Check for updates

METHOD ARTICLE

REVISED The development of a four-tier test to evaluate training in responsible conduct of research

[version 2; peer review: 1 approved, 1 approved with reservations]

Linda Zollitsch[1], Nicolaus Wilder[1], Julia Priess-Buchheit[2]

[1]University of Kiel, Kiel, Germany
[2]University of Applied Sciences Coburg, Coburg, Germany

v2 **First published:** 04 Aug 2021, **1**:90
https://doi.org/10.12688/openreseurope.13339.1
Latest published: 24 Jun 2022, **1**:90
https://doi.org/10.12688/openreseurope.13339.2

Abstract

Although higher education institutions across Europe and beyond are paying more and more attention to research integrity (RI) and responsible conduct of research (RCR) training, there are few studies and little evidence on what works and what does not work in these training sessions. One way to overcome this challenge is to evaluate such training with standardised instruments. Experts and trainers have used qualitative approaches to evaluate their training's successes, but it is difficult to compare their results with others. Sometimes they conduct standardised tests drawn from ethics education or other related fields, but these tests do not assess core themes of research integrity as outlined in the European Code of Conduct for Research Integrity (ECoC, 2017). At present, there is a lack of available standardised instruments designed to specifically evaluate success in this training. This article presents a pre-validated instrument for this purpose. The P2I questionnaire is a four-tier test based on the European Code of Conduct for Research Integrity. In it, testees choose a practice in line with research integrity to address an issue, justify their choice, and describe how confident they are with their answers. The development of the P2I questionnaire is outlined in three steps. First we describe the status quo and gaps in evaluating training success, then we illustrate how we designed the P2I questionnaire using practices and justifications in line (and not in line) with research integrity. In the third step, this P2I questionnaire is pre-validated and revised. This questionnaire is a first attempt to engage in a discourse on standardised research integrity instruments and is one step towards an evidence-based improvement of training sessions.

Keywords

research integrity, teaching and learning, evaluation, four-tier test, responsible conduct of research

Open Peer Review

Approval Status ✓ ?

	1	2
version 2 (revision) 24 Jun 2022		? view
version 1 04 Aug 2021	✓ view	? view

1. **Sophia Jui-An Pan**, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

2. **Guy Widdershoven**, Amsterdam UMC, Amsterdam, The Netherlands
Natalie Evans, Amsterdam UMC, Amsterdam, The Netherlands

Any reports and responses or comments on the article can be found at the end of the article.



This article is included in the Research on Research gateway.



This article is included in the Horizon 2020 gateway.



This article is included in the Research Ethics and Integrity collection.

Corresponding author: Linda Zollitsch (zollitsch@path2integrity.uni-kiel.de)

Author roles: Zollitsch L: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Visualization, Writing – Original Draft Preparation; **Wilder N**: Conceptualization, Formal Analysis, Investigation, Methodology, Project Administration, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Priess-Buchheit J**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Validation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This research was financially supported by the European Union's Horizon 2020 research and innovation programme under the grant agreement No [824488] (Rotatory role-playing and role-models to enhance the research integrity culture [Path2Integrity]).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2022 Zollitsch L *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Zollitsch L, Wilder N and Priess-Buchheit J. **The development of a four-tier test to evaluate training in responsible conduct of research [version 2; peer review: 1 approved, 1 approved with reservations]** Open Research Europe 2022, **1**:90 https://doi.org/10.12688/openreseurope.13339.2

First published: 04 Aug 2021, **1**:90 https://doi.org/10.12688/openreseurope.13339.1

REVISED **Amendments from Version 1**

In the second version of the article, we responded to the reviewers' comments and made linguistic adjustments. As a result, responsible conduct of research has become more prominent, and we have also emphasised the distinction between and common ground with research integrity. We have removed or changed misleading terms. For better comprehensibility, we added additional sources and topics like the difference between, for example, evaluation instruments.

Any further responses from the reviewers can be found at the end of the article

Introduction

In recent decades, research integrity (RI) and responsible conduct of research (RCR) have developed into a fundamental cross-cutting issue in research as well as in the training of (future) researchers (Abdi *et al.*, 2021; Gerber *et al.*, 2020).

Nevertheless, there is still no substantial evidence to show what works and what does not work in teaching and learning RI/RCR (Antes *et al.*, 2009; Godecharle *et al.*, 2013; Marusic *et al.*, 2016; Steneck, 2013).

We acknowledge that RI and RCR tend to overlap and blur into each other in practice, however we align with the European Code of Conduct for Research Integrity (2017). As such we describe RI for this project as a commitment to professional, legal, and ethical responsibilities, values, and principles to self-regulate research (cf. ECoC, 2017, 3ff.). RCR on the other hand is "simply conducting research in ways that fulfil the professional responsibilities of researchers, as defined by their professional organizations, the institutions for which they work and, when relevant, the government and public" (Steneck, 2006, 55), in particular "RCR education can help science take care of itself" (Roth, 2002).

RI describes therefore the framework, such as the ECoC (2017) while RCR concentrates on conducting research that is in line with RI. Since the Evaluation of Path2Integrity focusses on the question, if the practice and justification for this is in line with RI, we use the term RCR training to refer to training in the field of RI and RCR.

One of the main challenges in fostering a research integrity culture through training is that there is a lack of substantial evidence to show what works and what doesn't in teaching and learning RCR. One reason for the lack of evidence is that standardised instruments to evaluate the success of such interventions are rare, particularly for quantitative assessment. Most studies assessing successes of RCR training have used qualitative evaluation tools (e.g. Aubert Bonn & Pinxten, 2021; Hyytinen & Löfström, 2017; Sørensen *et al.,* 2021).

To overcome this lacuna, we designed the so-called P2I questionnaire. This instrument concentrates on practices and justifications in line with RI. It assesses the understanding of and argumentation for research integrity as conceptualised in the European Code of Conduct for Research Integrity (ECoC, 2017). We designed the P2I questionnaire to assess the learning successes of the P2I training programme (Prieß-Buchheit, 2020b) and to compare the test results in a pre-post test design from secondary school students older than 16 up to early career researchers.

The following shows the questionnaire's development process in three steps. After describing the status quo and gaps in evaluating RCR training successes first, this article illustrates how we designed the P2I questionnaire using practices and justifications in line (and not in line) with research integrity in a second step. In a third step, the P2I questionnaire is pre-validated and revised. Developing such a questionnaire is a first step towards evaluating RCR training's successes in a standardised way. By making each design transparent, we open a discourse to reach a standardised instrument at the end, quantitatively assessing learning successes in RCR training.

In particular, we acknowledge that different designs of assessment instruments significantly impact what they evaluate. We took that into account and elaborated on a specific design in the development process of the questionnaire. Whereas traditional multiple-choice questionnaire often assess knowledge on one tier, a multiple-tier structure can give insights into students' justifications. That is why we decided to use a multiple tier structure for the P2I questionnaire

We describe the development process of the questionnaire and its specific items as transparent as possible, to establish a comprehensible, valid, and reliable instrument. We collected answers in line (and not in line) with research integrity for different items as one key element in designing the P2I questionnaire. Furthermore, students and experts proofed each item in the pre-validation process. Some items were ambiguous and have been revised. The result is a pre-validated questionnaire. We publish this (early) design procedure to invite other experts to comment and improve the P2I questionnaire. Hence, this article does not yet present the final validation nor evaluation results.

Status quo and gaps in evaluating successes of RCR training

As previously mentioned, there are different qualitative instruments used for measuring the successes of RI/RCR training. For example, Rissanen & Löfström (2014) use vignettes, Vehviläinen *et al.* (2018) conduct qualitative semi-structured interviews, Poom-Valickis & Löfström (2019) administer interviews, Sarauw *et al.* (2019) use observations, and Berling *et al.* (2019) combine toolbox discussions with self-assessment.

Quantitative instruments to assess RCR training's successes remain rare, however attempts to implement quantitative instruments derive mainly from ethics education. This is because of overlapping training content and similar learning methods. Three of the most frequently used instruments are the

Defining Issues Test (DIT) (Rest, 1979), its revised version, the DIT-2 (Rest *et al.*, 1999), and the sensemaking approach (Mumford *et al.*, 2006).

While the DIT aims to assess cognitive moral reasoning schemes, which focus on dilemma situations, the sensemaking approach assesses specific metacognitive reasoning strategies (Mumford *et al.*, 2008). Although both aspects are demonstrably relevant in RCR training, they are derived from ethics education and do not capture the RCR training core theme as described in the ECoC (2017): RCR training aims to enable participants to become a part of the research community, including "the principles of research, … the criteria for proper research behaviour, [and] … the quality and robustness of research" (ECoC, 3).

What does that mean in particular? There are different ways to transform this overall aim to become part of the research community – into a training. Prieß-Buchheit *et al.* (2020a) describe that Path2Integrity's approach to training is to learn to argue in favour of RI. The theory presupposes that a responsible (future) researcher can argue in favour of practices in line with RI (Prieß-Buchheit *et al.,* 2020a). Hence, after a RCR training, participants should be able to choose practices in line with RI over other practices not in line with RI and should be able to justify their choice. Learning to argue in favour of RI entails different topics such as, how to cite correctly, how to handle research subjects, and hoe to work in collaborative teams etc. (see e.g. Hermeking & Prieß-Buchheit, 2022).

However, a common research integrity challenge is, that one practice can simultaneously be both in line and not in line with RI. For example, the practice to publish many articles can on the one hand, be driven by the purpose to get promoted and therefore have enough money to feed one's family; and on the other hand, could be driven by the purpose to timely inform others about one's findings. The former argument could lead to a salami publication tactic, whereas the latter does not run the same risk. How somebody justify their practices is crucial.

Another instrument, called the *Professional Decision Making in Research* (DuBois *et al.*, 2016) was developed based on the *Ethical Decision Making* (Mumford *et al.,* 2006). This assessment instrument focusses (not on justifications but) on decisions and "ethical problems that include factors such as incomplete knowledge, power discrepancies, and urgency – factors that may interfere with ethical decision making" (DuBois *et al.*, 2016). Also, the *Academic Motivation and Integrity Survey*, as for example presented in Stephens *et al.* (2021), and the *Academic Integrity Self-Evaluation Tools*, as reported in Gaižauskaitė *et al.* (2020) cover the field of research integrity. They all focus on ethical decisions and behaviour.

In distinction to these tools – which concentrate on ethical decisions and behaviour – our approach assesses, if (future) researchers argue in favour of research integrity. With this approach we ensure that the assessment reveals what practice in line with RI people chose in the research community and how they justify this practice. Participants show how responsible they are by choosing one option on each item representing the ECoC's topics of RI.

Until now, there has not been an established standardised instrument that builds on a common reference document such as the ECoC (2017) and covers peoples' argumentation in favour of RI. Nevertheless, some evaluations use a promising structure to assess how people argue in favour of RI. In 2007 Chou *et al.*, applied the so-called two-tier structure – a standard multiple choice question supplemented by a second tier in which the choice is to be justified – to a research integrity-related topic. They assessed students' justifications of cyber copyright laws. Also, Sun (2009) examined the paraphrasing strategies with a two-tier approach, while Pan & Chou (2015) used this structure to assess students' misunderstandings about ethics and behaviour. These approaches demonstrated that a two-tier test seems to be a good fit for a RCR questionnaire concentrating on practices and justifications (not) in line with RI.

Methods

The aim of the P2I questionnaire is to quantify a RCR training's success by looking at the number of answers from the respondents in line with the ECoC (2017). The instrument aims to assess what practice testees would choose in research settings and how they argue in favour of these practices. Practice in this sense means a concrete action, what to do to follow good research practice. The questionnaire also should observe how confident testees are in their answers. This helps to identify whether an answer was given due to strong believe or due to uncertainty. Trainers should be able to see both, how confident testees choose practices in line or not in line with RI and how confident they justify this practice to be in line with RI. This means that we focussed and limited the P2I questionnaire to what practices testees choose and how they argue for this practice in line with RI.

In the following we describe the development of the multi-layered-tier P2I questionnaire. As shown in Figure 1, we developed the questionnaire by following Chandrasegaran, Treagust, and Mocerino's (2007) approach and design, and pre-validated the P2I questionnaire in three steps.

Figure 1 describes the process from step one, defining RCR content to step two, identifying practices and justifications in line (and not in line) with research integrity to result in step three, developing the structure and items to evaluate the RCR training successes.

