Elements of proper conclusions

Pols, C.F.J.

DOI
10.1088/1361-6552/ad8f88

Publication date
2025

Document Version
Final published version

Published in
Physics Education

Citation (APA)
Pols, C. F. J. (2025). Elements of proper conclusions. *Physics Education*, *60*(1), Article 015009.
https://doi.org/10.1088/1361-6552/ad8f88

PAPER • **OPEN ACCESS**

Elements of proper conclusions

To cite this article: Freek Pols 2025 *Phys. Educ.* **60** 015009

View the article online for updates and enhancements.

OPEN ACCESS

PAPER

Phys. Educ. **60** (2025) 015009 (9pp)

iopscience.org/ped

Elements of proper conclusions

Freek Pols

Science & Engineering Education, Delft University of Technology, Delft, The Netherlands

E-mail: c.f.j.pols@tudelft.nl

CrossMark

Abstract

Investigating first-year physics students' ability to draw proper conclusions, we analysed 87 conclusions from the same experiment. Through rankings by teaching assistants we identified seven key elements of effective conclusions. These findings reveal a significant gap in students' skills, with about half of the conclusions deemed inadequate. This study underscores the necessity for targeted educational interventions to enhance conclusion-drawing capabilities in physics education. The set of seven elements might provide guidance to improve students' ability to draw proper conclusions.

Keywords: physics education, scientific writing, inquiry

Supplementary material for this article is available online

1. Introduction

In experimental physics, clearly and concisely communicating findings of an experiment, often in the form of a technical or scientific report, is an important skill [1–4]. However, students at all ages have difficulty with writing concise and clear reports [5–8]. Various issues that have been reported in literature, have been observed in our first-year physics lab course (FYPLC) at Delft University of Technology as well. For instance, our students have difficulty in deciding what information to include and how to effectively present that information. What is especially striking in students' first report is their inability to draw proper conclusions. They often merely restate the results, or draw superficial conclusions that are not (optimal) informative. This is in stark contrast with the idea that in science, more (or the most) informative and generalizable conclusions—those valid across multiple contexts—are to be preferred for their broader applicability [9, 10].

First year university physics students' apparent inability to draw proper conclusions may not be surprising as it is known that secondary school students lack this ability as well [8, 11, 12]. In their study Pols *et al* [11] find that grade 10 students draw conclusions that are superficial, unsubstantiated and without specification of limitations to their validity or reliability. The authors attribute this inability to a lack of data-analysis skills and to a lack of relevance for students to produce a useful and trustworthy answer to the research question. In follow-up studies they found that grade 9 students are able to understand that conclusions with more precision and details are more useful and trustworthy and therefore preferable [12]. However, the pupils still had difficulty in producing the *most informative, reliable* and *valid* answer to the research question within the given constraints and limits imposed by the feasibility of obtaining it [13].

Still, we may expect university students to have progressed in the intervening years and to have developed the data analysis skills that enable them to draw informative conclusions. However, our anecdotal evidence suggests that these students may not be demonstrating this ability yet, which provides a compelling rationale to further investigate the matter.

Unsatisfied with the number of students—after four hours of experimenting and many hours on analysing the data and writing a report—that draw conclusions that are superficial and qualitative in nature, we here investigate in detail the quality of their conclusion first. This first part of the research should justify and quantify (or refute) our hunch that the quality of students' conclusions is inadequate. If students indeed are unable to draw proper conclusions, then it seems useful to examine exactly what kind of conclusions we are looking for in order to bridge the educational gap, enhancing students' ability to conclude scientific inquiries meaningfully. Hence, in the second part of the study, we aim to identify key elements that constitute a proper conclusion. Our study is guided by the research questions:

RQ1. What is the quality of first year physics students' conclusions?

RQ2. What elements are considered to constitute a proper conclusion and what elements diminish the quality?

