Educational Psychology Review (2026) 38:94
https://doi.org/10.1007/s10648-026-10187-8

REVIEW ARTICLE

# How is the Impact of Responsible Conduct of Research Education Evaluated? A Review of Measures Used and a Content Analysis of Outcome Measures with Reported Validation Evidence

**M. van Loon**[1] · **D. Crean**[2] · **M. van den Hoven**[1] · **J. Katsarov**[3] · **N. Evans**[1]

Received: 2 December 2025 / Accepted: 2 June 2026


## Abstract
Education in Responsible Conduct of Research (RCR) aims to foster good research practices. Little, however, is known about how educators measure the impact of their courses, which outcome measures are used and if and how they are validated. In this review, we systematically identify measures used to evaluate RCR education and assess if and how they have been validated. We also analyse a selected subset of these measures. Specifically, those that include supporting evidence of validity and focus on research practices, in order to identify which Responsible Conduct of Research (RCR) concepts they address and which learning outcomes they are designed to assess. In total, 162 articles were included, of which 46 articles contained information about a measure described as 'validated'. Measures from these articles (and four additional measures suggested by experts) were assessed on accessibility, evidence of validation, and content related to research practice, resulting in 15 unique outcome measures with supporting validation evidence. All contained items covering topics which we categorised as pertaining to 'research integrity', whereas a smaller selection contained items covering topics which we categorised as pertaining to 'research ethics'. Whilst all 15 reported at least some form of reliability or validity evidence, only 3 met what we describe as 'strong' validation evidence. Our review revealed that the majority of RCR education evaluations in the published literature have not used measures related to research conduct with reported validation evidence. The comprehensive overview we provide can offer guidance for educators and researchers to select, or further develop measures aligned with core RCR competencies and targeted learning outcomes.

**Keywords** Responsible conduct of research · Research integrity · Research ethics · Scientific integrity · Good research practices

Extended author information available on the last page of the article

Published online: 23 June 2026

## Introduction

Responsible Conduct of Research (RCR) education is increasingly offered in research performing organisations and incentivized through guidelines and funding criteria (Resnik & Dinse, 2012; ALLEA, 2023; Mejlgaard et al., 2020). To ensure uptake, institutions often make RCR courses mandatory, at least for PhD students (Pizzolato & Dierickx, 2021). Education on RCR aims to, broadly, 'promote ethical behaviour (i.e. reliable, honest, respectful, and accountable; ALLEA, 2023) related to research' (Katsarov et al., 2021, p.934). Formal educational courses that come under the umbrella of RCR include those described as dealing with 'good scientific practices' (Fuerholzer et al., 2020), 'scientific/research integrity' (RI) (Abdi et al., 2021), or 'research ethics' (RE) (Todd et al., 2017). These courses often show a wide variety in learning goals, content, and delivery method. For example, RI courses usually focus more on adhering to professional standards in research in order to obtain trustworthy results, whereas RE courses usually focus more on the moral problems associated with research (Steneck, 2006). In this review, RCR is treated as the overarching construct, with RI and RE ethics analysed as distinct but overlapping subdomains.

Underlying the provision of RCR courses is the presupposition that education can help to foster responsible research practices, as RCR education aims to influence all aspects of decision-making through to actions; increasing knowledge and awareness of RCR issues and stimulating acting in accordance with criteria for responsible research (Goddiksen & Gjerris, 2022; Watts et al., 2017b; Kalichman & Plemmons, 2007). Specific approaches also come with their own presuppositions. For example, a course focused on knowledge and awareness of rules and procedures may assume that it is the deficit of these that leads to questionable research practices (QRPs) (Tammeleht & Löfström, 2025), whereas a course that focuses on building the professional integrity of researchers, may assume that professional attitudes and virtues (e.g. honesty) can, and need, to be cultivated through formal education (Evans et al., 2023).

However, such presuppositions are challenged by the finding that RCR courses can have both positive and negative outcomes on researcher behaviour (Antes et al., 2010) and be ineffective for the attainment of educational goals (Marusic et al., 2016; Krom & van den Hoven, 2022). Thus, testing the effectiveness of courses is essential for their continued development and credibility, and to do so requires appropriate and effective assessment measures (Watts et al., 2017a; Crean et al., 2024; Cahapay, 2021; van den Hoven et al., 2023; Smidt et al., 2009; Antes et al., 2009; Kalichman, 2013).

When considering course assessment, we make a distinction between the evaluation of a course in general, and summative assessment, i.e., how well individuals have performed in a course. Evaluation of a course might address elements of implementation, process (e.g. participants' engagement and satisfaction), as well as impact on participants and organisations. Measures of impact on participants particularly might be closely aligned to the content and specific learning outcomes of the course (overlapping somewhat with summative assessment aims), but they might also reflect more abstract constructs reflecting the underlying course presuppositions, for example "decision-making in research" or "moral sensitivity". Often, measures with reported validation evidence are used for these purposes. Validation is important for two separate reasons, (a) it improves the reliability of the evaluation of the single course, and (b) if researchers use the same measures with reported validation evidence, this can enable comparisons of effectiveness across interventions. In contrast, individual summative assessment is (or at least should be) always closely aligned to the specific learning outcomes and content of the course (Biggs, 2003). However, there is little research on how RCR educators evaluate the impact of their courses on participants, what type of measures are used, and if or how these measures are validated (Marusic et al., 2016; Katsarov et al., 2021). This review focuses on RCR-related measures that address research practice. Although aspects of ethics education such as moral learning and ethical reasoning are important outcomes of RCR education, these outcomes should be assessed in relation to the specific context of research. Measures that evaluate knowledge, skills, attitudes, or behaviours directly connected to research practice are therefore more closely aligned with the goals of RCR instruction.

A qualitative overview of the measures available to assess RCR education is needed by the research community, specifically, those involved in the development and deployment of RCR educational courses. Such an overview will enable educators to select appropriate measures that align with the presuppositions and learning outcomes related to their courses. In this review, we aim to systematically identify measures used for the evaluation of RCR education and to assess the extent to which these measures have been validated. We also aim to analyse the content of a subset of these measures, those with reported validation evidence and which contain items related to research (RI and RE), to describe the RCR concepts covered and the learning outcomes targeted.

The study is guided by the following research questions:

1. What measures are used for the evaluation of RCR education?
2. Which of them are specific to the research context and are supported by reported validation evidence and to what extent?
3. What is the content, learning outcome categories and structure of available RCR measures with reported validation evidence?

These questions are reflected in our specific objectives:

A. To systematically identify studies evaluating the impact of RCR education on participants or presenting the development or validation of a measure for the evaluation of RCR courses.
B. To identify the measures used in studies to evaluate the impact of RCR courses on course participants, and associated user manuals and/or articles on measures' development and validation (if available).
C. To assess the level to which those measures with supporting validation evidence and content related to research practice have been validated and with which target groups.
D. To conduct a quantitative and qualitative content analysis of those measures with supporting validation evidence and content related to research practice in relation to the learning outcomes targeted and the RCR concepts covered.
E. To critically compare the identified measures in relation to RE and RI content.

# Methods

## Protocol and Registration

This study builds on Katsarov et al., (2021) systematic review and meta-analysis of studies evaluating the effectiveness of RCR courses, which screened 1548 records that had been identified in four databases and through additional search methods. To identify and select studies, the PRISMA standard for systematic reviews (Moher et al., 2009) was followed (Katsarov et al., 2021:940). The current study extends Katsarov's review in two ways. First, it includes articles that focus on tests, validation, assessment, and games related to RCR education; and secondly, it extends the search period to March 2023. The protocol for the current study was drafted using the PRISMA-SCR guidelines for scoping reviews and pre-registered on OSF with amendments to the original protocol (https://osf.io/z7cjt/).

## Eligibility Criteria

The inclusion and exclusion criteria have been adapted to focus on the identification of tests and measures that can be used for the evaluation of RCR education (Table 1). Our search includes studies that present and/or validate a test, even if it has not been used in a course-related study thus far.

## Information Sources and Search Strategy

Three databases were searched (Web of Science, ProQuest One, and Google Scholar) for articles that include combinations of the following search terms in their title: ("research ethics" OR "research integrity" OR "scientific integrity" OR "responsible conduct of research" OR "RCR") AND (course OR lesson OR test OR validation OR assessment OR game OR teaching OR learning). ProQuest One and Google Scholar were chosen because of the higher likelihood to identify doctoral theses and grey literature. Searches were limited to the titles for efficiency, based on pre-testing of

**Table 1** Inclusion and exclusion criteria

| Criterion | Inclusion | Exclusion |
|---|---|---|
| Language | English | Non-English reported studies |
| Time period | January 1990 to March 2023 | Studies outside the time period |
| Dependent variable | Studies assessing the impact of RCR courses through relevant tests, and studies that present* or validate tests that have been tailored for the evaluation of RCR education | Tests and measures that do not specifically focus on the assessment of RCR course specific learning outcomes or content e.g., general measures of moral development |
| Availability of the study | The full study must be available via a journal or the internet | Unavailable and retracted studies |
| Availability of the test | The measure must be available or be made available to the researchers | Measures that were not published with related articles, and which were not made available to the researchers for review |

*During screening some relevant reviews of studies were included because they 'present' results related to the use of measures in various studies the search string for the previous review (Katsarov et al., 2021). Abstract search led to 5-digit numbers of publications, whereby all of the relevant publications identified in a search of the first 100 entries were also found when only searching the titles. The database searches were performed by JK in the identical manner as for the prior review, following the pre-registered search strategy (https://osf.io/z7cjt/). Each identified publication was added to the same Excel table used in the previous review, if it had not been identified previously. Findings from the previous review were kept in the database, in case they could not be retrieved a second time. The Excel table details, in which database each publication was found, and whether it was also considered in one of diverse previous reviews by Katsarov et al. (2021).