Step one "Defining the content" is twofold: first we identified RI topics which seem to be relevant in the research environment as well as for the target groups. Then we adapted these topics in regard to different target groups in RCR training. In step two we defined the answers for the first tier (practices in line and not in line with RI) and the third tier (justifications in line and not in line with RI) based on the specific RCR content that was identified in step one. Step three outlines the development

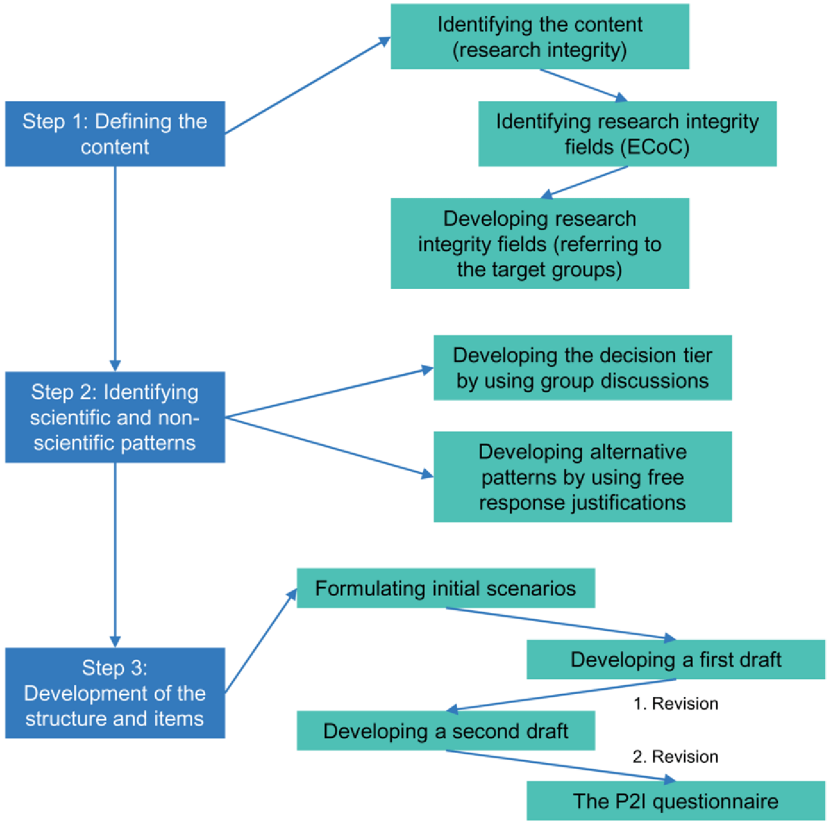


Figure 1. Scheme of development of the P2I questionnaire adapted with permission of The Royal Society of Chemistry, from Chandrasegaran *et al.* (2007); permission conveyed through Copyright Clearance Center, Inc.

of the structure of each item, starting with the pre-validation of each item and refining the items after the first tests runs and with the help of experts.

Ethics statement

All procedures performed in developing the questionnaire, as well as in the questionnaire were approved by the institutional research committee (Central Ethics Committee of the University of Kiel) approval number: ZEK-10/20.

Written informed consent for publication of data was obtained from all subjects before they conducted the P2I questionnaire. All testees proactively clicked on the 'next' button in the online questionnaire to proceed with the voluntary test and checked a mark that they voluntarily give data for this study.

Step one: Defining the content

Designing an accurate, reliable, and valid questionnaire requires considerable knowledge, skill, and experience in question-wording and questionnaire structure. Furthermore, our study requires that we then utilise this questionnaire to evaluate RCR training success which additionally requires critical knowledge on the subject area.

To identify research integrity topics in RCR training we followed Roth (2002), who says: "Since science is self-policing, it may be tempting to think that the scientific community can handle any matters of responsibility by its own methods. This is already rebutted by the creation of regulations to govern scientific research due to past failures of the scientific community to minimise and mitigate misconduct by some scientists. Moreover, RCR education raises issues for scientists to promote reflection and consciousness of their roles as members of the scientific community. Thus, RCR education can help science take care of itself".

An aim of this study is to quantify the success of RCR training. To do so, we have conceptualised our understanding of RCR as follows: First, we adopted an approach from ethics (education), which concentrates on establishing and explaining standards or principles of moral behaviour – which can be differentiated into values, norms, and virtues (Peels *et al.*, 2019) – within a community. Additionally, we focussed on integrity, emphasising the practices of agents who comply with these ethical standards or principles (Löfström & Pyhältö, 2019). We narrowed the questionnaire's focus to be in line with Shaw's (2019) understanding that "research integrity lies

in consistency with external rules" (p. 1087) and aimed to assess what practices testees choose in research settings and how they argue in favour of these practices.

To identify what content should be included in the questionnaire, we clarified what common ground we expect all RCR trainings to have. In a nutshell, Jordan (2013, 245–6) differentiates ethics from integrity. She describes ethics as "the standards of moral behaviors expected of autonomous humans living in a community, often through reference to a coherent system of thought or an ethical theory (e.g., deontology). [In opposition to that,] integrity describes how an individual or institution brings their moral positions and behaviors together in a logically coherent position that is brought into daily practice." This corresponds with Steneck (2006, 56), who describes research integrity as the "quality of possessing and steadfastly adhering to high moral principles and professional standards, as outlined by professional organisations, research institutions and, when relevant, the government and public". Following Jordan's statements above, learning RCR is not about blindly following rules (history has taught society many times the consequences of such attempt). Instead, learning RCR is about following rules due to a discourse based on rational arguments that legitimises all statements that claim normative validity and must be agreed upon by individuals (Habermas, 1990 and Gethmann, 2010). In Europe we have the ECoC (2017) as a common framework, displaying such a discourse. Such a perspective emphasises the individual's personal responsibility regarding research integrity (Frankel *et al.*, 2015), which is (only) compatible with RI/RCR training but not sufficient for fostering a culture of research integrity (Valkenburg *et al.*, 2021).

In identifying good research practices, we concentrated on the ECoC (2017), which names the following RI topics: Research Environment, Research Procedures, Safeguards, Collaborative Working, Publication and Dissemination, Data Management, Mentoring, and Reviewing, Evaluating and Editing. Our aim was to assess training from secondary school students to early career researchers. According to the European Framework for Research Careers, Reviewing, Evaluating and Editing are getting relevant for established researchers (R3) (European Commission, 2011, 10) Therefore, we decided to include all the fields from the ECoC (2017) mentioned above, except the field Reviewing, Evaluating and Editing.

As shown in Table 1, we included the fields Research Environment, Research Procedures, Safeguards, Collaborative Working, Publication and Dissemination, Data Management, as well as Mentoring and designed items. We created two versions of the questionnaire. The S and the M / Y versions adapt, transform, and reduce the above-described topics to two target groups by choosing concrete content from these topics. The two versions account for academic discrepancies between the target groups. Namely that secondary school students and bachelor students do not have an academic degree yet (S version) (Zollitsch *et al.*, 2020b), whereas master students and early career researchers hold an academic degree (M / Y version) (Zollitsch *et al.*, 2020a).

As a result, the P2I questionnaire contains two versions with several items. The items in each version adapt to pre-collegiate and collegiate learning circumstances. The testees see the items and each response possibility in random order to ensure there is no bias within the items. In this way, we reduce the possibility that there is a dependence between items, which might allow testees to answer an item more easily because of the previous item that they already finished.

Step two: Identification of practices and justifications which are and which are not in line with RI rules and principles

To identify which practices are in line and which practices are not in line with RI, we first designed different scenarios relating to the ECoC (2017) topics. To find suitable multiple choice answers and distractors to the designed scenarios, we then conducted group discussions with both students and researchers. In total, we conducted six group discussions with students majoring in education as well as with an interdisciplinary research group at Kiel University in Germany. The persons in the group discussions were not randomised; participation was voluntary. Participants were approached personally since this procedure involved a considerable investment of time for the participants and no compensation was offered for their participation. To make the procedure as efficient as possible students with high intrinsic motivation in research, and a lack of any particular connection to RCR were asked to help improve the questionnaire.

We presented the first case scenarios to each focus group and asked them what one should do to act in line with RI. We sought open and free-response questions. With this procedure, we collected attractive practices in line and not in line with RI for two different target groups: a) secondary school students and undergraduates and b) master students and (early career) researchers. Given that our focus group participants were both

Table 1. Content of the P2I questionnaire (referring to the European Code of Conduct).

	Version S	Version M / Y
Research Environment	X	X
Research Procedure	X	X
Safeguards	X	X
Collaborative Working	X	X
Publication and Dissemination	X	X
Mentoring		X
Data Management		X

students and researchers, we were able to collect attractive practices from their living environment.

In a second step, we analysed common justifications to find out which justifications are not in line with RI and are more common in a specific test group, culture, gender, or age. We designed a draft questionnaire with seven items and open-ended questions to identify justifications and justification patterns for the aforementioned practices (Draft of the P2I questionnaire, Version 1.0) (Zollitsch *et al.*, 2019a).

A total of 35 participants, from both target groups (aged 18 to 50) filled out the draft questionnaire. The participants were, mainly from the social sciences and were equally gender-mixed from Kiel University in Germany, they were approached personally, participation was voluntary, and the group was not randomised. In order to reach the participants, the lecturers of the Department of Education at Kiel University were asked to have students fill out the questionnaire within their courses. In the end, the participants came from six separate courses with four different lecturers. The participants answered the draft questionnaire in a pen-and-paper version.

The following displays some original justification examples:[1]

- Research means being creative and not always sticking to rules.
- In research, one should think outside the box and break away from the rules to discover new things.
- The regulations are often written in a complicated way and do not provide a good reference point for writing a paper.
- One should not refer purely to regulations because there is little practice associated with it.

We analysed all 232 free text responses carefully and elaborated repeating justifications by doing a qualitative content analysis, as described in Mayring (2000). The development of inductive categories revealed justification patterns from the everyday life of the target groups that we used. The answers were reduced to their content-bearing components and comparable answers were combined in different patterns. Finally, a counter check with the original answers confirmed the final structure. As shown in Table 2, the answers from the participants led to nine different patterns that contain different justifications as well as one justification pattern according to the ECoC.

We selected the justification pattern according to the ECoC as well as three justification distractors for each scenario. We reformulated the distractors as little as possible, as we wanted to stay close to the typical answer while also ensuring connectivity to the scenario of the item.

In the following, we use the term scientific common sense (SCS) when the statement is in line with the RI rules and principles from the ECoC. We combine different statements, which are in line with RI, in a justification pattern called the 'scientific common sense pattern'. Statements from this pattern are distinguishable from other statements, which are not in line with the scientific common sense, such as the rejection of binding codes or individual benefits (as presented in Table 2). We take into account that statements other than our predefined SCS statements could also be correct and expected in specific settings other than our RI-setting.

[1] The examples are translated from German to English by DeepL.

Table 2. Justification patterns (not) in line with RI.

Justification patterns	Answers
Scientific common sense (according to the ECoC)	It ensures reliable research results.
Hierarchy	The supervisor said X has to do it that way.
Structural conditions	The structural conditions have made this scientific practice necessary.
Individual Benefits	This scientific practice will support X's career and get him/her the next Nobel prize.
Community benefits	Only with this scientific practice can the highest benefit for the community be ensured.
Equal treatment of all	The same requirements should apply to everyone.
Duty	Good researchers must do this.
The others	As long as others do this, X can do it.
Quantitative majority decisions	Since the majority does so, X must adapt his/her scientific practice accordingly.
Rejection of binding codes	Due to the complexity and variety of research, there can be no binding codes and regulations.

Step three: Development of the structure and items

To assess the success of RCR training, we choose a so called multi-layer-tier structure for the questionnaire, specifically we used models from the co-called tier-test designs. Treagust (1988) designed the original structure of the two-tier test as an extension of classic multiple-choice tests to determine not only what students know but also how students understand a specific topic. This two-tier approach can help differentiate between multiple choice answers which are right by chance and which are right because there is an understanding of the concepts. Treagust's original two-tier test is a two part (a.k.a. tier) multiple choice series. It starts with a multiple-choice question that the testees respond to in the first tier; this extends to a second-tier where the testees reason their response from the first tier. He based his original design on Tamir's (1971) approach in which distractors were derived from existing misconceptions of the testees – this would make them more attractive. The two-tier test is an extension of classical multiple-choice tests and helps to determine not only what students know but also what rationale underpins that answer.

Such a multi-layered tier test provides quantitative data from both tiers and provides insights into the first tier's quality (Yang & Lin, 2015). This design suits the needs of this study as it allows statements about the testees' choice of appropriate practices in RCR and the justifications for these practices. Furthermore, it identifies frequently occurring and stable rationales. As Hestenes & Halloun (1995) or Yan & Subramaniam (2018) show, the two-tier test is able to show what action testees would suggest in a situation and, at the same time, why they would do that. Following this idea, we designed the P2I questionnaire as a multi-layered tier test, assessing the testees' choices of research practices and justifications, as Tsui & Treagust (2010), Chandrasegaran *et al.* (2007), or Treagust (1988) outlined.