2. Methodology

This explorative, single case, descriptive study [14] consists of two rounds, summarized in figure 1. In the first round the 23 teaching assistants (TAs) of the FYPLC were given a small subset of (anonymized) conclusions which were drawn by students in the previous year (see next section). Each of the 87 conclusions was given to at least two TAs. The TAs were asked to:

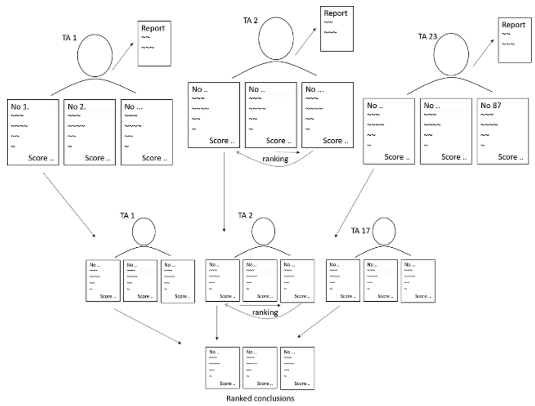

Figure 1. The study consists of two rounds in which teaching assistants are asked to rank a set of conclusions, ultimately leading two a ranked set of top six conclusions. The teaching assistants provide reports on the quality and elements that enhance or diminish the quality of the conclusions.

1. *Rank the conclusions in order of perceived quality (no further instruction given)*
2. *Indicate which conclusions were regarded as insufficient—that is, would probably be graded less than 6 out of 10*
3. *Use the ranked conclusion to indicate what elements constitute a proper conclusion and what elements diminish the quality of the conclusions*
4. *Send the responses to the teacher.*

Ranking the conclusions and marking those which are considered inadequate provides information pertaining to RQ1. Identifying elements that enhance or diminish the conclusion's quality provides information pertaining to RQ2.

The list of identified elements were then used by the teacher to draw his own conclusion.

In the second round, the TAs were given the top five rated conclusions and the teacher's conclusion. The TAs were again asked to rank these. Presumably the teacher's conclusion was ranked best as it constitutes all elements that were identified to enhance the quality.

2.1. Educational setting and participants

The study was carried out at Delft University of Technology's FYPLC, part of the bachelor applied physics program. This recently reformed six ECTS course, focuses on developing inquiry competences. In the initial four weeks, students acquire basic programming skills in Python and learn data visualization and analysis techniques. In the next four weeks the focus is on data collection, processing, and presentation, culminating in their first experimental project. Here, students are tasked with writing the results and conclusion sections to complement provided introduction and methods sections. Their reports are subsequently graded.

The course is facilitated by two instructors and supported by ∼20 TAs, who have completed their first year and received specific training in motivation, scientific writing, and lab work pedagogy. This study involved 23 TAs in the first round and 17 in the second, with most being in their third year or higher, indicating a well-experienced group capable of assessing the quality of scientific reports.

2.2. Experiment

The analysed conclusions stem from an experiment in which students investigate the relation between the force between two magnets as function of their mutual distance, see figure 2. The experimental setup and theoretical background is described in detail in [15].

To increase the relevance for students to do quality research and incite a need to produce a proper conclusion, we make use of a real-life context. Students are asked to validate a theoretical model that describes the force as function of the mutual distance of the magnets ($F = {}^{\alpha}/_{r^4}$) in the context of building a new Magnetic Levitation (MagLev) train. It is said that such a model serves as input for a simulation of the train's driving characteristics. No guidance on the criteria to assess the validity of the model is provided. However, prior to the experiment the students are told that they can be held responsible for any severe consequences that the outcomes of improper research implies. In the given context the practicality of consequences [16] demands high standards of validity and reliability of the evidence. Moreover, when students discuss what criteria to use, they are asked whether they would board the train when they know their suggested criteria are applied.

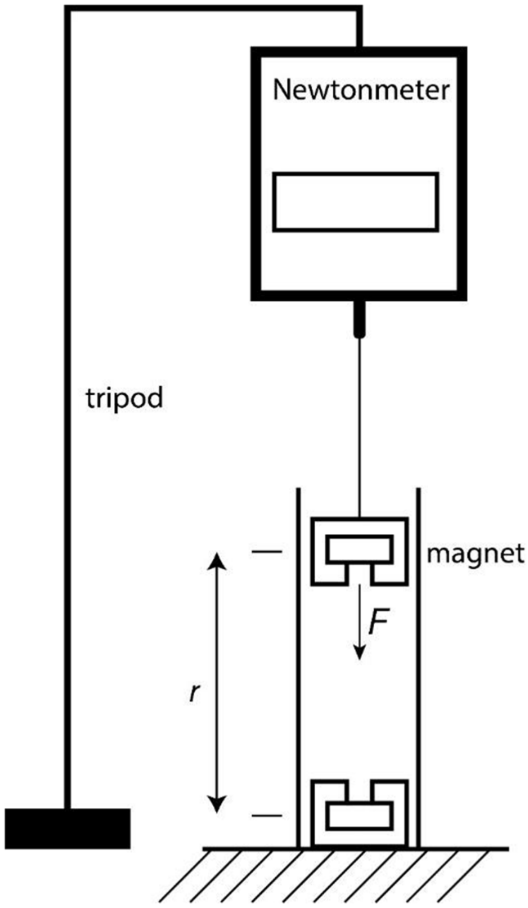

Figure 2. The experimental setup where students investigate the relation between the forces (F) acting between the magnets as function of their mutual distance (r).