As an additional search strategy, we consulted the websites of known expertise centres (The Online Ethics Center https://onlineethics.org/ and The Bioethics Research Center https://bioethicsresearch.org/) and contacted several known experts in the field to identify relevant outcome measures for RCR courses, which may have not been published yet.

### Selection of Sources of Evidence

**Part 1 - Articles** Search results were downloaded and managed in Excel. Duplicates were removed and articles were screened by MvL and DC on title and abstract to assess if they met the eligibility criteria. The first 50 retrieved articles were screened independently by both MvL and DC to ensure a shared understanding and application of the eligibility criteria. Differences were discussed in order to reach a shared understanding, after which independent screening was continued. MvL and DC screened equal numbers of retrieved articles. Full-text manuscripts were retrieved if the abstract was not sufficient for making a decision. Several checks and discussions between the researchers were held to ensure consistent application of the eligibility criteria throughout the whole screening process.

**Part 2 - Measures** Included articles were assessed for information about the measure used in evaluation studies or developed in measure development articles. Measures were then retrieved via three sources: first, the supplementary material associated with the article, second, via contacting the authors of manuscript or third by downloading from a website dedicated to the dissemination of the measure. Included measures were retrieved along with any user manuals, additional measure development and/or validation articles. Information about the extent to which the measures were validated was retrieved either via the included article, via contacting the authors or via additional materials associated with the measure (like user manuals or additional development articles identified via a web search).

### Data Charting Process and Items

A data charting form was developed for included articles by JK, MvL, DC, and NE. An additional data charting form was developed for measures with reported validation evidence by DC and MvL.

**Part 1 – Articles and associated measures** For all included articles with accessible measures associated with them, the following items were recorded: Authors; Article title; Year of publication; Geographic origin (country of lead author affiliation); Stated aim of article; Measure target group; Measure described as "validated" (yes or no); Measure accessible and with supporting validation evidence (yes or no); Measure content related to the conduct of research? (yes or no). Measures and/or validation evidence were categorized as inaccessible if they were not available online or from the first author of the measure development article subsequent to an email request and multiple reminders.

**Part 2 - Measures related to the conduct of research with reported validation evidence** For the measures with supporting validation evidence and content related to research practice additional items related to extent of validation were recorded: quantitative and qualitative tests of the reliability and validity; results of quantitative tests.

## Analysis

### Analysis of Measures' Approaches to Validation

For the purposes of this review, we made a clear distinction between measures that were only *described as validated* by their authors and measures that were *supported by reported validation evidence* (i.e., where empirical or procedural details were available). Whenever an article stated or implied that a measure was validated, we retrieved all associated materials—including supplementary files, user manuals, and earlier measure-development papers—to determine what forms of validation had been undertaken. Our coding of validation evidence was guided initially by the framework outlined by Fenn et al. (2020), which distinguishes between different forms of reliability and validity evidence. As we analysed the measures, we extended this framework slightly to accommodate additional categories that appeared in the source materials (e.g., global content validity indices, cognitive interview procedures). Importantly, our coding protocol focused on documenting only those forms of evidence that authors explicitly reported; we did not infer validation from statements of intent or from general descriptions of test development. Table 2 summarises the types of evidence we coded.

Reliability categories included internal consistency, split-half or parallel-forms approaches, and (in principle) test–retest reliability. Validity categories included content and face validity procedures, cognitive interviews, correlations with existing instruments (convergent validity), and any reported form of criterion or construct validity. Not all categories were observed in the included studies, but the complete framework enabled consistent classification and comparison. Using this scheme, each measure was coded independently by one reviewer and checked by a second reviewer. Disagreements were resolved through discussion, and coding rules were refined iteratively to ensure consistent application across all instruments. This process yielded a structured overview of which measures were supported by reported validation evidence, and which were only described as validated.

**Table 2** Categories of evidence on reliability and validity considered

| Reliability tests | Validity tests |
|---|---|
| 1. Test–retest reliability (TTR)<br>2. Parallel Forms Correlation (PFC)<br>3. Split-Half Reliability (SHR)<br>4. Internal Consistency Reliability (CA)<br>• e.g. using Cronbach's Alpha (CA)<br>• e.g. using Kuder–Richardson Formulas (KR-20)* | 1. Content validity (CentV*)*<br>• e.g. via Cognitive Interviews (CI) with content experts.<br>• e.g. using Global Content Validity Indicies (S-CVIs)*<br>2. Face Validity (FV)<br>3. Criterion Validity (CrionV)<br>4. Construct Validity (CuctV) |

*Criteria for reliability or validity testing not described in the initial framework by Fenn et al. (2020)

The reliability of measures informs us about the extent to which we can trust that test results are independent of aspects such as the personality of an auditor, the formulation of items, or the time of day. The reliability of quantitative tests can be assessed through diverse strategies: Test-retest reliability implies that a measure delivers the same results repeatedly (presuming that no change should be expected). Parallel and split half reliability approaches also test the reliability of responses of the same group, but avoid practice effects by testing an alternative form of the test (parallel testing) or two comparable halves of the testing (in Split-half testing). *Internal consistency reliability* refers to how well each item in a measure reflects the content or construct under consideration. *Cronbach's alpha* is commonly used to evaluate the internal consistency of a scale, for instance, whether several items can be said to measure different aspects of one underlying concept (e.g., whether people think of themselves as being fair). *Kuder–Richardson formulas* are used for evaluating internal consistency of items with dichotomous choices. A qualitative method often used to improve that reliability (and validity) of measures can be found in *cognitive interviews*. Here, people are interviewed while answering items or questions with the purpose of learning, what misunderstandings could arise, etc.

The validity of measures informs us about the extent to which they assess what they claim to assess. Validity can be safeguarded in various ways. One method is to establish *content validity*, e.g. by asking several experts to take the tests and to discuss the results with them. The goal is to reach agreement with the experts, whether the questions or items adequately gauge the relevant concept. This can be a qualitive approach, such as *cognitive interviewing*, or quantified using a measure which represents the proportion of items that achieved a high relevance rating from experts, such as *Global Content Validity Indices (S-CVIs)*. Face validity refers to the perception of what the measure is about from the perspective of the user and is important for uptake. *Criterion validity*, often divided into *concurrent* and *divergent validity* tests how a new measure correlates with findings of related tests. When positive correlations are expected, concurrent validity is tested. For example, one might expect that knowledge of principles for correct citation will correlate with performance in correct citation. When negative correlations are expected, divergent validity is tested. For example, one might expect positive attitudes towards science to correlate negatively with sloppy answers to research protocols. Finally, when measures aim at gauging concepts that are not directly measurable, the ultimate goal is to establish *construct validity*. Construct validity subsumes content validity and criterion validity.

## Content Analysis of Measures with Reported Validation Evidence

### Qualitative Content Analysis

A qualitative content analysis (Mayring, 2015) was used to identify (1) what RCR conceptsare represented in the outcome measures' items and scenarios, and (2) what types of learning outcomes are assessed. A deductive coding tree was developed, reflecting the RI and RE topics that are covered in the European Code of Conduct for research integrity (ALLEA, 2023) and the European Textbook on Ethics in Research (Hughes et al., 2010) respectively. These were chosen as they provide distinct and specific items related to RI and RE that are applicable across different research domains considered relevant by experts across 27 European countries.

The learning outcome categories of measures items was assessed using a learning outcomes framework developed by Katsarov et al. (2021), which builds on work on learning outcomes in the moral/ethical domain (Maesschalck & De Schrijver, 2015), abilities underlying moral agency (Christen et al., 2014) and learning objectives for RCR (Antes & DuBois, 2014).

All included validated outcome measures were analysed in MaxQDA. The unit of analysis was the individual outcome measure items and any associated scenarios. Items were coded according to the manifest content on RCR concepts and targeted learning outcomes. Due to the diversity in item type, some coding rules had to be developed to ensure content was consistently coded at the item level (Mayring, 2019). These coding rules are presented in Box 1.

Box 1. Coding rules

**What is an item?** Items are questions which are independently understandable and answerable. For example, alist of 10 individual questions about the same topics would still be considered 10 items, even if they are groupedin a section under that topic, whereas a set of stepwise questions all related to one scenario would be coded as one item

**What needs to be coded per item?** 1) content and 2) learning outcomes. All items need an individual learningoutcome code and content code. One or more learning and content codes can be coded per item

How to code 'foil' items? A foil (distractor) item is an incorrect response option for a question. It might be relatedto RCR, or it might be a non-RCR related item. The content of the RCR related foil options however should still becoded, because the respondent needs to understand the RCR issue to be able to judge that it is not applicable to the specific case

Each measure was initially coded by MvL and checked by NE for adherence to coding rules and alignment with code descriptions. Disagreements led to refinements in the coding rules and code descriptions. Inductive codes such as "academic integrity", "personal character", and codes related to items asking about the socio-demographic characteristics of participants were also developed at this stage. After finalizing the coding scheme and rules, NE recoded all measures. MvdH independently coded three measures, for which there was intercoder agreement of 81.35%

(minimum code overlap at the segment level 90%). Due to the high intercoder agreement, and the involvement of two researchers in the iterative coding of all measures, additional independent coding was not deemed necessary. The final coding scheme is presented in Table 3.

## Quantitative Content Analysis of Measures

Descriptive statistics were used to present the RCR concepts covered by each measure with supporting validation evidence and content related to research practice both individually and as an aggregate.