First pilot test and revision

For the first pilot run, 11 students conducted the two-layered P2I questionnaire. All 11 students were from the educational field (six bachelor students and five master students) at Kiel University in Germany, but were from different courses. They were given a a pre-version (Pre-Version of the P2I questionnaire, Version 1.0) (Zollitsch *et al.*, 2019b), derived from the draft questionnaire (Zollitsch *et al.*, 2019a). The group was not randomised; the participation was voluntary. In order to acquire participants as efficiently as possible, the lecturers who had already agreed to support the study before were approached again. To avoid distorting the results, we only selected courses in which participants had not already completed the test. This left two courses where the questionnaire could be distributed. The participants answered six items: two items for the category *Research Procedures*, two items for the category *Collaborative Working* and two items for the category *Publication and Dissemination*.

Their results in Table 3 show that at least six out of 11 participants answered in line with RI in both tiers in all six items. Item cd9 received the same number of answers in line with RI in the first tier and the second tier. All other five items show that even if the participants answered in the first tier in line with RI, they did not always choose the justification in line with RI.

Only in two items 72.73% of the participants choose their answer in line with RI in the second tier, while in the other four items only 54.55% of the participants choose their answer in line with RI in the second tier. For this reason we readjusted the item scenario as well as tier one.

After revising the pilot test, a second pilot run was conducted with four master students. These students were given the P2I questionnaire (P2I Questionnaire S, Vers. 1.0) (Zollitsch *et al.*, 2020b) in a pre-post-test design. They all came from interdisciplinary studies in Coburg University in Germany, and the age range was between 23 and 32 years. The group was not randomised and the participation was voluntary. To have a full pre-post-pilot run, subjects who participated in the first interventions of the P2I training were asked to conduct the questionnaire before and after their training. They were personally contacted by the instructor of the P2I training. Since the test was voluntary, only four participants finished the pre- and the post-test. Nevertheless, this test run exemplified that the complete questionnaire design, as presented in the P2I Questionnaire S, Vers. 1.0 (Zollitsch *et al.*, 2020b) works.

Figure 2 shows that in the adjusted questionnaire, more participants chose a practice in line with RI in the post test in the items cp5 and cc5. In the items cc1 and cd2, the participants answered with the same frequency choosing practices in line with RI in the pre-test as well as in the post-test, while for the items cp3 and cd9 practices in line with RI have been chosen less often in the post test.

Acknowledging that the test did not assess learning successes as wished, we decided to rearrange the structure by using the above-described tier models and rewrote the failing items.

Table 3. Results of the first test.

Scientific practice fields	Research procedure		Collaborative working		Publication and dissemination	
Item	cp3	cp4	cc1	cc5	cd2	cd9
Tier 1 practice in line with RI (%)	8 (72.73)	10 (90.91)	7 (63.64)	8 (72.73)	8 (72.73)	8 (72.73)
Tier 2 justification in line with RI (%)	6 (54.55)	8 (72.73)	6 (54.55)	6 (54.55)	6 (54.55)	8 (72.73)

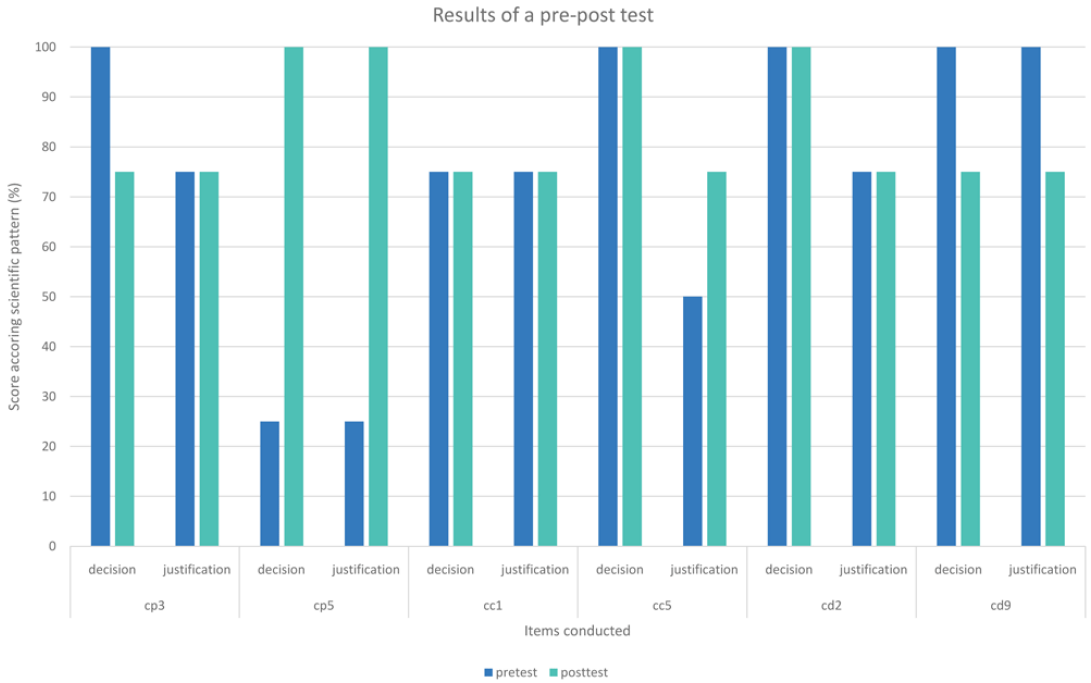


Figure 2. Comparison of the results of a pre- post-test.

We modified the original structure. Instead a binary response in the first tier (decision between yes or no), the new version of the P2I questionnaire includes four possible options. At the same time, the justification patterns (not) in line with RI that were previously based on whether the practice corresponds to a practice in line with RI or not were summarised from four justifications per answer (meaning eight in total: four yes/four no) into a total of four justification patterns. Thus, someone who chooses a practice not in line with RI can still select a rationale that fits a justification pattern in line with RI.

The first tier no longer requires testees to decide whether a described practice is or is not in line with RI, but rather testees must choose which of the four given practices is in line with RI.

We expanded the P2I questionnaire's multiple-tier approach to evaluate in detail if the testees can determine which practice is in line with RI and how to justify this decision. We added two confidence tiers (Caleon & Subramaniam, 2010; Peşman & Eryılmaz, 2010) to create a so-called four-tier structure. This approach is used to find out if testees are guessing their answers. These confidence tiers determine how sure the testees are with their responses. Information from four-tier tests provides valuable insights into the development and improvement of training. They enable conclusions, especially in cases of answers not in line with RI, for example, if testees are using a specific justification more often than others. Data gives insights into if there is simply a lack of knowledge – this would be expressed in low confidence – or whether the problem lies in a stable pattern (not) in line with RI. (P2I Questionnaire S, Vers. 2.0 (Zollitsch *et al.*, 2020b) and P2I Questionnaire M, Vers. 2.0 (Zollitsch *et al.*, 2020a))

Test for pre-validating and revision

To pre-validate the final four-tier structure of the P2I questionnaire, the developers conducted a content validation test (Miyata & Kai, 2009) with international research integrity and research ethics experts. In the first step, the RI/RE experts gave feedback on the content. In a next step, the validation test followed the hypothesis that international experts would respond with high confidence to the P2I questionnaire; and students (without specific training) would respond with less confidence than experts.

The test developers invited all 17 members of the Horizon 2020 project Path2Integrity and 12 additional research integrity and research ethics experts working with the training centre of Path2Integrity to validate the content. Invitations were extended via email only to those who did not have a connection to the learning units or the P2I questionnaire. Their credentials were accredited via their respective position in either the P2I advisory board, the P2I scientific board, or by being a contact from the P2I training centre. The validation procedure instructed all experts to complete both versions (the S and the M/Y version), to give feedback, to make the questionnaire's content valid as possible, and ask questions when needed. The participation was voluntary, and the group was not randomised.

We received 27 responses to the P2I questionnaire from the experts' group mentioned above. The experts came from various disciplines (including Humanities, Medicine, Law, and Natural Sciences) around the world. The experts' group was gender-balanced, and the age structure varied from 35 to 67.

Each expert received both versions of the questionnaire (P2I Questionnaire S, Vers. 2.0 ((Zollitsch *et al.*, 2020b) and P2I Questionnaire M, Vers. 2.0 (Zollitsch *et al.*, 2020a)) and all of them were asked to answer both and comment on them. 12 experts filled out version S (P2I Questionnaire S, Vers. 2.0) (Zollitsch *et al.*, 2020b), and 15 filled out version M / Y (P2I Questionnaire M, Vers. 2.0) (Zollitsch *et al.*, 2020a). From those 27 replies, only 12 contained content-specific comments. The test developers checked these comments and integrated them into the P2I questionnaire.

The experts' comments did not only refer to the content but rather often to aspects of design. For example, having an open field for the category country, not storing IP addresses, and providing a more prominent link to the data policies. We implemented those recommendations accordingly. One comment stated that items for data management are missing in version S. As shown in 'table 1. Content if the P2I questionnaire (referring to the European Code of Conduct)' data management is a topic for those who graduated. Therefore, we did not integrate those items in version S. Furthermore, two comments declared that they are missing the handling of dilemma situations. As explained in the status quo, this questionnaire does not focus on moral dilemmas but rather on contradicting practices, and on arguing in favour of research integrity, and therefore, does not contain dilemma situations. On top of that, experts commented on the test's difficulty; they stated that some items were difficult to answer.

Two comments directly addressed the content of one item, namely S1. These comments led to an item amendment. The developers extended the option "ombudsperson" to "ombudsperson/research integrity officer" because these terms are synonyms but used exclusively in some countries. Another comment was on item S4. Experts asked the developers to clarify the case as well as the answers. The developers shortened the item from "Sarah is in charge of the coordination of a research project" to "Sarah is in charge of a research project". Also, the answer "informs each group that it is responsible for establishing and adhering to its own principles" was changed into "informs each partner that they are responsible for establishing and adhering to their own principles". A third expert comment asked to give more than one response option. The developers declined this comment after discussing and weighing it against the central concept of the P2I questionnaire, since the participants should decide on only one practice and justify this practice. If they are not sure about their answer given, they have always the option to tell by using the slider in the second and fourth tier.

The procedure confirmed the content validity of the P2I questionnaires because none of the experts questioned the validity of the test and the few content-specific comments were adjusted in consultation with the experts. As there were no indications that other RI and RE experts would come to a fundamentally different conclusion regarding the content validity of the questionnaire, we considered the number of participants to be saturated and refrained from consulting further experts.

In the next step, we contacted six lecturers from Kiel University in Germany to acquire students from non-RI/RE related courses. Two lecturers offered 15 students that had not already completed the test. The P2I questionnaire was administered by the lecturers personally. The participation was voluntary, and the group was not randomised. Ten students participated in this second step of the validation study (five in Version S (P2I Questionnaire S, Vers. 2.0, (Zollitsch *et al.*, 2020b)) and five in Version M (P2I Questionnaire M, Vers. 2.0)) (Zollitsch *et al.*, 2020a). The students' group consisted of five bachelor students and five master students. They were between 23 and 29 years old, with seven females and three males (from Germany); most of them from the field of social science, some from humanities and economics.

The following describes the confidence (tier two and four) for a small sample of items from the S version of the P2I questionnaire as an example. Beyond these examples described, we pre-validated all items (and both versions) of the P2I questionnaire. Figure 3 shows the students' confidence with their answers from the first tier and the third tier combined compared to the experts' confidence (for the items in version S). (P2I Questionnaire S, Vers. 2.0) (Zollitsch *et al.*, 2020b)

Figure 3 shows the contrast of the average student's confidence and the average expert's confidence and the median. Students are less confident with their decisions in the first and in the third tier, except for Item S3. The median in all items is higher for the experts than for the students. Also, the range of confidence is much broader in the students' group than in the experts' group, which displays that the students might be confident within some decisions but not in all research integrity topics.

The result of this comparison is not ideal, because experts' confidence does not concede students' confidence in one out of five items. Nevertheless, comparing how confident the students and experts responded showed that students, in general, are less confident than experts. This comparison documents that the P2I questionnaire is (only) a pre-valid instrument to assess research integrity training, as the students are not as familiar with research integrity as experts.

Results

The result of the development and validation is a pre-validated P2I questionnaire, which needs to be validated in a second pilot study. The following describes the P2I questionnaire in its latest content and structure (P2I Questionnaire S, Vers. 2.1 (Zollitsch *et al.*, 2020b) and P2I Questionnaire M, Vers. 2.1 (Zollitsch *et al.*, 2020a)).

