The development of a four-tier test to evaluate research integrity training [version 1; peer review: 1 approved]

Linda Zollitsch, Nicolaus Wilder, Julia Priess-Buchheit

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

Abstract
Although higher education institutions across Europe and beyond are paying more and more attention to research integrity training, there are few studies and little evidence on what works and what does not work in such training. One way to overcome this challenge is to evaluate such training with standardised instruments. Experts/trainers have used qualitative approaches to evaluate their research integrity 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 research integrity’s core themes. At present, there is a lack of standardised instruments designed to specifically evaluate success in research integrity training. This article presents a pre-validated instrument for this purpose. Named the P2I questionnaire, it is designed as a four-tier test and based on the European Code of Conduct for Research Integrity; in it, testees choose a (scientific) practice to address an issue, justify their decision, and describe how confident they are with their decisions. The development of the P2I questionnaire is outlined in three steps. After describing the research integrity (alternatively, the responsible conduct of research) training successes, the article notes scientific and non-scientific patterns and then concludes with a pre-validated and revised version of the P2I questionnaire. This questionnaire is intended as a first step in a discourse on standardised research integrity measurements and is one step towards an evidence-based improvement of research integrity training.

Keywords
research integrity, teaching and learning, evaluation, four-tier test, responsible conduct of research
This article is included in the Science with and for Society gateway.

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: © 2021 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 research integrity training [version 1; peer review: 1 approved] Open Research Europe 2021, 1:90 https://doi.org/10.12688/openreseurope.13339.1

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

In recent decades, research integrity (RI) and responsible conduct of research (RCR) have developed into a fundamental cross-cutting issue in research and 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). This lack is one of the main challenges in fostering a research integrity culture through training. 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 of the tools for assessing the learning successes of RI/RCR training are qualitative evaluations.

This article describes the development of one instrument for this purpose, namely the P2I questionnaire, which evaluates RI/RCR training successes and gives feedback to individual trainers. The P2I questionnaire concentrates on scientific practices and scientific justifications. It assesses the understanding of and argumentation for research integrity as conceptualised in the European Code of Conduct for Research Integrity (ECoC, 2017). The P2I questionnaire aims to assess the learning successes of secondary school students older than 16 up to early career researchers.

The development process for this questionnaire followed three steps. After describing successes of RI/RCR training in a first step, the article illustrates in a second step scientific and non-scientific patterns. In a third step, the P2I questionnaire is validated and revised. The development of the questionnaire recognizes that the instrument’s structures significantly impact what it evaluates. Whereas simple multiple-choice questions and answers can measure knowledge, priority rankings can evaluate norms and values. That is why the P2I questionnaire uses a multiple tier structure, which allows insights into students’ justifications for responsible conduct of research.

In the content validation test, students and experts proofed each item. Some items were ambiguous and have been revised. The result is a pre-validated questionnaire that contains questions about what scientific practices are and how to justify these scientific practices. The collection of scientifically and non-scientifically based answers is one key element in designing the P2I questionnaire. Developing such a questionnaire is a first step towards evaluating RI/RCR training’s successes in a standardised way, considering scientific reasoning and justification. By taking each design step into account, the discourse about standardised quantitative evaluations can be enhanced.

Status quo and gaps in evaluating successes of RI/RCR training

As mentioned above, there are different qualitative instruments 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 tools to assess RI/RCR training’s successes derive mainly from ethics education because of overlapping training content and similar learning methods. Three of the most frequently used tests are the Defining Issues Test (DIT) (Rest, 1979), its revised version, the DIT-2 (Rest et al., 1999), and the sensemaking approach by Mumford and colleagues (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 RI/RCR training, they are derived from ethics education and do not capture the RI/RCR training core theme.