**Table 3** Final coding scheme

| Parent code | Sub code |
|---|---|
| Research Integrity | Research Environment |
| | Training, Supervision and Mentoring |
| | Research Procedures |
| | Data Practices and Management |
| | Collaborative Working |
| | Publication and Dissemination |
| | Reviewing, Evaluating and Editing |
| | Dealing with violations and allegations |
| Research Ethics | Conflict of interests |
| | Intellectual property |
| | Ethics oversight |
| | Ethical concerns with new bio tech and tech |
| | Societal considerations |
| | Justice considerations |
| | Balancing harms and benefits |
| | Privacy and confidentiality |
| | Consent |
| | Vulnerable and non-competent subjects |
| Personal character | |
| Academic integrity | |
| Learning outcomes | Sensitivity |
| | Knowledge |
| | Attitude |
| | Judgement |
| | Behaviour |
| Independent variables | Age |
| | Gender |
| | Expertise/experience |
| | General further education |
| | Exposure to RCR education |
| | Place of birth/Citizenship/Ethnicity |
| | Language |

## Results

### Included Studies

The systematic scoping review retrieved 1917 results, of which 162 studies met our inclusion criteria and reported the impact of RCR courses on participants (including 2 reviews) or the development or validation of an RCR measure (Fig. 1).

For each included article, the aim, the target group and degree of validation was analysed. This led to distinguishing 112 articles which used measures not described as validated measures (69%) or which were reviews of studies ($n$=2) or inaccessible for assessment ($n$=2). 46 (28%) articles used measures described as validated (Fig. 1). Tools focusing solely on evaluation of plagiarism were also excluded from analysis because they had no specific content but rather identified plagiarised content in assignments. From the 46 articles using or describing measures described as validated, 31 unique measures were identified. The number of studies and measures vary, because several studies include the same measure, and some studies include multiple measures. 18 of these measures were accessible and had supporting validation evidence. The complete list of included articles and measures can be found in Supplemental Material 1.1 'Overview of validated and unvalidated measures' on work sheets 1 and 2 respectively.

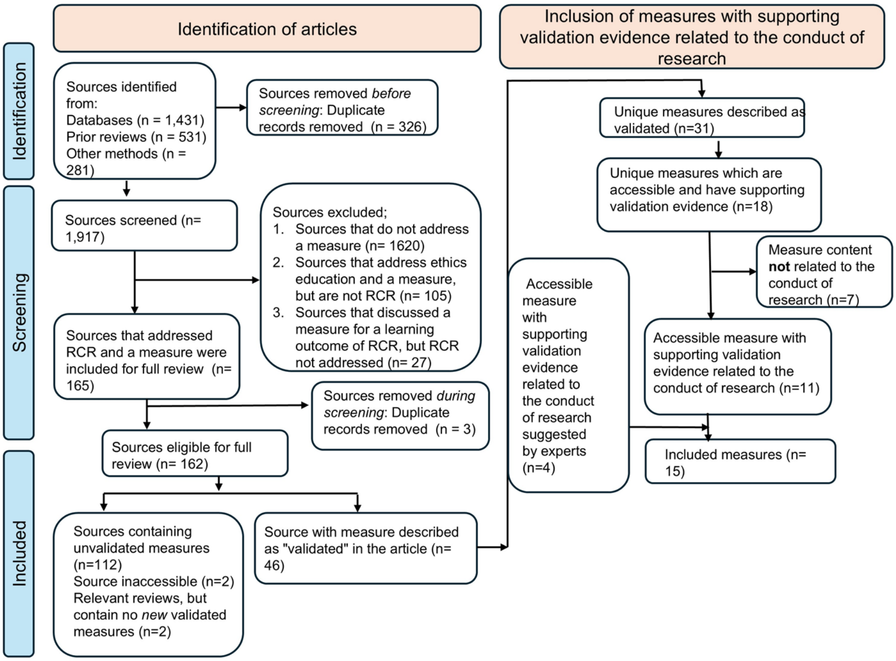


**Fig. 1** PRISMA flow chart

From the 18 measures which were accessible and had validation evidence, a further 7 were excluded, as per our exclusion criteria, because their content was not related to the conduct of research. These included well-known measures such as the Ethical Decision Making (EDM) instrument (Mumford et al., 2008) – used in 13 included studies – and the Defining Issues Test (DIT) (Rest et al., 1992; Bebeau, 2002). An additional 4 measures were included after contacting experts in RCR evaluation, including 1 measure (LAMPS and RTP) that had not been published yet at the time of analysis. A total of 15 measures were identified which were developed specifically for the evaluation of RCR courses and which reported at least some form of reliability or validity evidence (Table 4).

Geographically, these measures 8 of these measures originated from the USA, 1 from Croatia, one from Germany and one jointly from the USA and Taiwan.

**Table 4** Included measures with reported validation evidence

| | Measure | Reference |
|---|---|---|
| 1 | HIT-Res | The How I Think about Research (HIT-Res) test - DuBois et al. (2016a). |
| 2 | RCR Knowledge Test | Responsible Conduct of Research Knowledge Test(RCR Knowledge test); Bioethics Research Center. |
| 3 | VSW | Values in Scientific Work (VSW) English et al. (2018). |
| 4 | Online AI | Online Academic Integrity (Online AI) Chertok et al. (2014). |
| 5 | Rev-RCRRT | Reverse Responsible Conduct of Research Test (rev-RCCRT) Pan (2021). |
| 6 | P2I | Path 2 Integrity (P2I) Zollitsch et al. (2022). |
| 7 | PDR | Professional Decision-Making in Research (PDR) Dubois et al. (2016b). |
| 8 | Human Subjects and Rights in SA | Human Subjects and Rights in Saudi Arabia, Al-Madaney and Fässler (2023). |
| 9 | REKASA | Research Ethics Knowledge and Analytical Skills Assessment tool (REKASA) Taylor et al. (2012). |
| 10 | RCR Survey | Attitude scale regarding Responsible Conduct in Research (RCR Survey) Abd ElHafeez et al. (2022). |
| 11 | RCR Measure | Responsible Conduct of Research Measure (RCR Measure) Wester et al. (2008). |
| 12 | Ethics QNN | Students' Ethical Awareness and Conceptions of Research Ethics Questionnaire (Ethics QNN) Löfström (2012). |
| 13 | LCRE | Assessing the climate for research ethics in labs (LCRE) Solomon et al. (2022). |
| 14 | LAMPS and RTP | Leadership, management, and team practices in research labs (LAMPS and RTP) Antes et al. (2024). |
| 15 | ACP | Perceptions of Authorship Criteria (ACP) Hren et al. (2007). |

### Included Measures with Reported Validation Evidence and the Approaches to Validation

We provide an overview of the measures with reported validation evidence and extent of that evidence in Table 5.

Across the 15 measures with reported validation evidence, the scope and depth of validation varied substantially. Importantly, validation was rarely a single, uniform construct; instead, measures differed markedly in the types of reliability and validity assessments undertaken, the methodological rigor of those assessments, and the populations with whom they were carried out.

**Reliability testing** Twelve measures reported some form of reliability evidence, but in most cases, this was limited to a single indicator—typically internal consistency. Eleven measures relied on Cronbach's alpha to demonstrate internal consistency, often reporting values $\geq 0.80$. While these values suggest homogeneity of items, reliance on alpha alone leaves unaddressed questions about temporal stability or equivalence across forms. Only two measures provided more advanced reliability assessments: the Professional Decision-Making in Research (PDR) measure additionally reported parallel-forms reliability, and the Human Subjects' Rights and Research Ethics (SA) instrument reported split-half reliability. Notably, no measure reported test–retest reliability, meaning that stability over time—a key psychometric criterion—remains unknown for all measures. Two measures (P2I and ACP) did not report any reliability indices, limiting confidence in the reproducibility of their scores.

**Validity testing** Validity evidence was even more heterogeneous. Nine measures provided some form of validity evidence, but the nature of this evidence differed greatly. Most commonly, authors reported content or face validity established through expert review, though the procedures for expert review were often minimally described and rarely quantified. A smaller subset of measures supplemented content validation with cognitive interviews (e.g., the RCR Knowledge Test, VSW), but details about how interview data informed item refinement were usually absent. Only two measures—Ethics QNN and LCRE—reported convergent construct validity, demonstrating expected correlations with existing validated scales. No measure provided criterion validity evidence, and none reported factor-analytic examinations of dimensional structure. This absence is notable given that many measures purport to assess multi-component constructs such as ethical awareness, judgment, or decision-making.

**Variation across target populations** The validation populations also differed considerably. Some measures were developed with practicing researchers or NIH-funded investigators, others with postdoctoral researchers, graduate students, clinical researchers, or undergraduates. No measure was validated across multiple cultural contexts, and most were normed on U.S. samples. As a result, cross-study comparability is limited, and the external validity of several measures for diverse international RCR contexts remains uncertain.

