258/1355819021927674.

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

Kalichman, M.K. (2013). A Brief History of RCR Education. *Accountability in Research, 20*(5), 380–94, DOI:10.1080/08989621.2013.822260.

*Kim, E.-A., Park, E.-Y., Lim, S.-M., & Yang, I.-H. (2016). Development and application of an education program based on socio-scientific issues for enhancing students' understanding of the nature of science and moral sensitivity. *Information*, *19*(8), 3427–32.

*Kligyte, V., Marcy, R.T., Waples, E.P., Sevier, S.T., Godfrey, E.S., Mumford, M.D., & Hougen, D.F. (2008). Application of a sensemaking approach to ethics training in the physical sciences and engineering. *Science and Engineering Ethics, 14*(2), 251–78, DOI:10.1007/s11948-007-9048-z.

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), 311–28, DOI: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*, DOI:10.3389/fpsyg.2013.00863.

Macrina, F.L. (2014). *Scientific Integrity. Text and Cases in Responsible Conduct of Research* (4th Ed.). Washington: 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). New York: Routledge.

Martakis, K., Czabanowska, K., & Schröder-Bäck, P. (2016). Teaching Ethics to Pediatric Residents A Literature Analysis and Synthesis. *Klinische Pädiatrie, 228*(5), 263–68, DOI: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*, DOI:10.1002/14651858.MR000038.pub2.

*McCormack, W.T., & Garvan, C.W. (2014). Team-based learning instruction for responsible conduct of research positively impacts ethical decision-making. *Accountability in Research: Policies & Quality Assurance 21*, 34–49. DOI:10.1080/08989621.2013.822267.

*Melcer, E.F., Grasse, K.M., Ryan, J., Junius, N., Kreminski, M., Squinkifer, D., … 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*, DOI: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, DOI: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, DOI:10.1007/s11948-016-9803-0.

*Mumford, M.D., Connelly, S., Brown, R.P., Murphy, S.T., Hill, J.H., Antes, A.L., Waples, E.P., & Devenport, L.D. (2008). A sensemaking approach to ethics training for scientists: Preliminary evidence of training effectiveness. *Ethics & Behavior, 18*(4), 315–39, DOI:10.1080/10508420802487815.

*Ogunrin, O.A., Ogundiran, T.O., & 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), DOI:10.1186/1472-6939-14-1.

Pigott, T.D., & Polanin, J.R. (2015). The Use of Meta-Analytic Statistical Significance Testing. *Research Synthesis Methods, 6*(1), 63–73, DOI:10.1002/jrsm.1124.

*Powell, S.T., Allison, M.A., & Kalichman, M.W. (2007). Effectiveness of a responsible conduct of research course: a preliminary study. *Science and Engineering Ethics, 13*(2), 249–64, DOI: 10.1007/s11948-007-9012-y.

*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*(7), XC1–XC3, DOI:10.7860/JCDR/2014/8825.4561.

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), 1002–9, DOI: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*, 644–59, DOI:10.1037/0022-0663.91.4.644.

*Roberts, L.W., Warner, T.D., Dunn, L.B., Brody, J.L., Hammond, K.A.G., & Roberts, B.B. (2007). Shaping medical students' attitudes towards ethically important aspects of clinical research: Results of a randomized, controlled educational intervention. *Ethics & Behavior, 17*(1), 19–50, DOI:10.1080/10508420701309937.

*Roberts, L.W., Warner, T.D., Hammond, K.A.G., Brody, J.L., Kaminsky, A., & Roberts, B.B. (2005). Teaching medical students to discern ethical problems in human clinical research studies. *Academic Medicine, 80*(10), 925–30.

*Rozmus, C.L., Carlin, N., Polczynski, A., Spike, J., & Buday, R. (2015). The Brewsters: A new resource for interprofessional ethics education. *Nursing Ethics, 22*(7), 815–26, DOI:10.1177/0969733014547974.

*Seiler, S.N., Brummel, B.J., Anderson, K.L., Kim, K.J., Wee, S., Gunsalus, C.K, & Loui, M.C. (2011). Outcomes assessment of role-play scenarios for teaching responsible conduct of research. *Accountability in Research, 18*(4), 217–46, DOI:10.1080/08989621.2011.584760.

Steneck, N.H. (2007). *ORI Introduction to the Responsible Conduct of Research*. Washington: Office of Research Integrity.

*Strohmetz, D.B., & Skleder, A.A. (1992). The use of role-play in teaching research ethics: A validation study. *Teaching of Psychology, 19*(2), 106–8, DOI:10.1207/s15328023top1902_11.

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), Berlin: Springer.

Todd, E.M., Torrence, B.S., Watts, L.L., Mulhearn, T.J., Connelly, S., & Mumford, M.D. (2017). Effective practices in the delivery of research ethics education: A qualitative review of instructional methods. *Accountability in Research, 24*(5), 297–321, DOI: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. (2017). 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, DOI: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, DOI:10.1080/08989621.2016.1276452.

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, DOI: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, DOI: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, DOI: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, DOI: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, DOI:10.1080/10508422.2011.585591.