



## Critical thinking in electricity and magnetism: assessing and stimulating secondary school students

Jan Sermeus, M. De Cock & J. Elen

To cite this article: Jan Sermeus, M. De Cock & J. Elen (2021): Critical thinking in electricity and magnetism: assessing and stimulating secondary school students, International Journal of Science Education, DOI: [10.1080/09500693.2021.1979682](https://doi.org/10.1080/09500693.2021.1979682)

To link to this article: <https://doi.org/10.1080/09500693.2021.1979682>



Published online: 28 Sep 2021.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)



# Critical thinking in electricity and magnetism: assessing and stimulating secondary school students

Jan Sermeus <sup>a,b,c</sup>, M. De Cock <sup>b</sup> and J. Elen <sup>a</sup>

<sup>a</sup>Faculty of Psychology and Educational Sciences, Centre for Instructional Psychology and Technology, KU Leuven, Leuven, Belgium; <sup>b</sup>Department of Physics and Astronomy, KU Leuven, Leuven, Belgium;

<sup>c</sup>Planetarium, Royal Observatory of Belgium, Brussels, Belgium

## ABSTRACT

Critical thinking is one of the most desirable outcomes of education, yet it is often not well defined in curricula. Additionally, there are open questions concerning the domain specificity of critical thinking. In this work, we present two studies aimed at secondary education. Starting from Halpern's conceptualisation of critical thinking we developed a test for assessing critical thinking within the domain of physics (more specifically electricity and magnetism). In the second study, we conducted an intervention study in a quasi-experimental design. Together with experienced teachers, we designed lessons that elicit critical thinking based on the First Principles of Instruction of Merrill. Compared with a control group, the experimental group obtained a significantly higher score on the domain specific critical thinking (measured using the test of the first study), but there was no difference between the groups with regard to domain general critical thinking.

## ARTICLE HISTORY

Received 23 March 2021

Accepted 8 September 2021

## KEYWORDS

Critical thinking; assessment; physics education; quasi-experimental research

## Introduction

Critical thinking (CT) is one of the most desirable outcomes of education. CT involves, among others, taking different perspectives into consideration, recognising assumptions and logical thinking (Bailin, 2002). CT is associated with complex problem solving, decision making in ill-defined situations and citizenship (Butler et al., 2017; Paul & Binker, 1990). Hence, it is mentioned in most curricula as an educational goal.

However, despite its communality, CT is often not well, or not at all, defined in the curriculum (Pithers & Soden, 2000; Thompson, 2011). Additionally, even if it is well defined, it remains difficult to assess. Because of these reasons, and the time pressure educators' experience, CT is often not explicitly targeted or assessed. The result is that even at a university/college level students are not adequately prepared for CT (Arum & Roksa, 2011; Halpern, 2014; Pascarella & Terenzini, 2005).

In educational sciences, different definitions of CT have been proposed (Lai, 2011; Rudd, 2007). A non-comprehensive list is presented below.

- Paul and Elder (2001) defined CT as ‘the art of analysing and evaluating thinking with a view to improving it’ (p. 2).
- McPeck (1981) defined CT as ‘the propensity and skill to engage in an activity with reflective skepticism’ (p. 7).
- Ennis (2011) defined CT as ‘reasonable, reflective thinking that is focused on deciding what to believe or do’ (p. 1).
- Facione (1990) defined CT as ‘purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological or contextual considerations upon which that judgment is based’ (p. 2).
- Halpern (2003) defined CT as ‘the use of those cognitive skills or strategies that increase the probability of a desirable outcome. It is used to describe thinking that is purposeful, reasoned, and goal directed – the kind of thinking involved in solving problems, formulating inferences, calculating likelihoods, and making decisions, when the thinker is using skills that are thoughtful and effective for the particular context and type of thinking task.’ (p. 6).

While these definitions are different, they have some common features such as reflection on ones’ thoughts and thought processes (meta-cognitive skills), and they all mention that CT is used to make an informed decision or perform an action with purpose.

These different definitions reflect ample debate and much research on CT. Is CT a skill or a disposition? Is CT domain specific (DS) or domain general (DG)? Are CT skills transferable between domains? ... All these open questions result in a diversity of interventions aimed at improving CT and in a diversity of assessments aimed at measuring CT.