**Table 5** Overview of the measures with reported validation evidence and the extent of this evidence

| | Measure | Reliability | Validity | Validation population |
|---|---|---|---|---|
| 1 | HIT-Res | ICR: CA (0.92) | -- | Researchers and trainees |
| 2 | RCR Knowledge Test* | *[Reliability testing with PhD students in progress at the time of writing (Pers. Com]* | CIs | RCR experts |
| 3 | VSW | ICR: CA (greater than 0.70) | CIs | Experts (researchers) |
| 4 | Online AI | ICR: CA 0.93 | -- | Science students |
| 5 | Rev-RCRRT | ICR: KR-20 (0.77 (for the Chinese version) and 0.65 (for the English version)) | --- | Graduate students working on a master's degree or Ph.D. in Taiwan or the U.S were targeted for participation |
| 6 | P2I | -- | CentV | Master students and (early career) researchers |
| 7 | PDR | ICR: CA (0.84) and PFC (0.70) | CentV & CuctV | Group of NIH-funded researchers and research trainees who were diverse in terms of age, years of experience, types of research, and race |
| 8 | Human Subjects and Rights in SA | ICR: CA (0.77 for researchers' attitudes), SHR (0.755 for knowledge items), | FV & CentV: S-CVIs (greater than 0.78) | Clinical researchers |
| 9 | REKASA | ICR: CA (0.837 and two shortened versions of the measure were 0.72 and 0.655) | FV & CentV | Face validity and Content validity ("Three experienced bioethics faculty, two international research ethics scholars, and one bioethics PhD student reviewed the 74 draft items and provided feedback. |
| 10 | RCR Survey | ICR: CA (0.975 for the 16 items in the "attitudes toward the acceptability of RCR practices" scale and 0.754 for the four items in the "general attitudes toward research misconduct") | -- | (1) academic faculty; (2) individuals with master's and PhD degrees and postdoctoral students; and (3) senior undergraduate students and individuals working in research positions |
| 11 | RCR Measure | ICR: CA (ranged between 0.84–0.89) | -- | Expert panel of five investigators with knowledge and expertise on RCR |
| 12 | Ethics QNN *(several measures, Ethics QNN in analysis)* | ICR: CA (Ethical climate measure was 0.81 and 0.94, the organizational and professional socialization measures was 0.74 and 0.84, empathy measure was 0.78 and 0.8) | CuctV (Convergent) | Undergraduate and graduate (masters) students |
| 13 | LCRE | ICR: CA (0.95) | CuctV (Convergent) | Postdocs |
| 14 | LAMPS and RTP | ICR: CA (for RTP 0.82) | -- | RCR experts, postdoctoral researchers |
| 15 | ACP | -- | CentV | Graduate medical students, doctors, medical teachers experienced in scientific research |

*ICR* Internal Consistency Reliability, *CA* Cronbach's Alpha, *PFC* Parallel Forms Correlation, *SHR* Split-Half Reliability, *KR-20* Kuder–Richardson Formulas, *S-CVIs* The Global Content Validity Indices, *FV* Face Validity, *CentV* Content Validity, *CuctV* Construct Validity, *CIs* Cognitive Interviews

Together, these patterns indicate that while all 15 measures meet minimal criteria to be described as "validated," they do so to very different extents. In many cases, validation evidence is partial, fragmented, or narrowly focused on internal consistency. These differences underscore the importance of distinguishing between levels of validation evidence rather than applying a binary categorization.

## Content Analysis of Validated RCR Measures

The full coding scheme is presented in Supplemental material 2.1 and an overview of the full content analysis is provided in Supplemental material 2.2. Specific items (or questions) were categorized as being related to RI, RE, academic integrity or personal character. Items addressing RI topics, as we defined them, were included in all 15 measures. Ten measures (two of which were primarily on research ethics) also included items on research ethics, three measures included items on personal character, and one measure included items on academic integrity (Fig. 2). For coding frequencies per code, see Supplemental material 2.3.

### Research Integrity

Although RI was represented in all measures, the specific RI topics which were focused on varied greatly (Fig. 3).

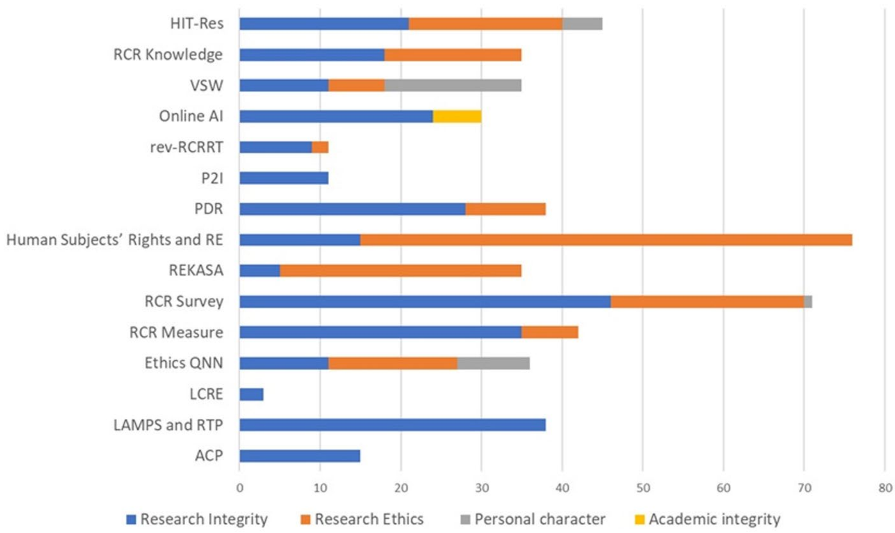


**Fig. 2** Overview of the items which covered research integrity, research ethics, personal character or academic integrity

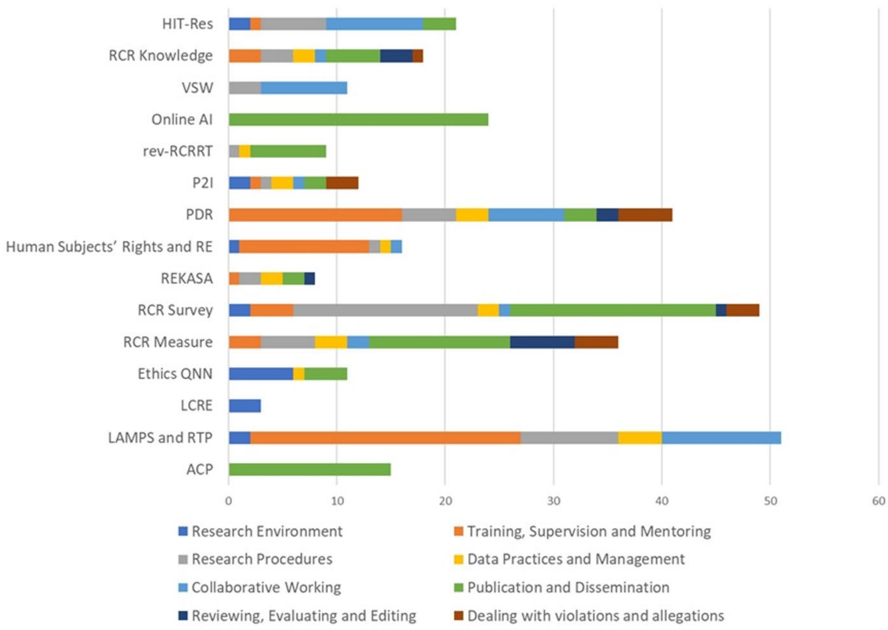


**Fig. 3** Aspects of RI covered by items

1. Research environment

Seven measures included items on the research environment (Fig. 3). The majority of items asked respondents about their attitudes on the explicit and implicit norms and standards related to how research is conducted within departments (such as research team practices) and institutions and attitudes on research team practices. Explicit items asked about institutional guidelines and policies, whereas implicit items asked respondents if they know how 'research is done here' (Supplemental material 2.2). For example this attitudinal item from the Ethics QNN asks respondents about their agreement with the statement: "I know what's considered (in)appropriate behavior in my department". Other topics, which reflect more recent foci of RI guidance such as equity, diversity and inclusion (EDI), social safety and whistleblower protections, were not addressed by any item in any measure.

2. Training, supervision and mentoring

Nine measures included items which touched on training, supervision or mentoring (Fig. 3). Behaviour and attitude items were frequent (Supplemental material 2.2) due to the questions in the LAMPS and RTP measure asking respondents about the supervision, mentoring and training practices of their principal investigator. Judgement items were also relatively frequent with respondents asked to choose a course of action related to hypothetical scenarios related to poor supervision. For example, this judgement item from the REKASA measure asks to choose a scenario most ‘appropriate with respect to your postdoc's prospective role as a co-author in a publication that might emerge from the project?’

3. Research procedures

Eleven measures included items on research procedures (Fig. 3). Items were mostly about behaviour or attitudes, (Supplemental material 2.2) with questions on the prevalence and perceived acceptability of poor research practices (e.g. dropping “outliers” without mentioning it, inadequate/selective reporting, or fabrication), or good practices (e.g. to keep accurate and detailed records, alignment between research questions and methods, collaborative analyses and interpretations, or awareness of and attention to potential biases). Research protocols featured in a number of items, for example the LAMPS and RTP measure asks respondents to rate the frequency in which the PI “Tells lab members to contact them when concerns arise in project protocols”. Public pre-registration, in contrast, was not touched on in any items. No items addressed the use of research funds, or more recent concerns such as the use of AI in research methods.

4. Data management practices

Ten measures included items on Data Management Practices (DMPs) (Fig. 3). The content of these items was mostly related to the behaviour of the respondent, their colleagues or PI in relation to the correct storage of data, or the respondent’s knowledge or judgement in relation to DMPs (Supplemental material 2.2). Other topics in data management included open (FAIR) data and when to share data with other researchers.

5. Collaborative working

Nine measures included items on collaborative working (Fig. 3). These were most frequently attitudinal, on interpersonal relationships within teams and joint responsibility/accountability for the quality and integrity of research, or behavioural, on collaborations within teams and departments and with other academic institutes. Only two measures include an item on collaborations with industry partners, and one on collaboration with a public partner (Supplemental material 2.2). Questionable research practices such as hampering work of others, bullying and harassment were not represented in any of the items.