RI/RCR training enables participants to become a part of the scientific community, including “the principles of research, … the criteria for proper research behaviour, [and] … the quality and robustness of research” (ECoC, 3). After RI/RCR training, students can justify scientific practices and question them. What does that mean in concrete terms? Learning RI/RCR entails, how to cite correctly, handle research subjects, and work in collaborative teams etc. In these cases, students need to know what to do, prioritise these practices over other contradicting practices, and be motivated to value scientific practices above others. The contradicting practices/arguments in Table 1 are a typical research integrity challenge, which exemplifies the motivational aspect of research integrity training: Whereas the argument for promotion and family support can lead to a salami publication tactic, the argument to inform others on time about findings, does not.

While assessments in ethics education focus on classical moral dilemmas and reasoning structures, RI/RCR training assessment needs to include the motivational aspect by looking at contradicting practices and incorporating scientific justification (scientific patterns). Whereas some quantitative evaluations such as the Academic Motivation and Integrity Survey (AMIS), as for example presented in Stephens et al. (2021), and the Academic Integrity Self-Evaluation Tools, as reported in Gaiauzkaite et al. (2020) cover the field of research integrity, they do not concentrate on (non-) scientific justifications.

| Why should I publish as many articles as possible? |
|--------------------------------------------------|
| To get promoted and therefore have enough money to feed my family. |
| To inform others on time about my findings. |
Until now, there is no established standardised instrument that covers this facet of RCR/RI training. Nevertheless, some evaluations use a promising structure to assess such scientific justifications and patterns. In 2007 Chou et al., applied the so-called two-tier structure to a research integrity-related topic. They assessed students’ patterns of cyber copyright laws. Also, Sun (2009) examined the paraphrasing strategies with a two-tier, while Pan & Chou (2015) used this structure to assess students’ misunderstandings about ethics and behaviour. These approaches demonstrated that the two-tier test seems to be a good fit for a RI/RCR questionnaire concentrating on (non-) scientific justifications and patterns.

**Methods**

The P2I questionnaire quantifies RI/RCR training’s successes via the amount of well-founded answers. It assesses in which way testees argue in research settings. Thus, true or false answers or specific cognitive abilities or strategies are not the focus of this questionnaire.

The following describes the development of the multi-layered-tier test. As shown in Figure 1, this specific questionnaire’s development follows Chandrasegaran, Treagust, and Mocerino’s (2007) approach and designs, and validates the P2I questionnaire in three steps.

Figure 1 describes the steps from one, defining the content to step two, identifying scientific and non-scientific patterns to result in step three, developing the structure and items to evaluate the RI/RCR training successes.

Step one combines identifying the content, identifying research integrity fields, and adapting the research integrity fields regarding different target groups. Step two unites the

---

**Figure 1.** Scheme of development of the P2I questionnaire adapted with permission of The Royal Society of Chemistry, from Chandrasegaran, Treagust, and Mocerino (2007); permission conveyed through Copyright Clearance Center, Inc.
development of the specific content and patterns for the first tier (scientific practice) answers and the third tier (scientific justification). Step three outlines the development of the structure of the P2I questionnaire, starting with the validation of the questionnaire with the help of experts and refining the questionnaire after the first tests runs.

Ethics statement
All procedures performed in the development, 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 voluntary give data for this study.

Step one: Defining the content
To design an accurate, reliable, and valid questionnaire and to evaluate RCR/RI training successes is not a simple task. It requires considerable knowledge, skill, and experience in question-wording and questionnaire structure on the one hand and the subject area of the evaluation on the other hand. “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” (Roth, 2002).

As mentioned above, participants who successfully participated in RI/RCR training can justify research practices and question them. They justify their research practices with well-founded (rational) arguments. While ethics 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, integrity emphasises the practices of agents who comply with these ethical standards or principles (Löfström & Pyhältö, 2019). So, “research integrity lies in consistency with external rules” (Shaw, 2019, 1087).

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 RI/RCR is not about blindly following rules, (History taught society many times the consequences of such attempts.) Instead, learning RI/RCR can be integrated into what Habermas (1990) and Gethmann (2010) say. Learning RI/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 to by individuals. 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).