One can, for example, consider CT as DG and hence teach it in a separate course without a specific focus on any domain. On the other hand, one might consider CT as DS, where both so-called immersion and infusion approaches have been proposed (Ennis, 1989). In immersion, it is assumed that students will learn the DSCT skills by learning the content and skills of a domain, all while the employed CT skills are not made explicit. In the infusion approach to teaching CT these skills are assumed to be DS but still need to be made explicit during teaching. For higher education, the work by Tiruneh et al. (2018) suggests infusion of CT skills in lessons that are designed based on the first principles of instruction (Merrill, 2002) results in higher gains of DSCT skills than immersion. Tiruneh et al. (2018) also found that while the DSCT skills improved, the domain general critical thinking (DGCT) skills did not.

We, and others (Davies, 2013; Robinson, 2011), adopt a synthesis, i.e. we consider CT as a combination of both DGCT skills and DSCT skills. In depth DS knowledge is needed to be able to apply these DSCT skills in a meaningful way. Additionally, we assume that CT requires both skills and the disposition to think critically. In taking this view, the instruction of CT should be infused within the teaching of content, and CT should be assessed using both DS and DG measures.

Assessing CT is difficult (Ennis, 1993), yet several tests measuring CT exist. These differ in format, psychometric characteristics and scope. Some require participants to write essays, others are predominantly multiple-choice. Most of them can be considered

DG tests (Ennis, 2009; Hatcher, 2011; Ku, 2009), i.e. the questions are related to everyday life or cover several subjects. There are less DS tests, and even fewer (see e.g. Walsh et al., 2019 or Sugiarti et al., 2017) that focus on physics. The work by Yanti, Suana, Maharta, Herlina and Distrik (2019) is the only one we found with a suggestion for a DSCT test aimed at secondary education in the context of electricity (i.e. they do not include magnetism).

### ***Aims and objectives***

Critical thinking is also considered an important outcome of secondary education in Flanders. In the curriculum goals (Vlaamse Overheid, 2010) CT is, in line with literature, outlined as follows:

‘The students

- can discuss data, practices, and arguments based on relevant criteria;
- are able to weigh alternatives and are able to make an informed choice;
- can approach issues from different perspectives.’ (p. 15)

As such CT is considered a DG skill. This is similar to curricula in other countries, see e.g. Changwong et al. (2018) who discuss CT in Thailand, van der Zanden et al. (2020) who discuss CT in the Netherlands or Dunn (2015) who discusses CT in Japan. There, however, is an increasing call for embedding CT into curricula as both a skill and disposition (Elen et al., 2019).

Given the importance of CT on the one hand, and the relatively low number of DSCT interventions (Abrami et al., 2015) and tests on the other, the objective of the current study is twofold. The first goal, see study 1, is to construct and validate a DSCT test for physics, more specifically electricity and magnetism (E&M). Such a test should allow to investigate the domain specificity of CT, as well as to assess students’ CT skills. The second goal, see study 2, is to design and assess an intervention that stimulates the DSCT skills of students in higher secondary education (ages 16–17).

Both the test and the intervention were limited to E&M. There are several reasons to limit the scope of this work. The foremost reason is that E&M is a high school subject for all students who choose a study track with a large science component. The second reason is that E&M is an intrinsically difficult subject. Students are required to think both at a microscopic and at a macroscopic level, the areas of electricity and magnetism are naturally interwoven, and the interactions are truly three dimensional (rather than two or one as is often the case in kinematics and dynamics). Finally, and not surprisingly, teachers often experience E&M as the most challenging part of the physics curriculum. This means that teachers have to spend more conscious effort on teaching the content, and hence spend less time on honing students’ cross curricular skills. These aspects make E&M a particularly interesting domain to focus on from the perspective of CT.

The paper is organised as follows. We start by presenting the theoretical framework which serves as the basis for both studies. We then report on the design and validation of the DSCT test on E&M. The next section describes the intervention study (study 2) we carried out to stimulate CT in physics in secondary education. After both studies, we discuss aspects specific to the studies. In the last part, we formulate an overall conclusion.

## Theoretical framework

The theoretical framework for this work is based on the framework of the Halpern Critical Thinking Assessment (Halpern, 2010), abbreviated to HCTA, and is developed in line with the theoretical framework of Tiruneh et al. (2017). The HCTA is a validated DGCT test that measures CT skills in everyday situations. The choice for the theoretical framework of Halpern over that of others (e.g. Ennis, Facione, ...) was made because 'the HCTA is based on CT skills that are commonly mentioned in various definitions of CT, and it includes adequate and well-structured items that appear to measure each of the identified CT skills' (Tiruneh et al., 2017, p. 669).