6. Publication and dissemination

Eleven measures included items on publication and dissemination. This was the most frequent RI item topic across all measures, with two measures entirely devoted to items on publication and dissemination(the Online AI and the ACP measure) (Fig. 3). All of the Online AI measure items asked respondents to rate the permissibility of specific practices, therefore sensitivity was the most frequent item learning objective overall (Supplemental material 2.2). Attitudinal items were also very frequent, for example all items from the ACP measure represented attitudinal items asking respondents about the importance of specific authorship criteria. Across all measures, the most frequent topics of items were appropriate and inappropriate authorship contributions and plagiarism of text or ideas. For example, attitudinal items asking about deciding in a hypothetical situation about authorship considerations. Other less frequent topics included salami slicing and duplicate publications, not publishing all results, and appropriate and inappropriate citation practices. Topics that featured in just one question included open access publishing, submitting without all authors' approval, and retraction/correction. No items addressed predatory journals, conference and paper mills, or the use of AI tools for writing manuscripts.

7. Reviewing, evaluating and editing

Five measures included items on reviewing, evaluating and editing (Fig. 3). Items were most frequently attitudinal or judgement items on topics such as potential conflicts of interest in peer review, either due to favouring colleagues or hampering competitors. A few items also addressed the importance of, and appropriate processes for, peer review.

8. Dealing with violations and allegations

Five measures included items on dealing with violations of research integrity (Fig. 3). These items generally assessed respondents' judgement or attitude, asking what a researcher should do if they became aware of the misconduct of a colleague or supervisee, if they were the victim of research misconduct, or if they were accused of misconduct themselves. For example, attitudinal items might ask respondents how likely they are to "confront a colleague that you know has engaged in some form of research misconduct" (RCR Measure),whereas judgement items might ask respondents to choose a course of action in a hypothetical case relating to accusations, e.g. this item from the RCR Survey: "You have received a manuscript for review from a journal editor. You believe the paper is very good and realize that it contains a new insight that is relevant to the content of a paper you are currently writing. Which of the following actions is most appropriate?" [four options]. Only two items were related to institutional procedures or the consequences of misconduct.

### Research Ethics

In contrast to research integrity, RE topics were not represented in all of the measures. Ten measuresincluded items on RE topics, with the specific topics again varying greatly (Fig. 4).

1. Conflicts of Interest

   Five measures addressed conflicts of interest (Fig. 4). Items were most frequently attitudinal or behavioural, on declaring conflicts of interest and on inappropriate interference in a study by the study sponsors.

2. Ethics oversight

   Eight measures included items on ethics oversight (Fig. 4). However, it was a frequent topic across the measures which included RE content and was the second most frequent RE topic in total. These were frequently knowledge items about definitions, principles, guidance and regulatory frameworks related to RE as well as attitudinal items related to the role and specific procedures of

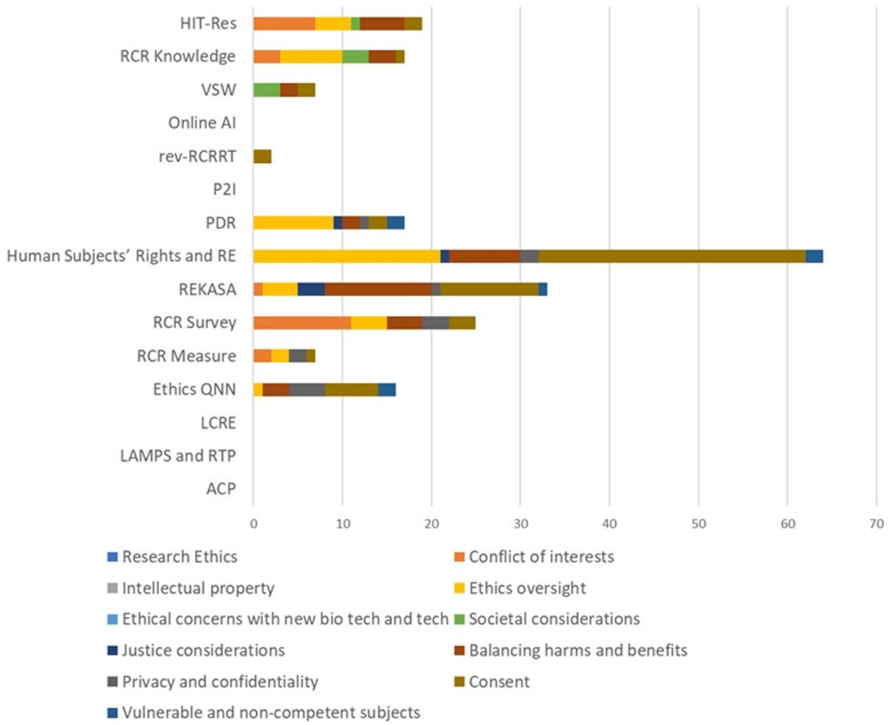


**Fig. 4** Aspects of RE covered by items in the measures

RE committees, e.g. an attitudinal item from the HIT-Res measure stating IRBs focus so much on rules, it could become impossible to do research.

3. Societal considerations

Only three measures contained any items on societal considerations in research (Fig. 4). The content of these items concerned attitudes towards the involvement of community members in setting research agendas or the need for research to have a social impact, or sensitivity to potential unintended consequences of research: “I really do not need community members or a council to tell me whether my research has social significance.” (Attitudinal item from the Hit-Res). Items related to judgment came exclusively from the RCR Knowledge measure, which asked about appropriate approaches to societal concerns in research practice; asking to choose the right option of four in a case of scientist discovering findings of research could result in serious unintended and potentially harmful consequences No measures included knowledge or behaviour items on societal considerations.

4. Justice considerations

Only three measures included items on justice considerations in research (Fig. 4). These were knowledge or judgement items related to the recruitment and retention of diverse participants for research, for example an item in the REKASA, Asking about providing ethics arguments to advocate for including more research with women. No items in any of the included measures addressed the sharing of research benefits more broadly or fairness in global north and global south research collaborations.

5. Balancing harms and benefits

Eight measures included items on balancing harms and benefits of participation in research (Fig. 4). These were most frequently knowledge or judgement items. Knowledge items assessed respondents' knowledge of specific terms, principles or risks related to research. Judgement items asked respondents what they should do in relation to potential harms associated with hypothetical studies. In terms of specific content, attitudes and responsibilities related to animal welfare were fairly frequent topics. Two items also specifically linked adequacy of study design for answering the research question to the ethical acceptability of a study, e.g. a judgement item from REKASA asking about why a study’s design is also relevant to the ethical acceptability of a study.

6. Privacy and confidentiality

Six measures included items on privacy and confidentiality (Fig. 4). These were frequently knowledge items related to identifying correct procedures for maintaining confidentiality, what constitutes a data breach, how to react to a breach of confidentiality, or sensitivity questions assessing if respondents can identify anonymity and confidentiality issues in hypothetical scenarios. An example knowledge item from Human Subjects Rights and RE asks if: “The researchers should notify the participant’s primary care physician of his/her participation in a research study” . No items dealt with more complex issues, such as when breaching confidentiality might be ethically justified in research.

7. Consent

Ten measures included items on consent to research participation, making it the most frequent RE item topic (Fig. 4). These were most frequently knowledge items about the elements that need to be included in consent forms. Attitudinal items focused on consent and researchers’ responsibilities during the consent process. Sensitivity items asked respondents to identify consent/voluntariness as potential ethical issues, and judgment items asked about the best course of action in a hypothetical case related to consent. Minor topics included questions about when consent is needed in relation to research vs. data collected for clinical care and what to do when informed consent forms might be missing.

8. Vulnerable and non-competent participants

Four measures addressed vulnerable and non-competent research participants (Fig. 4). Items targeted knowledge, sensitivity, or judgement. The content of these items related to vulnerable and non-competent research participants; six were related to children and just one was related to incompetent adults. All items were related to consent or parental assent.

9. Academic integrity

One measure (Online AI) included items on academic integrity (focused on student behaviour), for university students e.g.: “Student accesses internet sites to seek information while taking a computerized exam without instructor permission.” (Sensitivity item -Online AI). All items targeted respondents' sensitivity since the respondents were asked to rate the permissibility of the practices.

## Additional Outcomes

Other item topics that fell outside of RI and RE frameworks were related to the personal character of the respondent: general dispositions, personal (work) motivations, and attitudes towards wrongdoing. Some of these questions were also directly related to the research environment, whereas others were more general. All were attitudinal in nature. Two ethical topics which were included in the initial deductive coding were not addressed in any measures: intellectual property and ethical concerns with new biotech and technology in any items.

### Learning Outcomes

Learning outcomes measured by items varied across the included measures. Some measures only evaluated respondents' attitudes (Hit-Res, VSW, and LCRE), their ability to identify ethical issues via sensitivity items(Online AI), or their ability to choose the right course of action via judgement items (P2I and PDR). Only four measures included items that focused on respondents' knowledge (the Human subjects' rights and research ethics, the RCR Knowledge measure, the REKASA, and the ACP). The RCR Knowledge measure and the REKASA, as well as the RCR Survey, were the only measures to test three or more different learning outcomes (Fig. 5). 8 out of 15 measures only test a single type of learning outcome.

1. Knowledge

    Only four of the measures tested respondents' knowledge. However, knowledge items were still the second most frequent item type across all measures (109 items) - mainly due to the large number of knowledge items in the Human Subjects' Rights and RE measure. Knowledge questions included asking respondents to identify whether statements were true or false, or multiple-choice items.