The ECoC (2017) names the following scientific practice fields: Research Environment, Research Procedures, Safeguards, Collaborative Working, Publication and Dissemination, Data Management, Mentoring, and Reviewing, Evaluating and Editing. The P2I questionnaire assesses training from secondary school students to early career researchers. According to the European Framework for Research Careers, such training and their assessments exclude to “[d]emonstrate[s] critical analysis, evaluation and synthesis of new and complex ideas” (European Commission, 2011, 8) and to “[c]o-author[s] papers at workshop and conferences” (European Commission, 2011, 8). Therefore, the P2I questionnaire includes all the fields from the ECoC (2017) mentioned above, except the field Reviewing, Evaluating and Editing.

As shown in Table 2, the P2I developers included the fields Research Environment, Research Procedures, Safeguards, Collaborative Working, Publication and Dissemination, Data Management, as well as Mentoring and designed items. The developers created two versions of the questionnaire. The S and the M / Y versions adapt, transform, and reduce the above-described content to two target groups. The two versions consider that secondary school students and bachelor students have no 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

### Table 2. 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         |           |               |
and collegiate learning circumstances. The testees see the items and each response possibility in random order to ensure no bias within the items. This way, items are not easier to answer because their pre-items refer to them.

Step two: Identifying scientific and non-scientific patterns
The P2I questionnaire is a multi-layered tier test, assessing the testees’ patterns, as Tsui & Treagust (2010), Chandrasegaran, Treagust, and Mocerino (2007), or Treagust (1988) outlined. Treagust (1988) designed the original structure of a so-called two-tier test. He based his original design on Tamir’s (1971) approach that distractors are derived from testees’ existing misconceptions. The two-tier test is an extension of classical multiple-choice tests to determine what students know and additionally gain insights about their conceptions on their knowledge. Treagust’s original two-tier test design starts with a multiple-choice question that testees respond to in the first tier; this extends to a second tier where testees reason their response from the first tier designed as a multiple-choice item.

This structure provides quantitative data from both tiers and provides insights into the first tier’s quality (Yang & Lin, 2015). To use such a design allows statements about the testees’ knowledge of appropriate practices in RI/RCR on the one hand and justification patterns for these practices on the other hand. Furthermore, it identifies frequently occurring and stable misconceptions. As Hestenes & Halloun (1995) or Yan & Subramaniam (2018) show, the two-tier test is able to determine in which way testees decide to act in a situation and, at the same time, what concepts they hold about it.

To design multiple-choice responses and distractors which are attractive because they are based on the realities of the target groups, the P2I developers co-created various courses of action for the first tier with students and researchers. They conducted six group discussions with students majoring in education as well as with an interdisciplinary research group at Kiel University in Germany to identify those courses of action for the first tier, which correspond to the target groups’ academic everyday practice. The persons in the group discussions were not randomised; participation was voluntary. Since this procedure involved a considerable investment of time for the participants and no compensation was offered for their participation, those people were approached personally. To make the procedure as efficient as possible students with high intrinsic motivation in research, and without a particular connection to RI or RE were asked to help improve the questionnaire. The developers presented first case scenarios concerning research integrity and asked the focus groups what scientific practice should follow to the situation described in the case scenario. The developers sought open and free-response questions. With this procedure, the developers collected attractive non-scientific practices for two different target groups, a) secondary school students and undergraduates and b) master students and (early career) researchers.

The following displays some original justification examples that lead to justification patterns:

- 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.

The developers clustered these single justifications into patterns. As shown in Table 3, the answers from the participants led to nine different patterns that contain different justifications as well as one justification pattern according to the ECoC.

The developers selected always the justification pattern according to the ECoC and three justification distractors for each scenario according to the content fit. They reformulated the distractors deviating as minor as necessary from the typical answer while at the same time ensuring connectivity to the scenario of the item. In the following the ECoC justification pattern is called scientific justification pattern taking into account that other justifications patterns can also be scientifically.

1 The examples are translated from German to English by DeepL.
Step three: Development of the structure and items

Questionnaire structures have an immense impact on what they evaluate. Whereas simple multiple-choice questions and answers can measure knowledge, priority rankings can evaluate norms and values. Tsui & Treagust (2010, 1074) state that a test’s multiple tier structures can better understand students’ patterns than a simple multiple-choice test.