As shown in Figure 4, the P2I questionnaire opens with a scenario followed by a choice of which practice is in line with RI. This decision in the first tier consists of a multiple-choice question with four possible options to choose from. The first tier asks what the person in the scenario should do to follow a good research practice. The questionnaire's second tier asks on a scale from 1 (very unsure) to 100 (very confident), how confident the testees were with their choice in

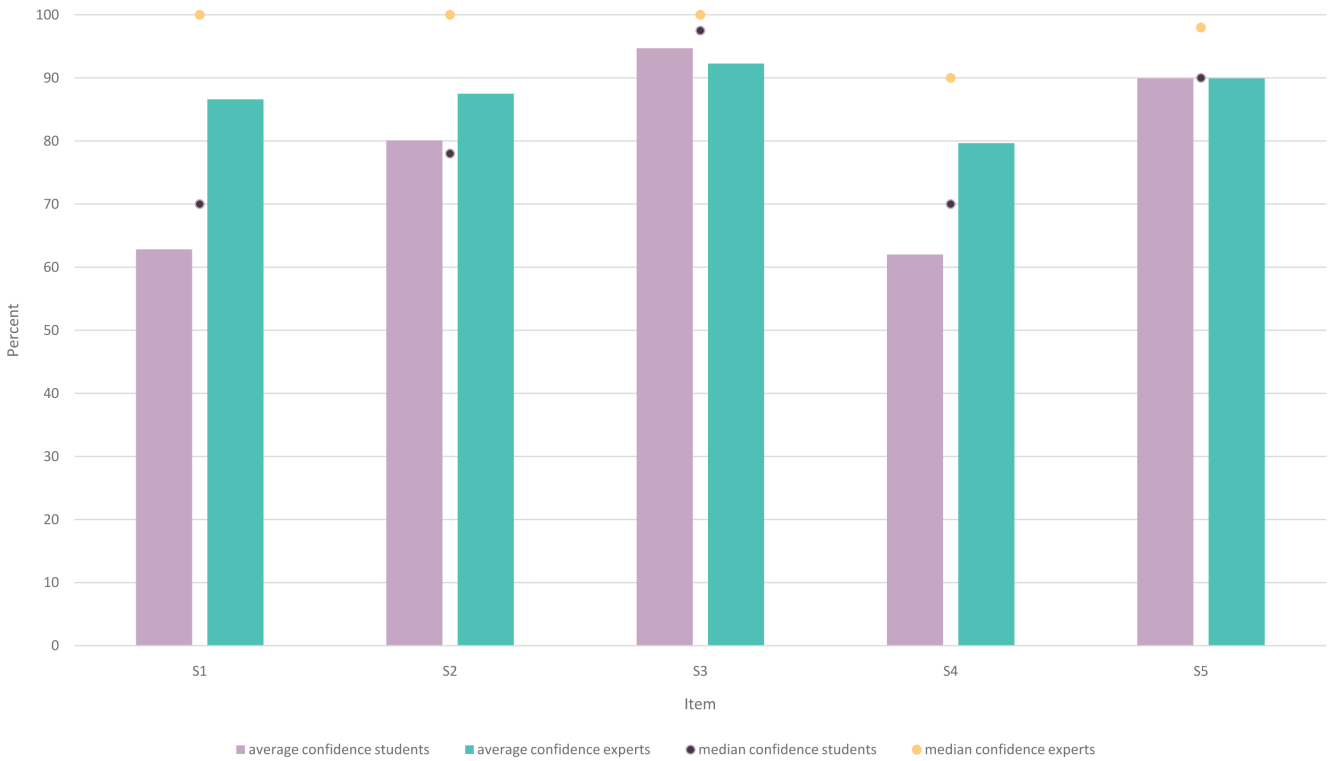


Figure 3. **Comparison of the confidence of experts and students.**

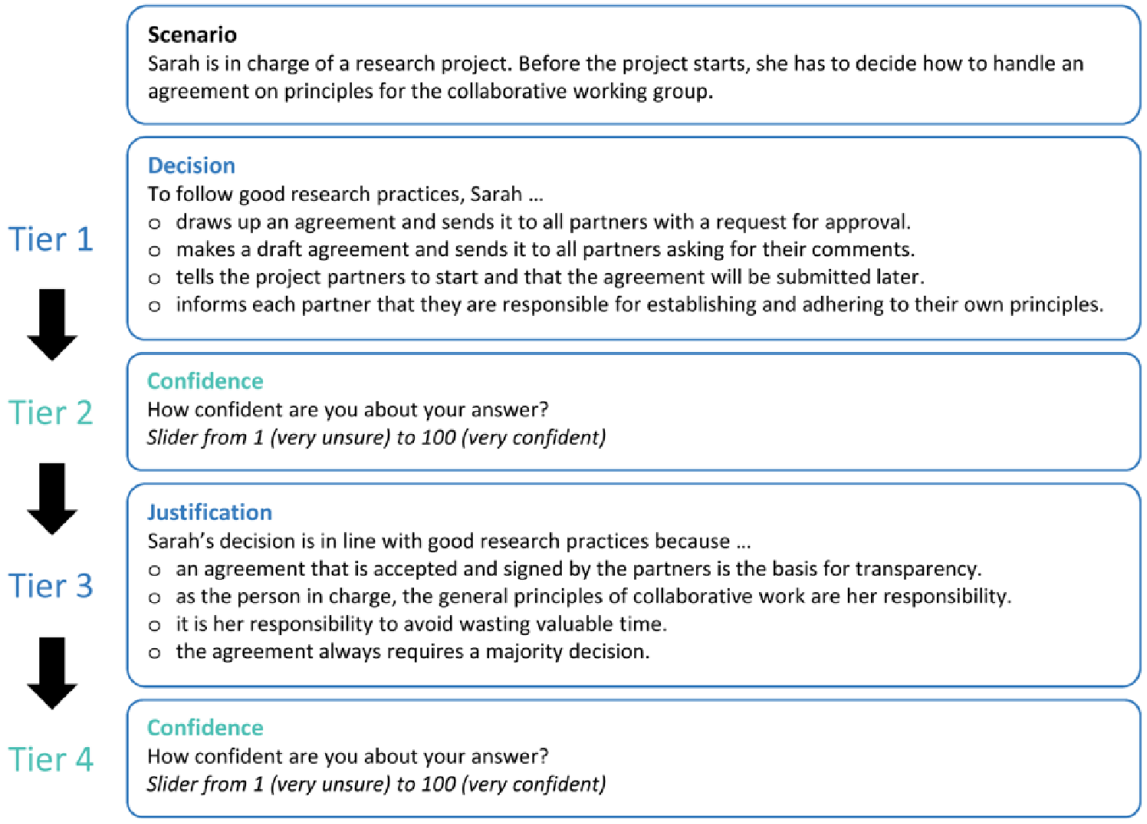


Figure 4. **Exemplary item of the P2I questionnaire from the field of collaborative working.**

the first tier. In the third tier, testees are asked a multiple-choice question with four alternatives to justify the practice chosen in the first tier. Again, the next tier asks on a scale from 1 to 100 how confident the testees are about their third-tier justification.

The confidence tiers are essential for determining if the testees are confident about both responses in tier one and tier three. Ideally, they choose a practice in line with RI, justify this choice with a justification in line with RI while confirming that they are confident with both answers.

To consider relevant differences that might occur, the P2I questionnaire also contains different independent variables to evaluate RCR training successes next to the tiers' responses.

Discussion

As stated earlier, one of RCR training's core themes is to enable participants to become a part of the research community. One way to assess this is to observe what practices (future) researchers choose in their research environment and how they justify their choice. As shown in the three different steps above, the assessment of justifications is – in particular – difficult. The P2I questionnaire is a first attempt to assess such practices and justifications (not) in line with RI.

The P2I questionnaire derives from Path2Integrity's RCR training, compared to the DIT-2, which originates from ethics education. The P2I questionnaire focuses on practices in line with RI in reference to the research community – in particular – in reference to the ECOC (2017). Both instruments start with a (case) scenario that testees should respond to. While the DIT-2 asks which practice is the most important, the P2I questionnaire asks to justify this practice. The P2I questionnaire, therefore, gives insights how testees argue in favour of specific practices.

The multiple-tier structure is a frequently used design for assessing RCR related topics in the Asia-Pacific region. In other educational fields, it is often used for natural science education, such as Geology (Chang *et al.*, 2020), Physics (Winarti *et al.*, 2017), or Mathematics (Fratiwi *et al.*, 2017). Based on the ideas of Chou *et al.* (2007) or Sun (2009), the P2I questionnaire transferred this multi-layered design to assess RCR training's successes.

Although the P2I questionnaire has been developed in the Horizon 2020 project Path2Integrity, it will also be applicable to other RCR training. The P2I questionnaire's framework ECoC (2017) is expandable and applicable to other reference documents such as the Singapore Statement, 2010, the Montreal Statement, 2013 etc. These codes refer to the same practices in line with RI and entail similar justifications. The P2I questionnaire focuses on its assessment on practices and justifications in favour of research integrity. It is therefore suitable for RCR training that focuses on dialogical approaches (Koterwas *et al.*, 2021).

The P2I questionnaire is a first step to offering a standardised instrument with which the RI community can achieve reproducible results and compare different RCR training. This article contributes to the dialogue about standardising and standardised instruments ton assess RCR training.

Limitations

The P2I questionnaire is a specific tool with limitations. It demands a high reading competence of the testees, as it is very much text based. Users need to consider that they also measure the literacy of the testees. Due to the complexity of the four-tier test structure, some testees also may drop out during the evaluation. Limitation of the pre-validation and testing: No secondary school students designed and pre-validated the P2I questionnaire. The leading group of involved participants – next to the 27 responses of international experts - affiliates to educational studies, and most of them identified their country as Germany.

This article described the development of the P2I questionnaire following Chandrasegaran, Treagust, and Mocerino's (2007) approach, designing and pre-validating the P2I questionnaire in three steps. As shown above, the expert student comparison did not receive ideal results. That is why further research with the altered P2I questionnaire version is needed. Before the questionnaire is used as a standardised assessment instrument for learning success in RCR training, a second validation is necessary.

Data availability

Underlying data

The data of all parts of the pilot test cannot be shared because it might jeopardise the achievement of the main objective of the Path2Integrity project as described in H2020 MGA article 29.3. As the data contains answers especially from experts, these might be used as orientation from future participants of the P2I test and this will jeopardise the result. All data will be shared at the end of the project, latest at the end of June 2022 in the data project included in the data statement below, Zenodo.

Zenodo: P2I Questionnaire Version MY, https://doi.org/10.5281/zenodo.4682660. (Zollitsch *et al.*, 2020a)

This project contains the following underlying data:

- P2I questionnaire MY Vers_1.pdf
- P2I questionnaire MY Vers_2.pdf
- P2I questionnaire MY Vers_2.1.pdf

Zenodo: P2I Questionnaire Version S, 10.5281/zenodo.4680216. (Zollitsch *et al.*, 2020b)

This project contains the following underlying data:

- P2I questionnaire S Vers_1.pdf
- P2I questionnaire S Vers_2.pdf
- P2I questionnaire S Vers_2.1.pdf

Extended data

Zenodo: Draft of the P2I questionnaire, https://doi.org/10.5281/zenodo.5059949. (Zollitsch *et al.*, 2019a)

Zenodo: Pre-Version of the P2I questionnaire, 10.5281/zenodo.5059965. (Zollitsch *et al.*, 2019b).

The data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

The data of all parts of the pilot test cannot be shared because it might jeopardise the achievement of the main objective of the Path2Integrity project as described in H2020 MGA article 29.3. As the data contains answers especially from experts, these might be used as orientation from future participants of the P2I test and this will jeopardise the result. All data will be shared at the end of the project, latest at the end of June 2022 at Zenodo. If prior access to the data is desired, this can be granted on the condition that neither the data nor analyses based on the data are published in any way. If access to the data is required before 30.6.2022, please contact the first author of the paper Linda Zollitsch (zollitsch@path2integrity.uni-kiel.de).

Acknowledgements

We thank the whole Path2Integrity Consortium Members and the Advisory Board, especially Dick Bourgeois-Doyle (in his function as chair of the advisory board) for his critical and constructive comments on the paper.