    The line between knowledge and attitudes was not always clear, however. For example, the ACP asked about attitudes regarding authorship contributions, but these were then compared to underlying ideas about a more objective measure of authorship. As a final question, respondents were asked whether they had knowledge of the ICMJE authorship criteria. Other items, which

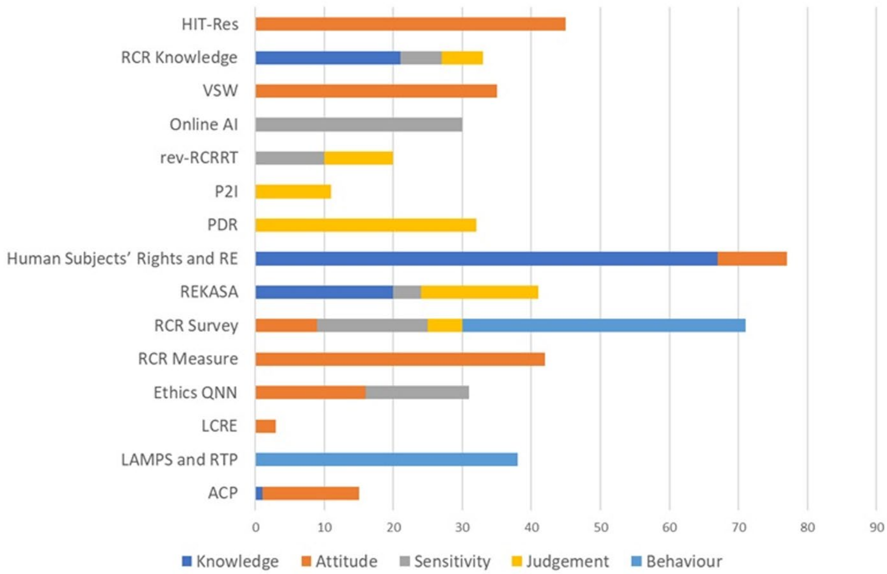


**Fig. 5** Learning outcomes across measures

were borderline knowledge items, included questions asking participants if they knew of, or had read certain guidelines. These do not evaluate actual knowledge but rather infer knowledge based on the awareness or reading of a particular guidance.

2. Attitude

Eight measures included items to test respondents' attitudes to RCR issues. These items consisted of statements to which respondents chose responses on a Likert Scale, e.g.: "Score how similar you are to the person in the following item: She believes that researchers should disclose all foreseeable risks to participants." (VSW).

3. Sensitivity

Six measures included items which aimed to evaluate respondents' ability to recognize an ethical problem. The sensitivity items assess respondents' abilities to identify actions or practices as ethically problematic by asking them to rate the acceptability of certain actions or practices on a Likert scale, or to identify an ethical issue in a hypothetical scenario via open response or multiple choice. In checking whether people can recognize ethical issues, sensitivity measures assess more than only knowledge or attitudes. This makes them more challenging higher-order learning outcomes. For example in RCR Survey: "Please rate the extent[..] to which you think the below behaviour is acceptable."

4. Judgment

Six measures included judgement items that evaluate respondents' ability to choose the right course of action. Judgement items were almost exclusively assessed via hypothetical scenarios with answer options relating to different courses of action. One measure (REKASA) also asked open responses to assess ethical reasoning in relation to hypothetical scenarios. For example: "[in relation to a described hypothetical scenario] Name one action the investigator(s) can take[..] to address the moral concern or challenge you identified above." Two measures (P2I and Pann-RCT) additionally incorporated a follow-up item asking respondents how sure they were of their response.

5. Behaviour

Two measures evaluated RCR behaviours. Respondent were asked about behaviours they had participated in themselves almost as frequently as behaviours they had witnessed, e.g. in LAMPS: "Please rate how frequently you observe the PI doing each of the behaviors [such as]. Interacts with lab members in a respectful manner."

## Discussion

Our review revealed that the majority of RCR course evaluations in the published literature have not used measures with reported validation evidence – instead often using questionnaires developed for individual studies which were, presumably, constructed by course developers. From the literature search, 46 sources contained measures described as validated, and most of the individual measures (20 of 31) did not containany specific RCR items or the measure or validation evidence were inaccessible. As a result, our findings do indicate a general lack of evidence regarding the comparability of RCR education evaluations, and there liability and validity of the measures used. Indeed the results indicate that the majority of RCR course evaluations in the published literature are at the level of evidence which might be appropriate for making iterative improvements of individual courses, an approach championed by approaches such as the‘scholarship of teaching and learning’ (SoTL; see e.g. Hubbal & Clarke, 2010a; Felten, 2013; Cox et al., 2023),but they do not reach the level of evidence to make any generalisable claims. Ultimately, the results indicate uncertainty related to what we know about the effectiveness of RCR education. This corroborates previous studies, which found that most studies do ‘not use standardized or validated outcome measures’ (Katsarov et al., 2021) and ‘overall there is very low-quality evidence that various methods of training in RI had some effects on participants attitudes to ethical issues but minimal (or short-lived) effects on their knowledge’(Marusic et al., 2016).

Although only a minority of studies used measures with reported validation evidence that were developed to evaluate RCR education, 15 such measures were identified. Potentially, these measures could be used more frequently to compare outcomes of RCR education. These 15 measures cover a wide range oftopics varying from supervision and mentoring, through day-to-day research ethics and integrity practices, to dealing with research misconduct. All measures contained items covering topics which we categorised as pertaining to ‘research integrity’, whereas a smaller selection of measures contained items covering topics which we categorised as pertaining to ‘research ethics’. This was interesting considering the longer history of research ethics as a subject of education (Kalichman & Plemmons, 2007; van den Hoven & Theunissen, 2021; van den Hoven et al., 2023).

Very broadly, research integrity items were fairly equally spread over a diversity of topics, variously assessing respondents’ knowledge of professional responsibilities, attitudes to implicit and explicit rules, sensitivity to good and detrimental research practices, judgement related to dilemmas, and frequency of one’s own and others practice which support or violate research integrity. Some aspects of research integrity were (almost) completely absent, these generally related to the inclusion and social safety(EDI, whistleblower protections, bullying and harassment) or open science practices. Also, more recent threats to research integrity, such as the use of AI in research methods and reporting, and knowledge of predatory journals were not evaluated in measures. There are therefore topics missing; since some more established topics are not found in validated measures, and new topics are emerging in the RCR field that have not been included in measures yet. We need new measures and items in order to also evaluate these missing topics and themes.

In relation to research ethics, there was a clear dominance of items related to ethics oversight and consent procedures, followed by balancing harms and benefits. The dominance of these topics reflects a very traditional understanding of research ethics (Hunt & Godard, 2013). Broader conceptions of research ethics, for example, related to societal and justice considerations, were infrequently represented across measures. The consideration of justice in research ethics relatively recent (Pieper & Thomson, 2013) and it was later added to the Declaration of Helsinki (World Medical Association et al., 2013). Knowledge of ethical issues related to new technologies (e.g.use of AI) was also not assessed, but is an increasingly important topic in the ethics of biomedical research (Stahl, 2021).

The overview provided by this study of the RCR topics covered by the 15 RCR measures with supporting validation evidence is an important resource for the RCR community to align group level evaluation closely to course content and for measure developers to see which topics are currently neglected. Of the 15, however, only 6 reported evidence for both reliability and validity (at the time of writing), indicating that RCR courses may have been evaluated by measures, 9 in total, that are either not reliable or not valid in relation to testing what the measure claims to test. There is a possibility then that the evidence produced is not meaningful, or may be inaccurate, possibly impacting generalizability of results and misrepresenting actual impact of RCR education. Furthermore, Cronbach's alpha reliability scores were over 0.9. While this was generally viewed as positive by the respective authors, it actually indicates that some of the items are too similar (Streiner, 2003). In these cases, highly correlated items could be deleted without a loss of information, which would help to increase the efficiency of the respective measures.

To characterise differences across measures more precisely, we can distinguish between (a)minimal evidence (e.g., internal consistency or basic expert review), (b) moderate evidence (e.g., multiple reliability checks or systematic content validation), and (c) stronger evidence (e.g., convergent construct validation or multiple complementary reliability assessments). Only a small subset of measures—PDR, Ethics QNN, and LCRE—meet criteria in the latter category. Even these measures, however, lack certain forms of validation (e.g., test–retest reliability, criterion validity, or cross-cultural examination), which limits the strength of claims about their overall psychometric robustness. More generally, the evidence base is constrained by limited reporting. Many studies provided Cronbach's alpha as the sole reliability indicator, and validity evidence was often restricted to expert judgment without accompanying details about review procedures or decision rules.

The field would benefit from more systematic and multi-method validation approaches, including longitudinal stability testing, factor-analytic assessment of dimensionality, and tests of construct validity using external behavioural or decision-making measures. Thus, while several instruments demonstrate stronger validation characteristics relative to others, the available data do not allow for fine-grained comparative judgments about which measures are "best supported" in an absolute sense. We therefore temper our conclusions accordingly and recommend that future research replicate validation work across different research populations, disciplinary settings, and cultural contexts to establish broader generalizability.

This being said, some measures were identified with comparatively stronger validation evidence covering a range of learning outcomes, which can aid further research into the promotion of quality in RCR training. Hopefully awareness of these measures will allow course developers to make informed decisions regarding which measures are appropriate in evaluating their courses.