For the purposes of this paper, it is relevant to highlight the five core CT-competencies that Halpern distinguishes (in *italic*):

- *Hypothesis testing*, which can be related to scientific thinking.
- *Verbal reasoning*, which can be interpreted as being able to distinguish between formal and everyday language, as well as the construction of own opinions and arguments.
- *Argument analysis*, which is different from verbal reasoning as here the goal is to assess arguments made by others. It can be related to logical thinking.
- *Likelihood and uncertainty analysis*, which can be related to statistical thinking.
- *Problem solving and decision making*.

We use the same (five) subcategories for the development of our test. For every category a set of objectives/outcomes can be listed. However, because the HCTA is a DG test, we revised the objectives/outcomes to account for the CT objectives specific to E&M. Some of the reinterpreted objectives are closely related to the components that Halpern distinguished, others require a more subtle interpretation. This translation from Halpern's objectives to the objectives in E&M is presented in Table 1.

## Study 1: development and validation of the test

To measure CT in physics in secondary school students, we designed and validated a test following the design principles set out by Adams and Wieman (2011). The test is meant for grade 11 students that have received instruction on E&M. It is, for practical constraints, administered using paper and pencil and should take no more than 45 minutes. As in the work of Tiruneh et al. (2017), we aimed to mimic the structure and format of the HCTA as much as possible.

We based our test on the CTEM (Tiruneh et al., 2017), which was aimed at first year tertiary education students. To adjust the test for secondary school students the questions were examined by both authors with a physics background. Two criteria determined whether a question was kept for our test. (1) The content of the question must be in the curriculum of the secondary school students, (2) the level of the question must be achievable by secondary school students. From the original 20 questions of the CTEM 12 were adopted, either with minor or major adjustments, 8 questions were rejected. Three new questions were designed, making a total of 15 questions.

**Table 1** . The DG objectives of the HCTA are linked to our interpretation of CT skills required for a critical physicist.

	HCTA		Interpretation towards physics
Argument Analysis	identify key parts of an argument: conclusion, reason, counterargument	AA1	identify key parts of a physical argument
	Identify the lack of information or key parts of an argument	AA2	identify the lack of information or key parts of a <i>physical</i> argument
	Questioning generalisations	AA3	questioning predictions/extrapolations
	Create an argument	AA4	create an argument/derivation
	Assess the value of a source	AA5	assess the value of a source
	provide an opinion	AA6	provide a hypothesis
Hypothesis Testing	correlation vs. cause and effect	HT1	identify cause and effect
	recognise the need for more factual information in order to make valid conclusions	HT2	recognise the need for more factual information or data in order to make valid conclusions
	recognise need for good experimental conditions: control group, unbiased sample selection, ...	HT3	recognise need for good experimental conditions
	Evaluate the value and correctness of an explanation	HT4	evaluate the value and correctness of a physical explanation
		HT5	evaluate the value and correctness of experimental data
Verbal Reasoning	Recognise ambiguity of terms	VR1	recognise ambiguity of terms
		VR2	recognise ambiguity in <i>data</i>
	identify vague ideas/terms	VR3	identify vague ideas
	Recognise invalid reasoning	VR4	recognise scientifically invalid reasoning
		VR5	recognise logically invalid reasoning
	recognise that personal opinion does not constitute an argument	VR6	recognise that personal opinion does not constitute an argument
	<i>*from the end goals in [regional] education*</i>	VR7	evaluate ideas from a different perspective
Likelihood and Uncertainty		LU1	understand the probability and likelihood of an event occurring
	understand the probability and likelihood of an event occurring	LU2	being able to understand the origin and limitation of noise in data and experimental error
		LU3	understand the limits of extrapolation
		LU4	make valid predictions
	Recognise assumptions	LU5	recognise assumptions
Problem Solving and Decision Making	recognise partial problems	PSDM1	recognise partial problems (as part of generic solving strategies)
	being aware of solving strategies	PSDM2	being aware of solving strategies in physics
		PSDM3	being aware of solving strategies that are not specific to physics
	generate reasonable, creative solutions to an everyday problem	PSDM4	generate reasonable, creative solutions to a physics problem

Some of the objectives can be adapted without altering, others require more attention. The objective 'evaluate ideas from a different perspective' was added to the objectives of Halpern in order to include the objectives set by the [regional] government.

Two of the three new questions were based on questions found in textbooks (Hiegelke et al., 2006; McDermott & Shaffer, 2002), and one question was constructed by the authors. These three questions, while entirely new, replaced three questions of the CTEM and probed the same CT competences.