References

Abdi S, Pizzolato D, Nemery B, *et al.*: **Educating PhD Students in Research Integrity in Europe.** *Sci Eng Ethics.* 2021; **27**(1): 5.
PubMed Abstract | **Publisher Full Text**

All European Academies (ALLEA): **The European Code of Conduct for Research Integrity.** Revised Edition. Berlin. 2017.
Reference Source

Antes AL, Murphy ST, Waples EP, *et al.*: **A Meta-Analysis of Ethics Instruction Effectiveness in the Sciences.** *Ethics Behav.* 2009; **19**(5): 379–402.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Aubert Bonn N, Pinxten W: **Rethinking success, integrity, and culture in research (part 2) - a multi-actor qualitative study on problems of science.** *Res Integr Peer Rev.* 2021; **6**(1): 3.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Berling E, McLeskey C, O'Rourke M, *et al.*: **A New Method for a Virtue-Based Responsible Conduct of Research Curriculum: Pilot Test Results.** *Sci Eng Ethics.* 2019; **25**(3): 899–910.
PubMed Abstract | **Publisher Full Text**

Caleon IS, Subramaniam R: **Do Students Know What They Know and What They Don't Know? Using a Four-Tier Diagnostic Test to Assess the Nature of Students' Alternative Conceptions.** *Res Sci Educ.* 2010; **40**(3): 313–337.
Publisher Full Text

Chandrasegaran AL, Treagust DF, Mocerino M: **The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation.** *Chem Educ Res Pract.* 2007; **8**(3): 293–307.
Publisher Full Text

Chang SC, Hsu TC, Kuo WC, *et al.*: **Effects of applying a VR-based two-tier test strategy to promote elementary students' learning performance in a Geology class.** *Br J Educ Technol.* 2020; **51**(1): 148–165.
Publisher Full Text

Chou C, Chan PS, Wu HC: **Using a two-tier test to assess students' understanding and alternative conceptions of cyber copyright laws.** *Br J Educ Technol.* 2007; **38**(6): 1072–1084.
Publisher Full Text

DuBois JM, Chibnall JT, Tait RC, *et al.*: **Professional Decision-Making in Research (PDR): The Validity of a New Measure.** *Sci Eng Ethics.* 2016; **22**(2): 391–416.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

European Commission: **Towards a European Framework for Research Careers.** Brussels. 2011.
Reference Source

Frankel MS, Leshner AI, Yang W: **Research Integrity: Perspectives from China and the United States.** In T. Bretag (Ed.), *Handbook of Academic Integrity.* Springer Singapore. 2015; 1–16.
Publisher Full Text

Fratiwi NJ, Kaniawati I, Suhendi E, *et al.*: **The transformation of two-tier test into four-tier test on Newton's laws concepts.** In: MATHEMATICS, SCIENCE, AND COMPUTER SCIENCE EDUCATION (MSCEIS 2016): Proceedings of the 3rd International Seminar on Mathematics, Science, and Computer Science Education. Bandung, Indonesia, 15 October 2016: Author(s) *AIP Conference Proceedings.* 2017; **1848**(1): 50011.
Publisher Full Text

Gaižauskaitė I, Glendinning I, Foltýnek T, *et al.*: **Academic Integrity Self-Evaluation Tools.** ENAI Report 3B-3E: revised version, May. 2020.
Reference Source

Gethmann CF: **Die Aktualität Methodischen Denkens.** In C. F. Gethmann and J. Mittelstraß (Eds.), Paul Lorenzen zu Ehren, UVK. 2010; 15–37.

Gerber A, Forsberg EM, Shelley-Egan C, *et al.*: **Joint declaration on mainstreaming RRI across Horizon Europe.** *J Responsible Innov.* 2020; **7**(3): 708–711.
Publisher Full Text

Godecharle S, Nemery B, Dierickx K: **Integrity training: Conflicting practices.** *Science.* 2013; **340**(6139): 1403.
PubMed Abstract | **Publisher Full Text**

Habermas J: **Discourse ethics: Notes on a program of philosophical justification.** In: *Moral consciousness and communicative action.* Cambridge: MIT Press, 1990; 43–115.

Hermeking N, Prieß-Buchheit J: **What's integrity got to do with it? Second-year experiences of the Path2Integrity e-learning programme.** *FACETS.* 2022; **7**: 110–125.
Publisher Full Text

Hestenes D, Halloun I: **Interpreting the force concept inventory: A response to March 1995 critique by Huffman and Heller.** *The Physics Teacher.* 1995; **33**(8): 502.
Publisher Full Text

Hyytinen H, Löfström E: **Reactively, Proactively, Implicitly, Explicitly? Academics' Pedagogical Conceptions of how to Promote Research Ethics and Integrity.** *J Acad Ethics.* 2017; **15**: 23–41.
Publisher Full Text

Integrity in research collaborations: **The Montreal Statement.** *Lancet.* 2013; **382**(9901): 1310.
PubMed Abstract | **Publisher Full Text**

Jordan SR: **Conceptual Clarification and the Task of Improving Research on Academic Ethics.** *J Acad Ethics.* 2013; **11**(3): 243–256.
Publisher Full Text

Koterwas A, Dwojak-Matras A, Kalinowska K: **Dialogical teaching of research integrity: an overview of selected methods.** *FACETS.* 2021; **6**: 2138–2154.
Publisher Full Text

Löfström E, Pyhältö K: **What Are Ethics in Doctoral Supervision, and How Do They Matter? Doctoral Students' Perspective.** *Scand J Educ Res.* 2019; **64**(4): 535–550.
Publisher Full Text

Marusic A, Wager E, Utrobicic A, *et al.*: **Interventions to prevent misconduct and promote integrity in research and publication.** *Cochrane Database Syst Rev.* 2016; **4**(4): MR000038.
PubMed Abstract | **Free Full Text**

Mayring P: **Qualitative Content Analysis [28 paragraphs].** *Forum Qualitative*

Sozialforschung / Forum: Qualitative Social Research. 2000; **1**(2): 20.
Publisher Full Text

Miyata H, Kai I: **Reconsidering Evaluation Criteria for Scientific Adequacy in Health Care Research. An Integrative Framework of Quantitative and Qualitative Criteria.** *Int J Qual Methods.* 2009; **8**(1): 64–75.
Publisher Full Text

Mumford MD, Connelly S, Brown RP, *et al.*: **A Sensemaking Approach to Ethics Training for Scientists: Preliminary Evidence of Training Effectiveness.** *Ethics Behav.* 2008; **18**(4): 315–339.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Mumford MD, Devenport LD, Brown RP, *et al.*: **Validation of Ethical Decision Making Measures: Evidence for a New Set of Measures.** *Ethics Behav.* 2006; **16**(4): 319–345.
Publisher Full Text

Pan SJA, Chou C: **Using a Two-Tier Test to Examine Taiwanese Graduate Students' Misunderstanding of Responsible Conduct of Research.** *Ethics Behav.* 2015; **25**(6): 500–527.
Publisher Full Text

Peels R, de Ridder J, Haven T, *et al.*: **Value pluralism in research integrity.** *Res Integr Peer Rev.* 2019; **4**(1): 18.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Peşman H, Eryılmaz A: **Development of a Three-Tier Test to Assess Misconceptions About Simple Electric Circuits.** *J Educ Res.* 2010; **103**(3): 208–222.
Publisher Full Text

Poom-Valickis K, Löfström E: **"Pupils should have respect for you, although I have no idea how to achieve this?": The ideals and experiences shaping a teacher's professional identity.** *Educ Stud.* 2019; **45**(2): 145–162.
Publisher Full Text

Prieß-Buchheit J, Aro AR, Demirova I, *et al.*: **Rotatory role-playing and role-models to enhance the research integrity culture.** *Res Ideas Outcomes.* 2020a; **6**: e53921.
Publisher Full Text

Prieß-Buchheit J: **Path2Integrity Learning Cards: First Year Experiences of an Educational Programme to Foster Research Integrity in Europe.** *EDUKACJA.* 2020b; 54–66.
Reference Source

Rest JR: **Development in Judging Moral Issues**. University of Minnesota Press, 1979.
Reference Source

Rest JR, Narvaez D, Thoma SJ, *et al.*: **DIT2: Devising and Testing a Revised Instrument of Moral Judgement.** *J Educ Psychol.* 1999; **91**(4): 644–659.
Publisher Full Text

Rissanen M, Löfström E: **Students' Research Ethics Competences and the University as a Learning Environment.** *Int J Educ Integr.* 2014; **10**(2): 17–30.
Publisher Full Text

Roth C: **Why teach Research Ethics, with contribution from Michael Kalichman and Dena Plemmons**. Resources for Research Ethics Education, 2002.
Reference Source

Sarauw LL, Degn L, Ørberg JW: **Researcher development through doctoral training in research integrity.** *Int J Acad Dev.* 2019; **24**(2): 178–191.
Publisher Full Text

Shaw D: **The Quest for Clarity in Research Integrity: A Conceptual Schema.** *Sci Eng Ethics.* 2019; **25**(4): 1085–1093.
PubMed Abstract | **Publisher Full Text**

Sørensen MP, Ravn T, Marušić A, *et al.*: **Strengthening research integrity: which topic areas should organisations focus on?** *Humanit Soc Sci Commun.* 2021; **8**: 198.
Publisher Full Text

Steneck NH: **Global Research Integrity Training.** *Science.* 2013; **340**(6132): 552–553.
Publisher Full Text

Steneck NH: **Fostering integrity in research: Definitions, current knowledge, and future directions.** *Sci Eng Ethics.* 2006; **12**(1): 53–74.
Publisher Full Text

Stephens JM, Watson PWSJ, Alansari M, *et al.*: **Can Online Academic Integrity Instruction Affect University Students' Perceptions of and Engagement in Academic Dishonesty? Results From a Natural Experiment in New Zealand.** *Front Psychol.* 2021; **12**: 569133.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Sun YC: **Using a two-tier test in examining Taiwan graduate students' perspectives on paraphrasing strategies.** *Asia Pacific Education Review.* 2009; **10**(3): 399–408.
Publisher Full Text

Tamir P: **An Alternative Approach to The Construction of Multiple Choice Test Items.** *J Biol Educ.* 1971; **5**(6): 305–307.
Publisher Full Text

Treagust DF: **Development and use of diagnostic tests to evaluate students' misconceptions in science.** *Int J Sci Educ.* 1988; **10**(2): 159–169.
Publisher Full Text

Tsui CY, Treagust D: **Evaluating Secondary Students' Scientific Reasoning in Genetics Using a Two-Tier Diagnostic Instrument.** *Int J Sci Educ.* 2010; **32**(8): 1073–1098.
Publisher Full Text

Valkenburg G, Dix G, Tijdink J, *et al.*: **Expanding Research Integrity: A Cultural-Practice Perspective.** *Sci Eng Ethics.* 2021; **27**(1): 10.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Vehviläinen S, Löfström E, Nevgi A: **Dealing with plagiarism in the academic community: Emotional engagement and moral distress.** *High Educ.* 2018; **75**(1): 1–18.
Publisher Full Text

Winarti, Cari, Suparmi: **Development of two tier test to assess conceptual understanding in heat and temperature.** *J Phys Conf Ser.* 2017; **795**: 12052.
Publisher Full Text

World Conferences on Research Integrity: **The Singapore Statement**. 2010.
Reference Source

Yan YK, Subramaniam R: **Using a multi-tier diagnostic test to explore the nature of students' alternative conceptions on reaction kinetics.** *Chem Educ Res Pract.* 2018; **19**(1): 213–226.
Publisher Full Text

Yang DC, Lin YC: **Assessing 10- to 11-year-old children's performance and misconceptions in number sense using a four-tier diagnostic test.** *Educ Res.* 2015; **57**(4): 368–388.
Publisher Full Text

Zollitsch L, Vilter N, Niesel D, *et al.*: **P2I Questionnaire Version MY.** (Version 2.1.0). *Zenodo.* 2020a.
http://www.doi.org/10.5281/zenodo.4682660

Zollitsch L, Vilter N, Niesel D, *et al.*: **P2I Questionnaire Version S.** (Version 2.1.0). *Zenodo.* 2020b.
http://www.doi.org/10.5281/zenodo.4682616

Zollitsch L, Vilter N, Niesel D, *et al.*: **Draft of the P2I questionnaire.** (Version 1.0.0). *Zenodo.* 2019a.
http://www.doi.org/10.5281/zenodo.5059949

Zollitsch L, Vilter N, Niesel D, *et al.*: **Pre-Version of the P2I questionnaire.** (Version 1.0.0). *Zenodo.* 2019b.
http://www.doi.org/10.5281/zenodo.5059965

Open Peer Review

Current Peer Review Status: ✔ ?

Version 2

Reviewer Report 07 November 2022

https://doi.org/10.21956/openreseurope.16077.r29599

© **2022 Widdershoven G et al.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Guy Widdershoven
Amsterdam UMC, Amsterdam, The Netherlands
Natalie Evans
Amsterdam UMC, Amsterdam, The Netherlands

Overall, the previous comments have been taken well into account. One point which has not been addressed adequately is a check of the English language. The development of the test and the validation procedure are now in general clearly described, and mostly well founded. We do have several comments. Most of them are minor editorial suggestions. Yet, some are more fundamental.

Abstract:

- Change 'training sessions' in 'training programmes' (a training is usually more than 1 session).
- The following sentence is unclear: "In it, testees choose a practice in line with research integrity to address an issue, justify their choice, and describe how confident they are with their answers." We suggest changing this sentence, by clarifying that the testees are invited to make a choice between 4 alternative actions, and choose the option that in their view is in line with the principles of RI, as laid down in the ECoC.
- Elsewhere also change 'practice' in 'action'.