During the data analysis in the subsequent lesson, various scaffolding questions are used to help students analyse the data in various ways. They should come to realize that the theoretical model works well only for larger distances as the model assumes zero size magnets. Ultimately students should arrive at a conclusion as the one stated below, written by the teacher:

This research was conducted to experimentally validate a theoretical model—for the force between two magnets as a function of their *mutual distance—in the context of designing a new MagLev train. From the analysis of the results, the expected fourth power relationship with a small systematic error of 0.3 ± 0.1 mm was established. Based on the established relationship, the remanent field was determined to be 1.3 ± 0.2 T. This does not conflict with the values specified by the manufacturer of the magnets (1.29–1.32 T). This validates the theoretical model for heart-to-heart distances larger than 1.3 cm. Because the distance of the magnets of a MagLev train is much smaller (in the order of mm), further research will have to focus on the validity of the model for these smaller distances. As the spatiality of the magnets is more important at smaller distances, this may mean that the magnets can no longer be regarded as a superposition of point dipoles.*

2.3. Data collection and analysis

In the first round, each TA was given a set of eight conclusions along with a description of the task. They were asked to digitally hand in their rankings and their answers to the questions:

1. *Which elements or aspects enhance the quality of the conclusion (and do you thus want to see reflected in other conclusions)?*
2. *Which elements or aspects diminish the quality of the conclusion (and do you thus not want to see in other conclusions)?*

To serve as input for the second round, the rankings were globally evaluated. Using the TAs input (and own experience) the teacher formulated his own conclusion.

In the second round, the teacher's conclusion supplemented with the top five conclusions from the first round, were given again to the TAs. They were asked again to rank the conclusions. To compare the quality of the conclusions, we conducted a quasi-quantitative analysis. Each conclusion was given the value of its ranking position with a score of 1 indicating the best conclusion and a score of 6 indicating the worst. The average score is regarded as a measure of the overall quality of the conclusions. Additionally, the standard deviation of the 17 scores per conclusion is regarded as a measure of the consensus on which conclusion was considered the best. Finally, the best ranked conclusion is analysed to verify whether it indeed contains the elements identified that are regarded to enhance the conclusion's quality.

3. Results

The results are presented in accord with the rounds.

3.1. Round 1

In the first round the TAs were given a set of numbered conclusions and were asked to rank these, indicate which were seen as insufficient and distil elements that constitute a proper conclusion, and elements that diminish the quality of a conclusion. The two conclusions with the lowest score are shown below:

Example 1: From the data of the previous chapter it can be concluded that the theoretical model of a superposition of dipoles is valid to describe the behaviour or two magnets at small distances.

Example 2: A reasonably good agreement was found between the theoretical model and the measurements, so we can say that the relationship between the distance and force between two magnets can be described according to the given model. A match was found between the expected and calculated value, but it probably has a systematic or some other error because it is not close enough to the expected value.

Roughly half of the conclusions (51.7%) was regarded as inadequate by the TAs. Important to note here is that there was not always clear agreement between TAs on which conclusions are inadequate. In three cases (one shown below) a conclusion obtained the highest and lowest score, for example:

Example 3: In this report, a model based on the superposition of magnetic dipoles is tested. The relationship of the force between the magnets and their mutual distance was determined by means of an experiment. This relationship does not conflict with the value from the theoretical model. The particular remanent field is also not inconsistent with the theory, so the theoretical model would *be valid. However, in a follow-up study, the distance should be determined more accurately to avoid a large systematic error. The uncertainty in the measuring equipment is also so great that a follow-up study is recommended to calculate the remanent field again with a reduced uncertainty.*

The TAs who assessed this conclusion were contacted and asked to substantiate their ranking:

TA1: Reading the conclusion again, I would score it lower than I did, though I still regard it an adequate conclusion. The conclusion was easy to read. Its structure suffices as it contains the goal, how this goal is achieved, and it includes results and recommendations for future research.