First test and revision

For the first pilot run, 11 students from the educational field (six bachelor students and five master students) from different courses of Kiel University in Germany conducted the P2I questionnaire in 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, only those courses were selected for which it could be ensured that the participants had not already completed the test. This left two courses where the questionnaire was distributed. They 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 4 show that at least six out of 11 participants answered according the scientific pattern in both tiers in all six items. Item cd9 received the same number of answers according the scientific pattern in the first tier and the second tier. All other five items show that even if the participants answered in the first tier according to scientific practice, they did not always choose the justification tier’s scientific pattern.

Because only in two items that 72.73% of the participants choose the scientific pattern in the second tier, while in the other four items only 54.55% of the participants choose the scientific pattern in the second tier, the developers readjusted the item scenario and decision tier.

This led to a second pilot run, after revision, where four master students conducted the P2I questionnaire (P2I Questionnaire S, Vers. 1.0) (Zollitsch et al., 2020b) in a pre-post-test design. They all came from educational courses of Coburg University in Germany, and the age structure 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 showed 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, items cp5 and cc5 have more scientific patterns in the post-test. Items cc1 and cd2 have the same number of scientific patterns in the pre-test and the post-test, while items cp3 and cd9 had lesser scientific patterns in the post-test.

Acknowledging that the test did not assess learning successes as wished, the developers decided to rearrange the structure and rewrite the failing items. The developers modified the original structure. Instead of two answers in the first stage (decision between yes or no), the P2I questionnaire includes in its new version four possible action options. At the same time, the

---

**Table 3. (Non-) scientific justification patterns.**

| 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 practise will support X’s career and get him/her the next Noble 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. |
(non-) scientific patterns that were previously based on whether the described action corresponds to scientific practice 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 non-scientific practice can still select a rationale that fits a scientific justification pattern.

The first tier no longer focuses on whether a described practice is a sound scientific practice, but instead on the fact that the testees must decide which of the four given actions are a scientific practice.

The P2I questionnaire enlarged its multiple-tier approach to evaluate in detail if the testees can determine what a scientific practice is and how to argue for it. The developers enlarged the two-tier structure to a so-called four-tier structure to avoid testees’ guessing their answers and gaining more information. They added two confidence tiers (Caleon & Subramaniam, 2010; Peşman & Eryılmaz, 2010). These 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 non-scientific answers. The information 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 (non-) scientific pattern. (P2I Questionnaire S, Vers. 2.0 (Zollitsch et al., 2020b) and P2I Questionnaire M, Vers. 2.0 (Zollitsch et al., 2020a))

Test for validating and revision
To validate the final four-tier structure of the P2I questionnaire, the developers conducted a content validation test with international research integrity and research ethics experts. In a

| Scientific practice fields | Research procedure | Collaborative working | Publication and dissemination |
|----------------------------|--------------------|-----------------------|------------------------------|
| Item                       | cp3                | cc1                   | cd2                          |
| Tier 1 scientific decision pattern (%) | 8 (72.73) | 7 (63.64) | 8 (72.73) |
| Tier 2 scientific justification pattern (%) | 6 (54.55) | 6 (54.55) | 6 (54.55) | 6 (54.55) | 8 (72.73) |

Figure 2. Comparison of the results of a pre- post-test.
first step, the content validation was checked by feedback from RI/RE experts. In a next step, the validation test followed the hypothesis that international experts respond to the P2I questionnaire with high confidence; and students (without specific training) respond with less confidence than experts.

For the content validation, the developers invited all 17 members of the Horizon2020 project Path2Integrity, and 12 research integrity and research ethics experts who were contact persons of the training centre work package of Path2Integrity. All had no connection to the learning units or the P2I questionnaire and were invited by email. Their position in either the P2I advisory board, the P2I scientific board, or being a contact from the P2I training centre documents their expertise in RI/RE. The validation procedure instructed all experts to complete both versions (the S and the M/Y version), give feedback, to make the questionnaires as content valid as possible, and ask questions if the situation arises. The participation was voluntary, and the group was not randomised.