The adapted and newly developed test items were then presented to experts and experienced teachers and adopted according to their feedback. In the next stage, the items were tested in think-aloud interviews with students and finally, the test was administered to a large group which allows statistical analysis.

### ***Expert judgement: round table discussion with physics teachers***

The test was presented to a group of five experienced physics teachers. During a round table discussion, the teachers were asked to first solve the entire test, and subsequently evaluate the test-items based on the following criteria:

- Is the language used in the item appropriate for students?
- Given the curriculum goals, should secondary school students be able to answer this item?

The goal of these questions was to find out whether the correct notation and terminology were used, and to assess the difficulty of the test item (given the curriculum goals). Based on the teachers' comments the questions were reformulated, two questions needed major adjustments, the other questions needed minor or no adjustments. One question was omitted as it closely resembled another question. The teachers indicated that they were confident that the adjustments would be sufficient to make the questions suitable and attainable for their students, and no further discussion was needed.

### ***Think aloud interviews with students***

The revised version of the test, that consists of 14 items, was administered in think aloud interviews with four students. During these interviews, we asked the students to complete the test while thinking out loud. The goal of these cognitive interviews was to figure out whether the students understood the questions as intended, whether the responses of the students were in line with the intentions of the authors, and to obtain an indication of the time needed.

One student was able to correctly answer almost all questions and all questions were answered correctly by at least one of the students. Based on these interviews, we made minor changes to some questions by slightly adjusting the phrasing. There was no time limit, allowing the students to take all the time they needed to think of and formulate an answer. The students spent between 1 h 24 min and 1 h 55 min on the test. To reduce the time needed, some items were shortened by removing repetitive elements from questions.

### ***Large group administration***

We administered the final version of the test to a convenience sample of 162 students in 9 classes in 6 different secondary schools. 88 students were male (54.3%), 72 were female (44.4%) and 2 students (1.2%) did not indicate their sex. The students were on average 17 years old (with  $\sigma=0.42$ ). They had on average 6.7 lessons of mathematics ( $\sigma=1.0$ ) and 2.6 lessons of physics ( $\sigma=0.58$ ) per week. A normal school week in Belgium comprises 32 lessons of 50 minutes.

We administered the test at the beginning of the school year to students whom all received E&M instruction in the previous school year. Students had to complete the test in one lesson block of 50 minutes. We asked the students to stay in their seat and work quietly after they finished the test. The students were seated in a larger than

normal classroom, allowing them to sit alone at a table, reducing the possibility for cheating. They had nothing on the table except the test bundle, which includes a formulary, and one pen. We gave all students the same basic instructions, including an explicit reference to the formulary. Half of the classes were asked to start the test at the back, working from the last question towards the first. A researcher was present during test administration. The data collection went without any noticeable hiccups, hence all data collected was treated as valid.

### Example question and scoring

The majority of questions are a combination of forced choice and open format (similar to HCTA). For example, a statement is presented, and we ask the student to indicate whether the statement is correct or wrong and subsequently explain that choice.

In [Figure 1](#) one of the questions is presented. It asks whether or not you can generally conclude that (electrical) resistivity increases with increasing temperature given the

**Question 10**  
The temperature dependence of the electrical resistivity of different materials was measured, and is presented in the table below.

Temperature (K)	Resistivity aluminium ( $10^{-8} \Omega\text{m}$ )	Resistivity gold ( $10^{-8} \Omega\text{m}$ )	Resistivity iron ( $10^{-8} \Omega\text{m}$ )	Resistivity copper ( $10^{-8} \Omega\text{m}$ )
1	0,0001	0,0220	0,0225	0,002
10	0,000193	0,0226	0,0238	0,00202
100	0,442	0,65	1,28	0,348
200	1,587	1,462	5,20	1,046
300	2,733	2,271	9,98	1,725
400	3,87	3,107	13,1	2,402
500	4,99	3,97	23,7	3,09
600	6,13	4,87	32,9	3,792
700	7,35	5,82	44,0	4,514
800	8,7	6,81	57,1	5,262

Can you, based on these measurements, conclude that in general 'the resistivity increases with increasing temperature'?

Yes       No

Why or why not? Explain your answer.

**Figure 1.** Example of a question. The text is translated from Dutch. This question probes elements of hypothesis testing (HT2) and of likelihood and uncertainty (LU3), see [Table 1](#).