TA2: I gave it a low score as I think the answer to the research question is inadequate: they talk about agreement with theory but the results upon which this conclusion is based are missing. I found it difficult though to choose what is more important: an incomplete conclusion or a too long conclusion where elements are included which should not be part of a conclusion.

Based on their rankings, each TA provided a list of aspects they thought enhances the quality of the conclusion. Two illustrative responses are given—with between brackets the elements given in the list below:

Response 1: Well-structured story that starts with the research goal/question (2), briefly tells what the results are and draws a logical conclusion (5). The conclusion should provide enough information so that the reader understands how the conclusion is logically deduce from the results (6).

Response 2: Answers research question (5), present results (6) and compare with the theory, repeat the research goal at the start (2), small discussion & recommendations (7).

From these responses, seven elements that constitute a proper conclusion were drawn. According to the TAs, a proper conclusion:

1. Can be read independently from previous text (6×)
2. Restates the research question or goal (16×)
3. Reminds the reader of the study's context (6×)
4. Describes the method briefly (4×)
5. Answers the question (using the given context & reference to literature values if possible) (14×)
6. Provides a substantiation for that conclusion (8×)
7. Provides the main recommendations and suggestions for future research (15×).

Elements 1 and 5–7 largely correspond with the criteria in the rubric used in previous years:

Answers the key question correctly. Conclusions follow naturally from the contents of the previous sections and are concise and accurate. Any recommendations follow on logically from the conclusions. Can be read as a text in its own right. Contains no references.

The responses to the question what elements diminish the quality of the conclusions were much more diverse than the elements that enhance the quality, as can be seen in the following responses:

Response 3: No references (1), useless improvements such as that the heat had a huge influence on the volume and that they should have prepared the test better, spelling mistakes (5), no uncertainties.

Response 4: Many storylines mixed up, so first ask a question and then provide an answer to a different question (2). No substantiation of the conclusions, for instance saying that something is contradictory without showing anything of a result. Rather than making a logical story immediately providing values so that you have no idea what this has to do with the conclusion (4). Too much focus on data analysis and too little on the real conclusion.

A meta-analysis of these responses resulted in the following list of elements that diminish the quality of a conclusions, according to the TAs:

1. Too much extraneous information (7), e.g. including repeating analysis (5×) and references to information in text (13×)
2. Lack of clarity, e.g. vague conclusions (2×), no substantiation of the conclusion (1×), unstructured conclusion (4×), and useless, farfetched recommendations (2×)

Table 1. The average ranking and standard deviation as scored by the 17 TAs for the six conclusions in round 2. A score of 1 would imply that all TAs ranked the conclusion as best.

Conclusion reference number	90	36	22	74	5	67
Average ranking	1,65	2,7	2,9	3,2	5,1	5,4
Standard deviation	0,9	1,2	1,1	1,5	0,9	0,8

3. Missing elements, e.g. no answer to the research question (5×) or missing an introduction (2×)
4. Not readable independently (6×)
5. General spelling and language errors (5×).

As might be expected, these negative results correlate to the elements that enhance the quality of the conclusion when observed from a positive perspective. For example, *not readable independently* (neg. 4) relates to *can be read independently from previous text* (pos. 1).

3.2. Round 2

Table 1 provides the results of the second ranking task in which the TAs were given the five best conclusions and the teacher's conclusion. As might be expected, the teacher's conclusion (no. 90) was ranked best on average. Moreover, the standard deviation indicates there is consensus between TAs on which conclusion is best. Noteworthy is that the standard deviation of the two lowest ranked conclusion are small as well, implying that there seems to be consensus on what conclusions are least adequate. However, the average scores of the middle three conclusions show minimal deviation from each other, and all three exhibit a larger standard deviation compared to the best and worst conclusion. This observation suggests that differentiating the quality among these conclusions is more challenging.

In discussing the difficulty of ranking the conclusions the TAs mentioned that especially weighing is difficult: they have to decide whether a lack of information is more important than providing too much information.

3.3. Analysis of the elements

Below is again the conclusion as written by the teacher. We analyse and present *where* and *how* the seven elements that were identified as enhancing the conclusion's quality are included. By starting the conclusion with restating what was done and why—i.e. its context (2,3), the text becomes independently readable (1). The answer to the research question is given (5,6) (**bold**) followed by a recommendation that fits with the context (7). Moreover, a limitation to the validity of the conclusion is presented in the last sentence (which we here consider part of the answer to the research question).