Introduction:

- What does the following sentence mean: "RI and RCR tend to overlap and blur into each other in practice"? Please clarify.
- "RI describes therefore the framework, such as the ECoC (2017)"; RI does not describe a framework; the ECoC is a framework for describing principles of RI and good research practices.
- "we use the term RCR training to refer to training in the field of RI and RCR"; this is a clear

choice, yet, it has not been done everywhere (RI/RCR training is also used).

Status quo and gaps

- "RCR training aims to enable participants to become a part of the research community"; the aim of RCR training is not community building but develop knowledge and skills of participants regarding RCR.

- The statement "a responsible (future) researcher can argue in favour of practices in line with RI" does cover only one aspect (an important one, and as such rightly targeted by P2I training); more generally, RCR trainings aim to develop awareness, knowledge and skills needed for RCR.

- You say: "one practice can simultaneously be both in line and not in line with RI. For example, the practice to publish many articles can on the one hand, be driven by the purpose to get promoted and therefore have enough money to feed one's family; and on the other hand, could be driven by the purpose to timely inform others about one's findings. The former argument could lead to a salami publication tactic, whereas the latter does not run the same risk. How somebody justify their practices is crucial." This raises questions. First, a practice is not right or wrong depending on the motivation alone. A practice can still be an example of salami slicing, and thus not be justified, if the motivation is to inform others. Second, and more fundamentally: if there are two conflicting values at stake (such as career versus informing others), there is often a dilemma involved, meaning that whatever one chooses, harm is done. So there often is no simple right choice. The P2I instrument has a limitation in that it presupposes there is always one right answer, and does not take into account the experiences of researchers regarding dilemmas in their research practice. This is not a problem for the instrument, as long as it is acknowledged that not all aspects of RCR (and its problems) are addressed.

- "Participants show how responsible they are by choosing one option on each item representing the ECoC's topics of RI." Participants do not show how responsible they are (as responsibility is an aspect of (research) action, not of making a choice), but how knowledgeable they are about principles and good practices outlined in the ECoC.

Methods

- "This means that we focussed and limited the P2I questionnaire to what practices testees choose and how they argue for this practice in line with RI." In this sentence the topic of confidence is missing.

- Figure 1 mentions 'scientific and non-scientific patterns', this needs correction.

- Under Methods, also results are presented. It would be clearer to indicate where results begin by adding a paragraph title: Results. This could be after the Ethics statement. (NB: the subtitle "Results" is positioned much later, but this might better be changed)

Step 1

- "To identify research integrity topics in RCR training we followed Roth (2002)"; it is unclear how the (long) quote of Roth helped identifying topics; this quote seems not relevant in step 1.

- The references to Jordan, Steneck and Habermas are relevant in clarifying the importance of justifications; yet, they are only mentioned, and the argument for justification is not well

elaborated.

- The ECoC does not display a discourse in the sense of Habermas, it does provide a frame of reference for identifying good research practices.

Step 2

- Information about the content of the focus group discussions is missing. It is therefore unclear how pregiven scenarios could lead to attractive practices in line and not in line with RI. Were the scenarios adapted based on the focus groups results? How were practices not in line developed; were participants asked to provide examples of practices both in line and not in line?

- “In a second step, we analysed common justifications”; this is within step 2, so it would be better not to say it in this way, but for instance use “Next, we analysed”

- It seems that the answers ‘not in line with RI’ were not thematized in the focus groups, but were the result of analysis by the researchers; this can be formulate more clearly (see comments above); also: how did the researchers come to such evaluations of the discussions in the focus groups? Was the evaluation afterwards mentioned during informed consent of participants? (It is not common to judge answers in focus groups from an external perspective)

- The same question can be raised regarding the questionnaires: were respondents informed about the judgment of their answers?

- Table 2 presents one justification in line with the ECoC, the others are said not to be in line. Yet, the justification focusing on duty might also be in line with the ECoC. It is also not immediately clear that ‘equal to all’ is against the ECoC.

- “we use the term scientific common sense (SCS) when the statement is in line with the RI rules and principles from the ECoC”; it is not clear why this is called ‘common sense’; actually the ECoC aims to do more than codifying common sense.

- “We take into account that statements other than our predefined SCS statements could also be correct and expected in specific settings other than our RI-setting.”; how is this taken into account?

Step 3

- The title of step 3 does not cover the testing of the instrument, which is part of this step.

- “we only selected courses in which participants had not already completed the test.”; the test was not completed by anyone before; maybe you mean: who had not participated in the focus groups or filled in the questionnaire?

- “Item cd9 received the same number of answers in line with RI in the first tier and the second tier.”; is this relevant?

- “Only in two items 72.73% of the participants choose their answer in line with RI in the second tier, while in the other four items only 54.55% of the participants choose their answer in line with RI in the second tier. For this reason we readjusted the item scenario as

well as tier one." It is unclear why a (low?) number of right answers would require a readjustment of the scenario (and the choice options); also: what exactly did the change entail?

- "After revising the pilot test"; do you mean: after revision of the scenario based on the pilot test?

- "Nevertheless, this test run exemplified that the complete questionnaire design, as presented in the P2I Questionnaire S, Vers. 1.0 (Zollitsch *et al.*, 2020b) works."; what does 'work' mean in this context? Feasibility?

- After the pre/post test, changes were made; again the question: why does a (low?) number of right answers require a readjustment of the scenario (and the choice options); also: what exactly did the change entail?

- "A third expert comment asked to give more than one response option. "; what is meant here? One right option?

- "As there were no indications that other RI and RE experts would come to a fundamentally different conclusion"; what could these indications be? How do you know this?

Results

- See earlier comment about the position of the heading 'Results'

- Figure 4 is unreadable, even if we try to enlarge it.

- "the P2I questionnaire also contains different independent variables"; the term independent variables is not adequate, they are data about the respondents (which could be used as independent variables in an analysis)

Discussion

- "enable participants to become a part of the research community."; see comment above

- "It is therefore suitable for RCR training that focuses on dialogical approaches"; it is unclear why this is mentioned here; also, it will depend on the learning goals of such a training. If these for example focus on virtues, the instrument will not be very helpful.

Limitations

- See our comment above: the P2I instrument has a limitation in that it presupposes there is always one right answer, and does not take into account the experiences of researchers regarding dilemmas in their research practice. This is not a problem for the instrument, as long as it is acknowledged that not all aspects of RCR (and its problems) are addressed.

- Another limitation is the presupposition that experts know more about RI than (trained) students. We addressed this in our previous review, and the authors agree in their response. We propose mentioning this under limitations.

- The need for further validation can be addressed more clearly under limitations. The potential design mentioned in the response to our previous review is interesting, but not fully convincing. How can results from groups discussions be compared to later test results? And would the respondents not be influenced by the previous group discussion? Maybe the

other way around (first doing the test individually, afterwards a group discussion) is a better way.

Competing Interests: No competing interests were disclosed.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Version 1

Reviewer Report 09 December 2021

https://doi.org/10.21956/openreseurope.14409.r27805

© 2021 Widdershoven G et al. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Guy Widdershoven
[1] Amsterdam UMC, Amsterdam, The Netherlands
[2] Amsterdam UMC, Amsterdam, The Netherlands
Natalie Evans
[1] Amsterdam UMC, Amsterdam, The Netherlands
[2] Amsterdam UMC, Amsterdam, The Netherlands

Outline of the article

Background and aim
Research integrity training aims to foster responsible conduct of research by providing students with knowledge and skills needed to uphold research integrity principles, as outlined in, for instance, the European Code of Conduct (ECoC). An important question is whether research integrity training actually leads to better knowledge and skills of students, and ultimately contributes to better research practice. This article presents an instrument for evaluating research integrity training. The instrument was developed in the Path2Integrity (P2I) project.

The instrument focuses on assessing understanding of and argumentation for research integrity as conceptualized in the ECoC. In the literature, this is called a two-tier approach, which entails a combination of assessing whether the student gives the right answer (first tier) and provides an adequate reason for the answer (second tier). In the course of the development, two more tiers were added, addressing the student's confidence. This results in four tiers, addressing whether the answer is right (tier 1), whether the student feels confident that the answer is right (tier two), whether the reasoning is right (tier three), and whether the student is confident that the reasoning is right (tier four).

Development of the instrument
After a study of relevant literature, the instrument was designed by deducing good actions and relevant justificatory arguments from the ECoC. This resulted in what the authors call 'scientific' actions and arguments. Wrong, or 'non-scientific' actions and justificatory arguments were derived from focus groups with students. The instrument was pilot tested and improved based on the results of the tests.

Instrument
The content of the evaluation instrument mirrors the central domains of the ECoC. For various domains (or best practices) in the ECoc, such as research environment, safeguards, collective working, and publication and dissemination, a short case is presented, followed by a multiple-choice question about which behaviour would be right (tier one, one out of four options), and a multiple-choice question about the underlying principle justifying the right course of action (tier three, again one out of four options). Thus, the instrument measures, for relevant domains of the ECoC, whether the student knows what is the right action, and what principle underlies the action.

Validation
The instrument was validated through a content validation procedure in which students and experts commented upon items. Comments were examined, to see whether the items needed to be changed. The instrument was also tested by measuring results before and after training, and differences between experts and students. Results of the tests were used to improve the instrument.

Comments

1. The goal of the paper is important and the steps taken are relevant. Yet, the structure of the paper is not fully clear, and the aim and the steps taken are described in various, not always congruent ways. Concepts are introduced without explanation. Some core terms are not common, which makes it not easy to follow the argument. Also, the text is difficult to read, with misspellings and sentences which do not flow naturally. Overall, the paper would benefit from editing and a language check.

2. In the introduction, the authors state that most tools for assessing learning successes of RI/RCR are qualitative. However later in the article they address some validated measures which have been used (Academic Motivation and Integrity Survey (AMIS) and the Academic Integrity Self-Evaluation Tools) but miss out other measures which are specific to RI/RCR and suitable for pre-post test (e.g. The professional decision-making in research measure - Dubois et al 2016). Discussing these instruments, and their limitations, in the introduction would provide a clearer background for the need of a new instrument.

3. The authors treat RI and RCR as synonymous concepts. An explanation of the concepts when first used and their relationship would be helpful.

4. The procedures described in the article are based on literature on evaluation. More explanation of core concepts in the literature, for instance the two-tier approach, would make the paper more accessible for the reader.

5. The steps in the development of the instrument are relevant. Yet, the terms referring to the

steps are not always adequate:

1. In the Abstract and the Introduction, the first step (study of the literature on evaluation) is inadequately named “describing successes of RI training”.
2. The second step, deduction of good actions and relevant justificatory principles from the ECoC, is described as “identification of scientific and non-scientific patterns”. The terms ‘scientific’ and ‘non-scientific’ are problematic, since right actions are not *per se* scientific, and wrong actions not *eo ipso* non-scientific. This is further complicated by the authors’ own statement that “the ECoC justification pattern is called scientific justification pattern taking into account that other justifications patterns can also be scientifically.” Also, it is not clear what the term ‘pattern’ entails. It would be better to say: “identification of actions and justifications which are in line with RI rules and principles” and “identification of actions and justifications which are not in line with RI rules and principle”.
3. In the third step, the instrument is said to be “validated and revised”. Since the results of the content validation procedure and the tests are used to improve the instrument, the last version is not validated. So the term ‘pre-validation’ (used also in the article) is more adequate.

6. In describing the instrument, the authors refer to multiple-choice questions which measure knowledge, and “priority rankings which can evaluate norms and values”. Yet, the instrument contains only multiple-choice questions for measuring knowledge about right actions and knowledge about underlying principles. Priority rankings are not involved; also it is unclear what kind of evaluation of norms and values such rankings would entail.

7. The instrument has a clear and well-considered structure. Taking the ECoC as a point of reference, the student is assessed on the knowledge about what actions are right, and also on their knowledge of the principles that make the action right. This approach is attractive, as it is based upon an important and elaborated document for the European research community. Yet, this also entails a limitation, as the instrument is -at the moment- restricted to the principles and best practices presented in the code. Other principles and practices can also be relevant for research integrity teaching and good research (see for instance the Themes section of The Embassy of Good Science).