The experts’ group from above answered the P2I questionnaire 27 times, and the individuals 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. All experts 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 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 developers checked these comments and integrated them into the P2I questionnaire.

The experts’ comments referred more to aspects of design than aspects of content, such as having an open field for the category country, storing no IP addresses, and a more prominent link to the data policies. The developers implemented those recommendations accordingly. One comment stated that items for data management are missing in version S. As already stated in ‘Step one: defining the content’ data management in an interdisciplinary context is a topic for those who graduated. Therefore, the developers did not intend to integrate those items in version S. Furthermore, two comments declared that they are missing the handling of dilemma situations. As already stated in the status quo, the developers do not focus on moral dilemmas but on contradicting practices, and therefore, the test 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 both 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 for more than one response option. The developers declined this comment after discussing and weighing it against the central concept of the P2I questionnaire.

The procedure confirmed the 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 validity of the questionnaire, we considered the number of participants to be saturated and refrained from consulting further experts.

In the next step, the developers contacted six lecturers from Kiel University in Germany to acquire students from non-RI/RE related courses. Two lecturers came up with 15 students that had not already completed the test. The P2I test was contributed by the lecturers to the students personally by their lecturers. 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 consists 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 for a small sample of items from the S version of the P2I questionnaire as an example. Beyond these examples described, the developers validated all items (and both version) of the P2I questionnaire. Figure 3 shows the students’ confidence 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 the whole research integrity context.

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 still not as familiar with research integrity as experts are.

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

The confidence tiers are essential for determining if the testees are confident about both responses in tier one and tier three. Ideally, they justify their decision with a scientifically substantiated argument while confirming that they are confident with this response.

The P2I questionnaire contains different independent variables to evaluate RI/RCR training successes next to the tiers’ responses. The P2I questionnaire considers relevant differences that might occur.

Discussion
As already stated earlier, one of RI/RCR training’s core themes is scientific practices and their justifications. As shown in the three different steps above, the assessment of such scientific patterns is difficult. The P2I questionnaire is the first step to assess such (non-) scientific patterns.

The multiple-tier structure is a frequently used design for assessing RI/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 RI/RCR training’s successes.

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

Compared to the DIT-2, which originates from ethics education, the P2I questionnaire derives from RI/RCR training. The P2I questionnaire focusses on scientific practices based on rational arguments agreed on by researchers. Both instruments start with a (case) scenario that testees should respond to. While the DIT-2 asks which action is the most important, the P2I questionnaire asks about a justification for the action. The P2I questionnaire, therefore, gives insights how testees argue for their scientific practices.

Although the P2I questionnaire has been developed in the Horizon 2020 project Path2Integrity, it will also be applicable to other RI/RCR training. The P2I questionnaire focusses on scientific practices based on rational arguments agreed on by researchers. Both instruments start with a (case) scenario that testees should respond to. While the DIT-2 asks which action is the most important, the P2I questionnaire asks about a justification for the action. The P2I questionnaire, therefore, gives insights how testees argue for their scientific practices.

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.

This article described the development of the P2I questionnaire following Chandrasegaran, Treagust, and Mocerino’s (2007) approach, designing and 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 an assessment instrument for learning success in RI/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 jeopardize 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 jeopardize the result. All data will be shared at the end of the project, latest 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 jeopardize 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 jeopardize the result. All data will be shared at the end of the project, latest 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 of 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. Published 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. Published 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

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. Publisher Full Text

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, Dienerick 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. 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(9): 502. Publisher Full Text

Integrity in research collaborations: The Montreal Statement. Lancet. 2013; 382(9901): 1310. 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

Löffström E, Pyhältö K: What Are Ethics in Doctoral Supervision, and How Do They Matter? Doctoral Students’ Perspectives. Scandinavian Journal of Educational Research. 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 | Publisher Full Text | Free Full Text

Mayring P: Qualitative Content Analysis [28 paragraphs]. Forum Qualitative
Sophia Jui-An Pan
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.