This research was conducted to experimentally validate a theoretical model—for the force between two magnets as a function of their mutual distance—(2) in the context of designing a new MagLev train (1,3). From the analysis of the results, the expected fourth power relationship with a small systematic error of 0.3 ± 0.1 mm was established. Based on the established relationship, ***the remanent field was determined to be 1.3 ± 0.2 T. This does not conflict with the values specified by the manufacturer of the magnets (1.29–1.32 T). This validates the theoretical model for heart-to-heart distances larger than 1.3 cm (5,6)****. Because the distance of the magnets of a MagLev train is much smaller (in the order of mm), further research will have to focus on the validity of the model for these smaller distances (7). As the spatiality of the magnets is more important at smaller distances, this may mean that the magnets can no longer be regarded as a superposition of point dipoles (5).*

One of the TAs, who did not consider this conclusion the best, pointed out that the description of the method is still inadequate. Again, the issue of assigning weights to various elements differed amongst TAs. Moreover, the analysis above indicates that not all identified elements are independent on each other. For instance, by providing the purpose and context and research goal, the conclusion becomes independently readable.

4. Discussion

At secondary school level, students frequently engage in experimental work and are asked to report their findings. However, merely engaging in practical work and frequently writing a report does not ensure that they are or become able to draw proper conclusions when engaged in experimental work at a later stage: roughly half of the conclusions used in this study were regarded inadequate (albeit assessed in a single case). This indicates that students still lack a solid foundation for drawing proper conclusions. At secondary school level, we may attribute this to their lack of data-analysis skills [11]. However, our university students were trained for three weeks to analyse data and should have attained the required skills to make sense of the data. Therefore, the limited quality of their conclusion is likely attributed to their understanding of the scientific purpose of an experiment (to find and defend the best answer obtainable in the given circumstances [9]) and their understanding of what constitutes a proper conclusion.

Hence, it is essential to enhance students' understanding of the scientific purpose of an experiment *and* enable them to reflect on the quality of conclusions. As there are exist already activities that focus on the first [12], we focus on the latter and identified in this study which elements constitute a proper conclusion.

Reviewing these elements, the first three (*1. can be read independently from previous text; 2. restates the research question or goal; 3. reminds the reader of the study's context*) show great coherence. Given the typical reader's approach—first reading the abstract and, if intrigued, then the conclusions before any other sections—it is crucial that the conclusion be comprehensible independently of the other sections. The conclusion becomes independent readable if the study's context is summarized (*what is the problem?*) and the main research question is repeated, though not necessarily phrased as a question. Moreover, the other four elements (4. describes the method briefly; 5. answers the question (using the given context & reference to literature values if possible; 6. provides a substantiation for that conclusion; 7. provides the main recommendations and suggestions for future research) show coherence as well and seem to be a further specification what is meant by 'conclusions with more precision and details are more useful and trustworthy' (see introduction). These four elements provide the ultimate, warranted claim. It provides the information that allows the reader to, provisionally, establish whether (s)he considers the claim valid. On a more abstract level, these elements seem to constitute the field-invariant elements of Toulmin's argumentation model [17]. The latter fits with the contemporary idea that scientific inquiry can be regarded as the building of a scientifically cogent argument for a claim [18].

Although the first-year students were not yet able to draw proper conclusions, the high degree of consensus in the scores in the first round amongst TAs demonstrates progress in their ability to recognize the quality of conclusions during their studies. The presence of overlapping elements that either enhance or diminish the quality of conclusions further indicates the students' growing capability to assess the quality of conclusions over time. *Where, when* and *how* students develop these skills should be further investigated. Especially since these students seem to have their capabilities but not had these elements at their disposal.

With these seven elements, we can design educational interventions that support students in developing their understanding of what is regarded as a quality conclusions. We are in the explorative stage of designing such activities and study its contribution to students' ability. This seems especially required since conflicting rankings still existed amongst more senior students. Moreover, we asked our TA to rank the conclusion and not write one themselves. Recognizing quality is an important aspect, it does not guarantee that students are able to independently draw proper conclusions [13].