8. The authors state that “true or false answers or specific cognitive abilities are not the focus of this questionnaire.” This is incorrect. The instrument focuses on cognitive skills. The first tier addresses knowledge of rules of good conduct. The third tier assesses the student’s knowledge of the foundation of the rules in principles. Although the authors acknowledge that the goal of RI/RCR training includes that students “be motivated to value scientific practices above others”, motivation is not assessed by the instrument. By examining knowledge of rules and underlying principles, the instrument does not show whether these rules and principles are actually endorsed by the students themselves. When a student says that a certain action is right because of, for instance, transparency, this does not show that the student him- or herself values transparency as a precondition for good research. Also, it does not show that the student would deem it important to realize transparency in their own research. The instrument is not able to assess the student’s attitude and character building. For research integrity training to result in better research practice, not only knowledge of codes and identification of underlying principles is needed, but also the development of virtues (Pennock, 2018). It is not easy to develop an instrument which can

evaluate this, but for measuring the impact of a training on students in their work as (future) researchers, we need to go beyond assessing their knowledge of best practices and principles. An example of an instrument which aims to measure knowledge, skills and attitudes in the domain of clinical ethics is the Euro-MCD (De Snoo-Trimp *et al*, 2020).

9. In the development processes there are some limitations which should be acknowledged. Namely the lack of participation of the target audience (secondary school students) in the development and piloting processes and the lack of diversity in disciplinary field of participant researchers in the whole development process.

10. A content validity procedure, including expert comments, is important. Yet, the number of expert comments which resulted in changes is low. This may indicate that the instrument was of good quality. However, it might also mean that the procedure did not foster in-depth investigation and interpretation of the instrument.

11. The tests presented in the article contribute to validation by investigating whether the instrument detects differences between groups which can be expected to have different results. The assumption is that experts score higher than students on confidence, and that student after training score higher than before. Given that a gold standard is absent, this is a useful approach (Van Melle *et al*, 2019). Yet, it also contains a circularity, since the assumption is that experts know more, and that training improves knowledge. This is a limitation of the study which deserves attention.

12. As already mentioned, the validation procedures are used to improve the instrument. This means that the final instrument still needs to be validated, as the authors acknowledge. It would be nice to hear more about the way in which the authors want to validate the instrument in the future.

References

1. DuBois JM, Chibnall JT, Tait RC, Vander Wal JS, et al.: Professional Decision-Making in Research (PDR): The Validity of a New Measure.*Sci Eng Ethics*. 2016; **22** (2): 391-416 PubMed Abstract | Publisher Full Text
2. van Melle AL, Voskes Y, de Vet HCW, van der Meijs J, et al.: High and Intensive Care in Psychiatry: Validating the HIC Monitor as a Tool for Assessing the Quality of Psychiatric Intensive Care Units. *Adm Policy Ment Health*. **46** (1): 34-43 PubMed Abstract | Publisher Full Text
3. Pennock R.T.: Beyond research ethics: How scientific virtue theory reframes and extends responsible conduct of research. In D. Carr (ed.) Cultivating Moral Character and Virtue in Professional Practice. *Routledge.*2018. 166-177
4. de Snoo-Trimp J, de Vet H, Widdershoven G, Molewijk A, et al.: Moral competence, moral teamwork and moral action - the European Moral Case Deliberation Outcomes (Euro-MCD) Instrument 2.0 and its revision process. *BMC Medical Ethics*. 2020; **21** (1). Publisher Full Text

Is the rationale for developing the new method (or application) clearly explained

Yes

Is the description of the method technically sound?

Partly

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Ethics, Reserach Integrity, Evaluation research, Training and education

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 07 Jun 2022

Linda Zollitsch

Reviewer Comments in italics
The goal of the paper is important and the steps taken are relevant. Yet, the structure of the paper is not fully clear, and the aim and the steps taken are described in various, not always congruent ways. Concepts are introduced without explanation. Some core terms are not common, which makes it not easy to follow the argument. Also, the text is difficult to read, with misspellings and sentences which do not flow naturally. Overall, the paper would benefit from editing and a language check.
Answer from the authors: Based on the comments below, we revised the article and clarified core terms and concepts. In addition, we have had the article proofread once again.

In the introduction, the authors state that most tools for assessing learning successes of RI/RCR are qualitative. However later in the article they address some validated measures which have been used (Academic Motivation and Integrity Survey (AMIS) and the Academic Integrity Self-Evaluation Tools) but miss out other measures which are specific to RI/RCR and suitable for pre-post test (e.g. The professional decision-making in research measure - Dubois et al 2016). Discussing these instruments, and their limitations, in the introduction would provide a clearer background for the need of a new instrument.
Answer from the authors: The article focuses on assessing learning successes in research integrity. Besides qualitative tools to assess RI/RCR training learning successes, there are also quantitative tools. Our desk research showed that it is common to adapt tools from ethics education for this purpose. Thank you for pointing out how important the PDR

(Professional Decision Making in Research) (DuBois et al. 2016) is in evaluating decision making in research. We acknowledge that and added to our text: “Another instrument, called the Professional Decision Making in Research (DuBois et al. 2016) was developed based on the Ethical Decision Making (Mumford et al. 2006). This assessment instrument focusses (not on justifications but) on decisions and “ethical problems that include factors such as incomplete knowledge, power discrepancies, and urgency – factors that may interfere with ethical decision making” (DuBois, 2016). Also, the Academic Motivation and Integrity Survey, as for example presented in Stephens et al. (2021), and the Academic Integrity Self-Evaluation Tools, as reported in Gaižauskaitė et al. (2020) cover the field of research integrity. They all focus on ethical decisions and behaviour. In distinction to these tools – which concentrate on ethical decisions and behaviour – our approach assesses, if (future) researchers argue in favour of research integrity. With this approach we ensure that the assessment reveals what practice in line with RI people choose in the research community and how they justify this practice. Participants show how responsible they are by choosing one option in each item representing the ECoC’s topics of RI.” (p. 3)

The authors treat RI and RCR as synonymous concepts. An explanation of the concepts when first used and their relationship would be helpful.
Answer from the authors: To explain concepts more in detail, we added: "We acknowledge that RI and RCR tend to overlap and blur into each other in practice, however we align with the European Code of Conduct for Research Integrity (2017). As such we describe RI for this project as a commitment to professional, legal, and ethical responsibilities, values, and principles to self-regulate research (cf. ECoC 2017, 3ff.). RCR on the other hand is “simply conducting research in ways that fulfil the professional responsibilities of researchers, as defined by their professional organizations, the institutions for which they work and, when relevant, the government and public” (Steneck 2006, 55), in particular “RCR education can help science take care of itself” (Roth, 2002).

The procedures described in the article are based on literature on evaluation. More explanation of core concepts in the literature, for instance the two-tier approach, would make the paper more accessible for the reader.
Answer from the authors: We derived the idea and procedure for the P2I questionnaire from the literature on evaluation, especially the literature regarding the two-tier test. The authors have edited the paper to give more insight into the two-tier approach and its core concepts: "Treagust (1988) designed the original structure of the two-tier test as an extension of classic multiple-choice tests to determine not only what students know but also how students understand a specific topic. This two-tier approach can help differentiate between multiple choice answers which are right by chance and which are right because there is an understanding of the concepts. Treagust's original two-tier test is a two part (a.k.a. tier) multiple-choice series. It starts with a multiple-choice question that the testees respond to in the first tier; this extends to a second-tier where the testees reason their response from the first tier. He based his original design on Tamir's (1971) approach in which distractors were derived from existing misconceptions of the testees – this would make them more attractive. The two-tier test is an extension of classical multiple-choice tests and helps to determine not only what students know but also what rationale underpins that answer." (p. 6).

The steps in the development of the instrument are relevant. Yet, the terms referring to the steps are not always adequate:

In the Abstract and the Introduction, the first step (study of the literature on evaluation) is inadequately named "describing successes of RI training".
Answer from the authors: Thank you for pointing that out. We changed the wording into the following: "The following shows the questionnaire's development process in three steps. After describing the status quo and gaps in evaluating RCR training successes first, this article illustrates how we designed the P2I questionnaire using practices and justifications in line (and not in line) with research integrity in a second step. In a third step, the P2I questionnaire is pre-validated and revised." (p. 3)

The second step, deduction of good actions and relevant justificatory principles from the ECoC, is described as "identification of scientific and non-scientific patterns". The terms 'scientific' and 'non-scientific' are problematic, since right actions are not per se scientific, and wrong actions not eo ipso non-scientific. This is further complicated by the authors' own statement that "the ECoC justification pattern is called scientific justification pattern taking into account that other justifications patterns can also be scientifically." Also, it is not clear what the term 'pattern' entails. It would be better to say: "identification of actions and justifications which are in line with RI rules and principles" and "identification of actions and justifications which are not in line with RI rules and principle".
Answer from the authors: We renamed the second step into: "Identification of practices and justifications which are and which are not in line with RI rules and principles". We also added the following: "In the following, we use the term scientific common sense when the statement is in line with the RI rules and principles from the ECoC. We combine different statements, which are in line with the scientific common sense, in a justification pattern called 'scientific common sense pattern'. Statements from this pattern are distinguishable from other statements, which are not in line with the scientific common sense (SCS) such as the rejection of binding codes or individual benefits (as presented in Table 2). We take into account that other statements than our predefined SCS statements can also be correct and expected in specific settings other than our RI-setting." (p. 6).

In the third step, the instrument is said to be "validated and revised". Since the results of the content validation procedure and the tests are used to improve the instrument, the last version is not validated. So the term 'pre-validation' (used also in the article) is more adequate.
Answer from the authors: Thank you. The term validated is changed into pre-validated. (See p. 3).

In describing the instrument, the authors refer to multiple-choice questions which measure knowledge, and "priority rankings which can evaluate norms and values". Yet, the instrument contains only multiple-choice questions for measuring knowledge about right actions and knowledge about underlying principles. Priority rankings are not involved; also it is unclear what kind of evaluation of norms and values such rankings would entail.
Answer from the authors: The P2I questionnaire has a multi-tier structure. This structure contains: Traditional multiple-choice questions. Multiple-choice questions, which refer to (already) given answers. Sliders to rate how confident the testees are in giving their answers. We see that our example relating to priority rankings is misleading in this context

and changed the sentences into. "Whereas traditional multiple-choice questionnaires often assess knowledge on one tier, a multiple-tier structure can give insights into students' justifications. That is why we decided to use a multiple tier structure for the P2I questionnaire." (See p. 3).

The instrument has a clear and well-considered structure. Taking the ECoC as a point of reference, the student is assessed on the knowledge about what actions are right, and also on their knowledge of the principles that make the action right. This approach is attractive, as it is based upon an important and elaborated document for the European research community. Yet, this also entails a limitation, as the instrument is -at the moment- restricted to the principles and best practices presented in the code. Other principles and practices can also be relevant for research integrity teaching and good research (see for instance the Themes section of The Embassy of Good Science).
Answer from the authors: The P2I questionnaire focuses on the ECoC as the reference document for Europe. As written in the discussion, the P2I questionnaire "is expandable and applicable to other reference documents such as the Singapore Statement, 2010, the Montreal Statement, 2013 etc. These codes refer to the same practices in line with RI and entail similar justifications. The P2I questionnaire focuses on its assessment on practices and justifications in favour of research integrity. It is therefore suitable for RCR training that focuses on dialogical approaches (Koterwas et al. 2021)." (See p. 11).

The authors state that "true or false answers or specific cognitive abilities are not the focus of this questionnaire." This is incorrect. The instrument focuses on cognitive skills. The first tier addresses knowledge of rules of good conduct. The third tier assesses the student's knowledge of the foundation of the rules in principles. Although the authors acknowledge that the goal of RI/RCR training includes that students "be motivated to value scientific practices above others", motivation is not assessed by the instrument. By examining knowledge of rules and underlying principles, the instrument does not show whether these rules and principles are actually endorsed by the students themselves. When a student says that a certain action is right because of, for instance, transparency, this does not show that the student him- or herself values transparency as a precondition for good research. Also, it does not show that the student would deem it important to realize transparency in their own research. The instrument is not able to assess the student's attitude and character building. For research integrity training to result in better research practice, not only knowledge of codes and identification of underlying principles is needed, but also the development of virtues (Pennock, 2018). It is not easy to develop an instrument which can evaluate this, but for measuring the impact of a training on students in their work as (future) researchers, we need to go beyond assessing their knowledge of best practices and principles. An example of an instrument which aims to measure knowledge, skills and attitudes in the domain of clinical ethics is the Euro-MCD (De Snoo-Trimp et al, 2020).
Answer from the authors: Thank you for describing your view on this section. The P2I questionnaire "assesses in which way testees argue in research settings" (original article p. 4). We also pointed out that "true or false answers or specific cognitive abilities or strategies are not the focus of this questionnaire". Why do we point that out in our article? Because multiple-choice tests typically assess these. By reading your description, we realized that we failed in describing the concept of the P2I questionnaire. We designed the P2I questionnaire to help trainers support students' learning successes. In assessing how students argue in favour of scientific practice and how they justify their scientific practice, we learn what

students describe as a scientific practice, how sure they are about this, how they justify their scientific practice and how sure they are with their justification. Having this knowledge is a great advantage when you teach students to conduct responsible conduct of research. To clarify our thoughts, we added: “The instrument aims to assess what practice testees would choose in research settings and how they argue in favour of these practices. Practice in this sense means a concrete action, what to do to follow good research practice. The questionnaire also should observe how confident testees are in their answers. This helps to identify whether an answer was given due to strong believe or due to uncertainty. Trainers should be able to see both, how confident testees choose practices in line or not in line with RI and how confident they justify this practice to be in line with RI. This means that we focussed and limited the P2I questionnaire to what practices testees choose and how they argue for this practice in line with RI.” (See p. 4) Even though virtues are relevant for RI/RCR, the P2I questionnaire does not assess these factors. To make this limitation more visible, we added: "The P2I questionnaire focuses on its assessment on practices and justifications in favour of research integrity. It is therefore suitable for RCR training that focuses on dialogical approaches." (p. 11)

In the development processes there are some limitations which should be acknowledged. Namely the lack of participation of the target audience (secondary school students) in the development and piloting processes and the lack of diversity in disciplinary field of participant researchers in the whole development process.
Answer from the authors: The P2I questionnaire, as well as its development, have some limitations. We added the following part to the limitation section to address them more appropriately: “Limitation of the pre-validation and testing: No secondary school students designed and pre-validated the P2I questionnaire. The leading group of involved participants – next to the 27 responses of international experts – affiliates to educational studies, and most of them identified their country as Germany.” (See p. 11.)