A broader consideration, however, is that greater use of established measures is not always pedagogically appropriate (Diaz-Martinez, 2019). While instruments with reported validation evidence offer clear advantages—such as improved reliability, greater transparency about measurement quality, and the possibility of comparing outcomes across interventions—RCR courses differ substantially in their intended learning outcomes, pedagogical priorities, disciplinary contexts, and institutional cultures. These differences mean that existing measures may not fully align with what individual courses aim to achieve. In many cases, locally developed instruments may be better suited to capturing course-specific competencies or addressing context-sensitive issues that standardised measures overlook (Diaz-Martinez, 2019). Cultural variation is another factor: because none of the measures were validated across multiple cultural settings, their applicability may be limited outside the populations in which they were developed. Thus, while validated instruments can strengthen the evidence base for RCR education and facilitate cross-study comparability, their use should complement—not replace—tailored assessments that reflect the distinctive aims and contexts of individual courses.

Evaluation of RCR education might require using several instruments for assessment of different goals, a multi-level approach might be suited (Mumford et al., 2015; Steele et al., 2016). E.g. the LAMPS could be used to evaluate leadership, while at the same time the REKASA is used to evaluate research ethics skills. Used in combination, some of the measures could enable deeper insights into the knowledge and attitudes that drive ethical sensitivity, judgement, and behaviour.

It is worth noting that measures were often used in pre-and-post-test designs, measuring relatively short-term outcomes only, before and after taking a training/course. Evaluating longer-term outcomes (for example 3 months, 1 or 2 years) is important for understanding how RCR courses impact RCR educational learning outcomes in the long term. Looking at changes in behaviours connects to broader discussions on evaluating impact of ethics training (Steele et al., 2016). Furthermore, ideally, we would evaluate changed RCR behaviours, in real-world settings, understanding how participants in RCR education apply learned knowledge and skills to their everyday research practice.

A notable limitation was the in accessibility of reportedly validated measures. Considering the commitment in RCR circles to open science approaches, it is regrettable that some tools are unavailable for educators and researchers.

Our study provides the first systematic overview of how the impact of RCR education is evaluated. Strengths include the systematic approach to identifying studies and the critical assessment of associated measures. Based on the results of this review on RCR education measures, we developed a measures checklist, in order to help educators and researchers in the field of RCR education evaluation, recommending how authors can report on the use of RCR measures in the future. This can help in aligning future research in this area (Table 6).

**Table 6** Measures checklist

| | | |
|---|---|---|
| Q1 | Name of assessment measure used | |
| Q2 | Why is the measure appropriate for your use? e.g., does the measure align with the course learning outcomes. | |
| Q3 | Has evidence on the validation of the measure been reported? | Yes / No |
| Q4 | If Yes to Q3, was reliability tested? | Yes / No |
| Q5 | If Yes to Q3, was validity tested? | Yes / No |
| Q6 | If Yes to Q4, please state whether the tests where quantitative, qualitative, or both. | |
| Q7 | If Yes to Q5, please state whether the tests where quantitative, qualitative, or both. | |
| Q8 | If Yes to Q4, please provide further details of the type of test(s) used and their results. | |
| Q9 | If Yes to Q5, please provide further details of the type of test(s) used and their results. | |
| Q10 | If Yes to Q3, what was the target population used for the validation? | |
| Q11 | Is the target population used in the validation the same as your current target group being assessed? | Yes / No |
| Q12 | If No to Q11, why is this measure appropriate for your current target group? | |

## Conclusion

This review demonstrates that RCR education is, mostly, currently evaluated with unvalidated measures.Some good practices were identified, and the comprehensive overview we provide of measures withreported validation evidence addressing RCR concepts, distinguished in RI and RE topics, can offer guidance for educators and researchers to select or further develop measures aligned with core RCR competenciesand targeted learning outcomes.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s10648-026-10187-8.

**Acknowledgements** The authors wish to thank like to Joshua Gualtieri for assistance in identifying and sourcing measures for the review and the anonymous reviewers whose suggestions improved the article.

**Author Contributions** All authors made substantial contributions to conception, analysis of the data, and drafting the work, and final approval of the version to be published. All authors agreed to be accountable for all aspects of the work. In our methodology is described who contributed specifically to which part. Miriam van Loon: conceptualization, selection of inclusion measures, analysis of validated measures, content analysis, methodology, writing – original draft preparation and review and editing final version. Daniel Crean: conceptualization, methodology, selection of inclusion measures, writing and review and editing final version. Mariette van den Hoven: conceptualization, content analysis, original draft prepara-

tion and review and editing final version. Johannes Katsarov: conceptualization, methodology, search strategy and selection of measures, validation of measures, writing and review final version. Natalie Evans: conceptualization, content analysis, methodology, writing – original draft preparation and review and editing final version.

**Funding** This study is partly funded by the IRECS project, Horizon Europe grant number: 101058587.

## Declarations

**Ethical Standards** This study complies to ethical standards with regards to rigorous review methodology. No human subjects were involved in this study, since this study is a systematic review of previously published literature, it did not require ethical approval or informed consent. The review protocol was published on OSF (Open Science Framework) OSF protocol: https://osf.io/z7cjt/. We adhered to the international reporting standard PRISMA, a PRISMA flow diagram is provided in the manuscript. Search strategies are described in full, inclusion and exclusion criteria are reported.

**Conflict of Interest** All authors declare no conflicts of interests.



## References

64th WMA World Medical Association, General, & Assembly (2013). *Declaration of Helsinki: Ethical principles for medical research involving human subjects*. Fortaleza, Brazil. https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-sub

Abd ElHafeez, S., Salem, M., & Silverman, H. J. (2022). Reliability and validation of an attitude scale regarding responsible conduct in research. *PLOS ONE*, *17*(3), e0265392. https://doi.org/10.1371/journal.pone.0265392

Abdi, S., Pizzolato, D., Nemery, B., & Dierickx, K. (2021). Educating PhD students in research integrity in Europe. *Science and Engineering Ethics*, *27*(1), 5. https://doi.org/10.1007/s11948-021-00290-0

Al-Madaney, M. M., & Fässler, M. (2023). Development and validation of a tool to assess researchers' knowledge of human subjects' rights and their attitudes toward research ethics education in Saudi Arabia. *BMC Medical Ethics*, *24*(1), 94. https://doi.org/10.1186/s12910-023-00968-z

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

Antes, A. L., English, T., Solomon, E. D., Wroblewski, M., McIntosh, T., Stenmark, C. K., & DuBois, J. M. (2024). Leadership, management, and team practices in research labs: Development and validation of two new measures. *Accountability in Research*, *32*(8), 1496–1523. https://doi.org/10.1080/08989621.2024.2412772

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

Antes, A. L., Wang, X., Mumford, M. D., Brown, R. P., Connelly, S., & Devenport, L. D. (2010). Evaluating the effects that existing instruction on responsible conduct of research has on ethical decision making. *Academic Medicine*, *85*(3), 519–526. https://doi.org/10.1097/ACM.0b013e3181cd1cc5

Bebeau, M. J. (2002). The Defining Issues Test and Four Component Model: Contributions to professional education. *Journal of Moral Education*, *31*(3), 271–295. https://doi.org/10.1080/0305724022000008115

Biggs, J. (2003). Aligning teaching and assessing to course objectives. *Teaching and Learning in Higher Education: New Trends and Innovations*, *2*(4), 13–17.

Cahapay, M. (2021). Kirkpatrick model: Its limitations as used in higher education evaluation. *International Journal of Assessment Tools in Education*, *8*(1), 135–144. https://doi.org/10.21449/ijate.856143

Chertok, I. R. A., Barnes, E. R., & Gilleland, D. (2014). Academic integrity in the online learning environment for health sciences students. *Nurse Education Today*, *34*(10), 1324–1329. https://doi.org/10.1016/j.nedt.2013.06.002

Christen, M., Ineichen, C., & Tanner, C. (2014). How moral are the principles of biomedical ethics?–a cross-domain evaluation of the common morality hypothesis. *BMC Medical Ethics*, *15*(1), 47.

Cox, W., Wettergreen, S. A., Savage, A., & Brock, T. (2023). Research and scholarly methods: The scholarship of teaching and learning. *Journal of the American College of Clinical Pharmacy*, *6*(9), 1071–1077. https://doi.org/10.1002/jac5.1845

Crean, D., Gordijn, B., & Kearns, A. J. (2024). Impact and assessment of research integrity teaching: A systematic literature review. *Science and Engineering Ethics*, *30*(4), 30. https://doi.org/10.1007/s11948-024-00493-1

Diaz-Martinez, L. A. (2019). Recommendations for effective integration of ethics and responsible conduct of research (E/RCR) education into course-based undergraduate research experiences: A meeting report. *CBE—Life Sciences Education*. https://doi.org/10.1187/cbe.18-10-0203

DuBois, J. M., Chibnall, J. T., & Gibbs, J. (2016a). Compliance disengagement in research: Development and validation of a new measure. *Science and Engineering Ethics*, *22*(4), 965–988. https://doi.org/10.1007/s11948-015-9681-x

DuBois, J. M., Chibnall, J. T., Tait, R. C., Wal, V., Baldwin, J. S., Antes, K. A., A. L., & Mumford, M. D. (2016b). Professional decision-making in research (PDR): The validity of a new measure. *Science and Engineering Ethics*, *22*(2), 391–416. https://doi.org/10.1007/s11948-015-9667-8

English, T., Antes, A. L., Baldwin, K. A., & DuBois, J. M. (2018). Development and preliminary validation of a new measure of values in scientific work. *Science and Engineering Ethics*, *24*(2), 393–418. https://doi.org/10.1007/s11948-017-9896-0

ESF-ALLEA. (2023). *The European code of conduct for research integrity*. European Science Foundation and All European Academies. https://www.allea.org/code-of-conduct/

Evans, N., Schmolmueller, A., Stolper, M., Inguaggiato, G., Hooghiemstra, A., Tokalic, R., & Widdershoven, G. (2023). VIRT2UE: A European train-the-trainer programme for teaching research integrity. *Research Ethics*. https://doi.org/10.1177/174701612311612. Advance online publication.