5. Conclusion

As our first year physics students were seen to have difficulty with drawing proper conclusions, this study was set to identify elements that constitute a proper conclusion of a student's science report with the aim to provide scaffolds in teaching students to draw conclusions from their experiments. Using two ranking tasks, TAs identified the best conclusions that were distilled from a simple experiment conducted in a first year physics lab course. The analyses revealed that half of the conclusions were deemed inadequate, suggesting that students still lack a thorough understanding of the scientific purpose of an experiment and the elements of a proper conclusion. Furthermore, a further analysis of the 'best' conclusions led to the identification of seven elements that enhance the conclusion's quality.

The identification of these elements provide a mere initial step in enhancing students' ability to draw proper conclusions. Based on these elements, we may design activities that help students in gaining a feeling for the quality of conclusions. I believe that there is some merit in using the seven elements as a checklist which can be used by the students to evaluate their own conclusions, but also believe that merely using these elements in a checklist does not necessarily students' ability to write proper conclusions independently. How to further scaffold students in writing proper conclusions using these seven elements, surely needs further exploration.

Data availability statement

All data that support the findings of this study are included within the article.

Ethical statement

This study has been approved by the University's ethical commission. All teaching assistants participated on a voluntarily basis on condition of anonymity. Consent was obtained for the results to be published, and permission for the answers to be used verbatim in this paper.

ORCID iD

Freek Pols https://orcid.org/0000-0002-4690-6460

Received 25 April 2024, in final form 28 October 2024
Accepted for publication 6 November 2024
https://doi.org/10.1088/1361-6552/ad8f88

References

[1] Almeida L M D S, Becker K H and Villanueva I 2021 Engineering communication in industry and cross-generational challenges: an exploratory study *Eur. J. Eng. Educ.* **46** 389–401

[2] Hoehn J R and Lewandowski H 2020 Framework of goals for writing in physics lab classes *Phys. Rev. Phys. Educ. Res.* **16** 010125

[3] Joint task force on undergraduate physics programs 2016 *Phys21: Preparing Physics Students for 21st-Century Careers* ed A.P.S.A.A.P. Teachers

[4] Kozminski J *et al* 2014 (*AAPT recommendations for the undergraduate physics laboratory curriculum*) ed A.A.o.P. Teachers p 29

[5] Alaimo P J, Bean J C, Langenhan J M and Nichols L 2009 Eliminating lab reports: a rhetorical approach for teaching the scientific paper in sophomore organic chemistry *WAC J.* **20** 17–32

[6] Haagen-Schuetzenhoefer C 2012 Improving the quality of lab reports by using them as lab instructions *Phys. Teach.* **50** 430–3

[7] Lillis T and Turner J 2001 Student writing in higher education: contemporary confusion, traditional concerns *Teach. High. Educ.* **6** 57–68

[8] Sampson V, Enderle P, Grooms J and Witte S 2013 Writing to learn by learning to write during the school science laboratory: helping middle and high school students develop argumentative writing skills as they learn core ideas *Sci. Educ.* **97** 643–70

[9] Pols C F J, Dekkers P J J M and de Vries M J 2022 Defining and assessing understandings of evidence with assessment rubric for physics inquiry—towards integration of argumentation and inquiry *Phys. Rev. Phys. Educ. Res.* **18** 010111

[10] Chalmers A F 2013 *What Is This Thing Called Science?* (Hackett Publishing)

[11] Pols C F J, Dekkers P J J M and de Vries M J 2021 What do they know? Investigating students' ability to analyse experimental data in secondary physics education *Int. J. Sci. Educ.* **43** 1–24

[12] Pols C F J, Dekkers P J J M and de Vries M J 2022 "Would you dare to jump?" Fostering a scientific approach to research in secondary physics education *Int. J. Sci. Educ.* **44** 1481–505

[13] Pols C F J, Dekkers P J J M and de Vries M J 2023 Integrating argumentation in physics inquiry: a design and evaluation study *Phys. Rev. Phys. Educ. Res.* **19** 020170

[14] Cohen L, Manion L and Morrison K 2013 *Research Methods in Education* (routledge)

[15] Forringer E R 2022 Measuring and modeling the force between permanent magnets *Phys. Teach.* **60** 546–8

[16] Gott R *et al* 2003 Research into understanding scientific evidence (available at: www.dur.ac.uk/rosalyn.roberts/Evidence/cofev.htm)

[17] Toulmin S E 2003 *The Uses of Argument* (Cambridge University Press)

[18] Pols C F J 2023 Development of a teaching-learning sequence for scientific inquiry through argumentation in secondary physics education *Science Education & Communication* (Delft University of Technology Delft)