A content validity procedure, including expert comments, is important. Yet, the number of expert comments which resulted in changes is low. This may indicate that the instrument was of good quality. However, it might also mean that the procedure did not foster in-depth investigation and interpretation of the instrument.
Answer from the authors: We acknowledge that this procedure is not in line with the concept of reproducibility. Our goal is opening the discourse on the four-tier test as quantitative instrument for measuring RI by presenting a transparent design process. Therefore, your interpretation is possible. On the other hand - in a not entirely irony-free way – your comment implies specific negligence of the RI experts, which we do not expect. Nevertheless, the final validation will identify such errors and ensure a reproducible procedure.

The tests presented in the article contribute to validation by investigating whether the instrument detects differences between groups which can be expected to have different results. The assumption is that experts score higher than students on confidence, and that student after training score higher than before. Given that a gold standard is absent, this is a useful approach (Van Melle et al, 2019). Yet, it also contains a circularity, since the assumption is that experts know more, and that training improves knowledge. This is a limitation of the study which deserves attention.

Answer from the authors: Yes, we acknowledge that fact. We designed example practices and justifications not in line with research integrity from answers students gave us and derived practices and justification in line with research integrity from the ECoC. We assume that experts for RI are more in line with RI practices and justification than students. We acknowledge that the pre-validation can be seen as circular, and therefore, it is necessary to develop it further. We hope to receive different replies and reactions from the different fields of expertise related to RI by publishing this article.

As already mentioned, the validation procedures are used to improve the instrument. This means that the final instrument still needs to be validated, as the authors acknowledge. It would be nice to hear more about the way in which the authors want to validate the instrument in the future.
Answer from the authors: Since the article describes the pre-validation of the P2I questionnaire, this needs still validation. Such validation is complex, and we didn't present it within the article.
Briefly outlined, the final validation step looks as follows: The test aims to identify specific patterns of practices and justifications in research environments. Therefore, we will ask representatives of the different target groups to assess the cases described in the items in group discussions. Subsequently, the same persons will complete the test. In the analysis, we will compare the persons' statements from the group discussions with the multiple-choice answers from the questionnaire. If the results about scientific practice and justifications match, we assume that the test provides valid information about persons arguing in favour of research integrity.

References:

- de Snoo-Trimp, J.C., de Vet, H.C.W., Widdershoven, G.A.M., Molewijk, A.C., Svantesson M: Moral competence, moral teamwork and moral action - the European Moral Case Deliberation Outcomes (Euro-MCD) Instrument 2.0 and its revision process. BMC Med Ethics 21(1):53 (2020). doi: 10.1186/s12910-020-00493-3.
- DuBois, J.M., Chibnall, J.T., Tait, R.C. et al: Professional Decision-Making in Research (PDR): The Validity of a New Measure. Sci Eng Ethics 22, 391–416 (2016). https://doi.org/10.1007/s11948-015-9667-8.
- Kaltakci-Gurel, D., Eryilmaz, A., McDermott L.C: Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics. Research in Science & Technological Education. 35:2 (2017). 238-260. DOI: 10.1080/02635143.2017.1310094.
- Koterwas, A., Dwojak-Matras, A., and Kalinowska, K: Dialogical teaching of research integrity: an overview of selected methods. FACETS. 6(2021): 2138-2154. https://doi.org/10.1139/facets-2021-0045.
- Pennock, R.T: Beyond research ethics: How scientific virtue theory reframes and extends responsible conduct of research. In Cultivating Moral Character and Virtue in Professional Practice, 166-177 (2018). Routledge.
- van Melle A.L., Voskes, Y., de Vet, H.C.W., van der Meijs, J., Mulder, C.L., Widdershoven, G.A.M: High and Intensive Care in Psychiatry: Validating the HIC Monitor as a Tool for Assessing the Quality of Psychiatric Intensive Care Units. Adm Policy Ment Health 46(1), 34-43 (2019). doi: 10.1007/s10488-018-0890-x.

Competing Interests: No competing interests were disclosed.

Reviewer Report 14 September 2021

https://doi.org/10.21956/openreseurope.14409.r27517

© 2021 Pan S. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Sophia Jui-An Pan
[1] Research Center for Humanities and Social Sciences (RCHSS), National Yang Ming Chiao Tung University, Hsinchu, Taiwan
[2] Research Center for Humanities and Social Sciences (RCHSS), National Yang Ming Chiao Tung University, Hsinchu, Taiwan

This manuscript does a great job developing an assessment instrument for research integrity (RI) training. In general, I believe that the study is sound. The title and abstract are appropriate for the content of the article. In addition, the manuscript is well constructed, the analyses are well conducted and presented in the tables and figures. Each version of the P2I questionnaire has been released on Zenodo. Overall, this work is a good reminder for RI researchers and instructors to ensure that they use practical standardized instruments for evaluating the success of RI education. I have a few suggestions to improve the manuscript, as follows:

- **p. 3 –** The study is off to a good start. The authors indicate that the P2I questionnaires (MY/S) are intended as a first step in the development of standardized research integrity measurements. Therefore, it would be useful to present, maybe in the Introduction, the procedures that the authors will employ to finalize the standardized instruments, so that others could replicate the process and make their own assessments. Also, this improvement will likely provide the audience with a basic understanding of the developmental process at the beginning of the article, as well as enhance their understanding as to how the current study has contributed to – and what the role of the study is in – the developmental process.

- **p. 3, Table 1 –** I did not see the necessity for Table 1 since the information included in the table is limited. The authors might simply relate the information in an appropriate paragraph.

- **p. 8 –** What scoring method was employed for the P2I questionnaire? Have the item discrimination and difficulty been determined?

- **p. 10, Figure 3 –** The authors employed a rating scale of 1-very unsure to 100-very confident to determine how confident the subjects were with their decisions and justifications. I am curious about why the authors applied this rating method instead of the commonly-used 5 or 7-Likert scale (i.e., 1-very unsure to 5/7-very confident).

- **p. 11, Figure 4 –** The P2I questionnaire asked the subjects to select one of the four response options in both the decision and justification tiers; more than one selection on each tier is not permitted. However, there is always more than one good way to respond to RI-relevant issues; the decision-makers must make a fair choice among all of the good options, which can be challenging. In this vein, is it possible that the "good research practices" that the subjects would like to use to respond to the given scenarios go beyond the four options, making it difficult to make a choice reflecting their authentic understanding of RI? If so, how can this situation be resolved? This should be considered when revising the questionnaires.

Is the rationale for developing the new method (or application) clearly explained
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: research integrity; curriculum design, development, and assessment

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 07 Jun 2022

Linda Zollitsch

p. 3 – *The study is off to a good start. The authors indicate that the P2I questionnaires (MY/S) are intended as a first step in the development of standardized research integrity measurements. Therefore, it would be useful to present, maybe in the Introduction, the procedures that the authors will employ to finalize the standardized instruments, so that others could replicate the process and make their own assessments. Also, this improvement will likely provide the audience with a basic understanding of the developmental process at the beginning of the article, as well as enhance their understanding as to how the current study has contributed to – and what the role of the study is in – the developmental process.*
Answer from the authors: For the better orientation of the readers, the following sentence

was added to the introduction: “By making each design step transparent, we open a discourse to reach a standardised instrument at the end, quantitatively assessing learning successes in RCR training.” Furthermore, the article is indeed intended as an opening of discourse that hopes for critical-constructive feedback from the scientific community. It must always be kept in mind that this instrument was developed in the context of the evaluation of P2I learning units and that during the development process we realised that there might be a potential in the instrument to transfer the questionnaire or at least the test structure to other RI trainings. And we want to openly discuss this potential without already having the answer.

p. 3, Table 1 – *I did not see the necessity for Table 1 since the information included in the table is limited. The authors might simply relate the information in an appropriate paragraph.*
Answer from the authors: We deleted table 1 and added the following part: “However, a commonly research integrity challenge is, that one practice can simultaneously be both in line and not in line with RI. For example, the practice to publish many articles can on the one hand be driven by the purpose to get promoted and therefore have enough money to feed one’s family; and on the other hand, could be driven by the purpose to timely inform others about one’s findings. The former argument could lead to a salami publication tactic, whereas the latter does not run the same risk. How somebody justifies their practices is crucial.”

p. 8 – *What scoring method was employed for the P2I questionnaire? Have the item discrimination and difficulty been determined?*
Answer from the authors: In the course of test development, we counted the preferred responses in tier 1 and tier 3 independently of each other. For the analysis of the evaluation results, however, there is an evaluation matrix in which all combination possibilities as well as the confidences are taken into account. The item difficulty results directly from the solution probabilities in the results shown. It can be seen that the item difficulty is very heterogeneous depending on the item, i.e. the content area. The underlying RI construct in tier 1 refers to the 8 content fields covered in the ECoC. Although these fields add up to RI, they are not necessarily interdependent. An item with negative discriminatory power does not necessarily have to be a bad item, but can rather indicate that deficits in certain content areas occurred particularly frequently. In tier 3, the situation is different, as the same question is always asked here, which is aimed at scientific justification ability. As soon as a negative discriminatory power is detected, this would not lead to the deletion of the item, but to a distractor analysis.

p. 10, Figure 3 – *The authors employed a rating scale of 1-very unsure to 100-very confident to determine how confident the subjects were with their decisions and justifications. I am curious about why the authors applied this rating method instead of the commonly-used 5 or 7-Likert scale (i.e., 1-very unsure to 5/7-very confident).*
Answer from the authors: We acknowledge that in most literature about the multiple-tier test, a Likert scale is used. We decided otherwise since a scale from 1 to 100 is continuous and gives the testees the opportunity to place their confidence on it, while a Likert scale is designed in steps (1 very unsure, 2 unsure, 3 something in the middle, 4 confident, 5 very confident). But we cannot say for sure that the distance from 1 to 2 is the same as from 2 to 3 – which might lead to a problem and it might be possible that participants chose 4 on the

Likert scale in the pre test and also 4 in the post-test, as they might not be very confident, but we cannot identify if the confidence if 4 means the same in both tests, or if they might have a little bit more confidence in the post test. The decision was thus based on the consideration of being able to identify even small differences or shifts between the independent variables, such as pre- or post-test, gender or culture, etc.

p. 11, Figure 4 – *The P2I questionnaire asked the subjects to select one of the four response options in both the decision and justification tiers; more than one selection on each tier is not permitted. However, there is always more than one good way to respond to RI-relevant issues; the decision-makers must make a fair choice among all of the good options, which can be challenging. In this vein, is it possible that the "good research practices" that the subjects would like to use to respond to the given scenarios go beyond the four options, making it difficult to make a choice reflecting their authentic understanding of RI? If so, how can this situation be resolved? This should be considered when revising the questionnaires*.
Answer from the authors: It is possible that testees would decide on a different action or a different justification when they would have been asked to give an open text response (for example). We decided to not give this option, since the P2I questionnaire was designed to measure if the participants know the good research practice and can argue for this in a way that is common in the scientific community. Nevertheless, we have used free text responses when designing the option for the justification. Furthermore, the test does not claim to make statements about authentic understanding of RI. Rather, it is about seeing whether the testees can recognise and decide on scientific patterns of action and justification. To this end, an artificial unambiguity and reduction of decision-making possibilities was applied to the items in several revision loops. This was a great challenge, especially in the discussion with the RI experts, who repeatedly pointed out ambiguities in the first informal item drafts.

Competing Interests: No competing interests were disclosed.