Felten, P. (2013). Principles of good practice in SoTL. *Teaching and Learning Inquiry: The ISSOTL Journal*, *1*(1), 121–125. https://doi.org/10.2979/teachlearninqu.1.1.121

Fenn, J., Tan, C. S., & George, S. (2020). Development, validation and translation of psychological tests. *BJPsych Advances*, *26*(5), 306–315. https://doi.org/10.1192/bja.2020.33

Fuerholzer, K., Schochow, M., Peter, R., & Steger, F. (2020). Medical students' acquaintance with core concepts, institutions and guidelines on good scientific practice: a pre-and post-questionnaire survey. *Science and Engineering Ethics*, *26*(3), 1827–1845.

Goddiksen, M. P., & Gjerris, M. (2022). Teaching phronesis in a research integrity course. *FACETS*, *7*, 139–152. https://doi.org/10.1139/facets-2021-0064

Hren, D., Sambunjak, D., Ivaniš, A., Marušić, M., & Marušić, A. (2007). Perceptions of authorship criteria: Effects of student instruction and scientific experience. *Journal of Medical Ethics*, *33*(7), 428–432. https://doi.org/10.1136/jme.2006.017160

Hubball, H., & Clarke, A. (2010). Diverse methodological approaches and considerations for SoTL in higher education. *Canadian Journal for the Scholarship of Teaching and Learning*, *1*(1), 2. https://doi.org/10.5206/cjsotl-rcacea.2010.1.2

Hughes, J., Hunter, D., Sheehan, M., Wilkinson, S., & Wrigley, A. (2010). *European textbook on ethics in research*. Publications Office of the European Union.

Hunt, M. R., & Godard, B. (2013). Beyond procedural ethics: Foregrounding questions of justice in global health research ethics training for students. *Global Public Health*, *8*(6), 713–724. https://doi.org/10.1080/17441692.2013.796400

Kalichman, M. (2013). A brief history of RCR education. *Accountability in Research*, *20*(5–6), 380–394. https://doi.org/10.1080/08989621.2013.822260

Kalichman, M. W., & Plemmons, D. K. (2007). Reported goals for responsible conduct of research courses. *Academic Medicine*, *82*(9), 846–852.

Katsarov, J., Andorno, R., Krom, A., & van den Hoven, M. (2021). Effective strategies for research integrity training—a meta-analysis. *Educational Psychology Review*, *33*(1), 1–21. https://doi.org/10.1007/s10648-021-09630-9

Krom, A., & van den Hoven, M. (2022). A quality checklist for responsible conduct of research (RCR) education: A proposal to complement the predictive modeling tool. *Accountability in Research*, *29*(1), 26–44.

Löfström, E. (2012). Students' ethical awareness and conceptions of research ethics. *Ethics & Behavior*, *22*(5), 349–361. https://doi.org/10.1080/10508422.2012.679136

Maesschalck, J., & De Schrijver, A. (2015). Ethics in public policy and management, chapter: Researching and improving the effectivenessof ethics training, pp. 197–211.

Marusic, A., Wager, E., Utrobicic, A., Rothstein, H. R., & Sambunjak, D. (2016). Interventions to prevent misconduct and promote integrity in research and publication. *The Cochrane database of systematic reviews*, *2016*(4), MR000038.

Mayring, P. (2015). Qualitative content analysis: Theoretical background and procedures. In *Approaches to qualitative research in mathematics education* (pp. 365–380). Springer. https://doi.org/10.1007/978-94-017-9181-6_13

Mayring, P. (2019). Qualitative content analysis: Demarcation, varieties, developments. *Forum: Qualitative Social Research*, *20*(3), 1–26.

Mejlgaard, N., Bouter, L. M., Gaskell, G., Kavouras, P., Allum, N., Bendtsen, A. K., & Veltri, G. A. (2020). Research integrity: nine ways to move from talk to walk. *Nature*, *586*(7829), 358–360.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Bmj*, *339*.

Mumford, M. D., Devenport, L. D., Brown, R. P., Connelly, S., Murphy, S. T., Hill, J. H., & Antes, A. L. (2008). Validation of ethical decision-making measures: Evidence for a new set of measures. *Ethics & Behavior*, *16*(4), 319–345. https://doi.org/10.1207/s15327019eb1604_4

Mumford, M. D., Steele, L., & Watts, L. L. (2015). Evaluating ethics education programs: A multilevel approach. *Ethics & Behavior*, *25*(1), 37–60. https://doi.org/10.1080/10508422.2014.917417

Pan, S. J. A. (2021). Taiwanese and American graduate students' misconceptions regarding responsible conduct of research: A cross-national comparison using a two-tier test approach. *Science and Engineering Ethics*, *27*(2), 20. https://doi.org/10.1007/s11948-021-00297-7

Pieper, I., & Thomson, C. J. (2013). Justice in human research ethics: A conceptual and practical guide. *Monash bioethics review*, *31*(1), 99–116.

Pizzolato, D., & Dierickx, K. (2021). Stakeholders' perspectives on research integrity training practices: A qualitative study. *BMC Medical Ethics*, *22*(1), 67. https://doi.org/10.1186/s12910-021-00637-z

Resnik, D. B., & Dinse, G. E. (2012). Do U.S. research institutions meet or exceed federal mandates for instruction in responsible conduct of research? A national survey. *Academic Medicine*, *87*, 1237–1242.

Rest, J. R. (1992). *Development in judging moral issues.* University of Minnesota Press.

Smidt, A., Balandin, S., Sigafoos, J., & Reed, V. A. (2009). The Kirkpatrick model: A useful tool for evaluating training outcomes. *Journal of Intellectual and Developmental Disability*, *34*(3), 266–274. https://doi.org/10.1080/13668250903093125

Solomon, E. D., English, T., Wroblewski, M., DuBois, J. M., & Antes, A. L. (2022). Assessing the climate for research ethics in labs: Development and validation of a brief measure. *Accountability in Research*, *29*(1), 2–17. https://doi.org/10.1080/08989621.2021.1881891

Stahl, B. C. (2021). Ethical issues of AI. In *Artificial Intelligence for a better future: An ecosystem perspective on the ethics of AI and emerging digital technologies* (pp. 35–53). Springer. https://doi.org/10.1007/978-3-030-69978-9

Steele, L. M., Mulhearn, T. J., Medeiros, K. E., Watts, L. L., Connelly, S., & Mumford, M. D. (2016). How do we know what works? A review and critique of current practices in ethics training evaluation. *Accountability in Research*, *23*(6), 319–350. https://doi.org/10.1080/08989621.2016.1186547

Steneck, N. H. (2006). Fostering integrity in research: Definitions, current knowledge, and future directions. *Science and Engineering Ethics*, *12*(1), 53–74. https://doi.org/10.1007/PL00022268

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99–103. https://doi.org/10.1207/S15327752JPA8001_18

Tammeleht, A., & Löfström, E. (2025). How do you know if research ethics and integrity training is effective? – Overview of prior literature. *Ethics & Behavior, 1–14*. https://doi.org/10.1080/10508422.2025.2493055

Taylor, H., Kass, N., Joseph, A., Sisson, S., & Bertram, A. (2012). Development of a research ethics knowledge and analytical skills assessment tool. *Journal of Medical Ethics*, *38*(4), 236–242. https://doi.org/10.1136/medethics-2011-100025

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.

van den Hoven, M., Lindemann, T., Zollitsch, L., & Prieß-Buchheit, J. (2023). A taxonomy for research integrity training: Design, conduct, and improvements in research integrity courses. *Science and Engineering Ethics*, *29*(3), 14. https://doi.org/10.1007/s11948-022-00425-x

van den Hoven, M., & Theunissen, B. (2021). Taking RCR seriously in teaching scientific integrity. *Bioethica Forum*, *14*(1), 89–94.

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

Watts, L. L., Todd, E. M., Mulhearn, T. J., Medeiros, K. E., Mumford, M. D., & Connelly, S. (2017b). Qualitative evaluation methods in ethics education: A systematic review and analysis of best practices. *Accountability in Research*, *24*(4), 225–242. https://doi.org/10.1080/08989621.2016.1274975

Wester, K. L., Willse, J. T., & Davis, M. S. (2008). Responsible conduct of research measure: Initial development and pilot study. *Accountability in Research*, *15*(2), 87–104. https://doi.org/10.1080/08989620801946891

World Conference on Research Integrity (2010). *Singapore statement on research integrity*. https://www.wcrif.org/guidance/singapore-statement

Zollitsch, L., Wilder, N., & Priess-Buchheit, J. (2022). The development of a four-tier test to evaluate training in responsible conduct of research. *Open Research Europe*, *1*, 90.



## Authors and Affiliations

**M. van Loon[1] · D. Crean[2] · M. van den Hoven[1] · J. Katsarov[3] · N. Evans[1]**

✉ M. van Loon
m.loon@amsterdamumc.nl

[1] Department of Ethics, Law and Humanities, Amsterdam Public Health Institute, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

[2] School of Veterinary Medicine, University College Dublin, Dublin, Ireland

[3] Faculty of Management & Technology, Institute for Management & Organization, Leuphana University in Lueneburg, Lueneburg, Germany