**Table 2.** An overview of which question probes which CT-skill. The codes (e.g. AA1, HT4, ...) refer to the codes from Table 1.

Question	AA	HT	VR	LU	PSDM
Q1	AA1				
Q2	AA2	HT4			
Q3	AA2	HT4	VR4		
Q4					PSDM3, PSDM4
Q5	AA4	HT1	VR1, VR5		
Q6	AA2	HT2		LU5	
Q7	AA3	HT4		LU5	
Q8			VR1		
Q9	AA5	HT4	VR7		
Q10		HT2		LU3	
Q11	AA2	HT2			
Q12		HT2	VR2	LU5	
Q13					PSDM2, PSDM3
Q14		HT2	VR6, VR7		PSDM4

Notes: AA stands for argument analysis, HT, hypothesis testing; VR, verbal reasoning; LU, likelihood and uncertainty; PSDM, problem solving and decision making.

presented data. It is categorised as needing ‘understanding the limits of extrapolation’ (LU3) and ‘recognising the need for more factual information or data in order to make valid conclusions’ (HT2). It is clear from this example that a question might probe more than one CT-skill. This was true for 12 out of 14 questions. An overview of which question probes which CT-skill is presented in Table 2.

To this question a student might for example answer that ‘no, you cannot draw this conclusion because in semi-conductors the (electrical) resistivity decreases’. While this is physically correct it does not answer the question that was posed, which asks whether you could draw the conclusion *based on the given data*. Hence a student would not receive full credit for this answer. The scoring therefore requires a detailed scoring scheme. The layout of the scoring scheme follows that of the Dutch version of the HCTA test (Evens et al., 2014). The scheme specifies the prompts that should appear in the response of a student. These can be indicated by the student either clearly, poorly or not at all. The student then receives respectively 2, 1 or 0 points for that prompt. The scoring for this example question is presented in Figure 2.

The scoring scheme was developed together with the development of the test-items and was refined after the large group administration. In a first round of scoring the large dataset, two researchers scored all answers of one class of 20 students while they adhered to the original scoring scheme. After a comparison of their scores, the differences were discussed and the scoring scheme was adjusted and refined where necessary. Finally one researcher scored all answers of all students with regard to half of the questions, the other researcher scored all other answers. Throughout the scoring the researchers refrained from adding personal interpretation to the answers of the students.

## Results

The internal reliability of the test is typically expressed by Cronbach’s alpha,  $\alpha_C$ . However, it is argued that this is a poor estimate of the reliability because the conditions necessary for the correct calculation of  $\alpha_C$  are almost never met and that other estimates should be reported (Revelle & Zinbarg, 2009; Sijtsma, 2009; Widhiarso & Ravand, 2014).

Question 10 /2		
No points are awarded for “yes” and any subsequent explanation. Up to 2 points can be awarded for an explanation following an indicated “no”. The supporting explanation should mention either:		
You can only make a statement regarding these materials.		
Points awarded	How clearly is this prompt present	example
2	Clearly marked	Based on these measurements we can only say this is true for aluminium, gold, iron and copper.
1	In a limited way	-These are all metals, hence this is only true for metals. -This is not true for semi-conductors
0	Not marked	
OR		
You can only make a statement regarding this temperature range.		
Points awarded	How clearly is this prompt present	example
2	Clearly marked	We can't say anything for temperatures higher than 800 K.
1	In a limited way	This is only true for these temperatures. <i>(note: this implies it is only for the given temperatures, not even the entire temperature range. )</i>
0	Not marked	

**Figure 2.** Example of the scoring scheme used to score the question of Figure 1. The text is translated from Dutch.

We estimate reliability by the greatest lower bound (glb) using the Psych package (Revelle, 2019) in R (R core team, 2017). It can be interpreted in the same way as  $\alpha_C$ . We found a reliability of  $glb = 0.72$ , which is considered acceptable.

In view of analysing the structure of the test, an exploratory factor analysis was conducted. The test was designed to account for the five dimensions of CT that Halpern distinguishes (see Table 2), however no clear common factors could be found to cluster the data. This might be because we do not have enough data; it might also be because the dimensions as set out by Halpern are not strictly separated. One can imagine that someone who is able to analyse arguments can also create their own arguments, which is categorised under ‘verbal reasoning’.

Despite the adjustments in the early development stages the provided time to complete the test was not enough for a large fraction of students. 29 out of 162 did not complete the entire test. To account for the missing data, the analysis was compared between the reduced data obtained from listwise deletion and the data employing regression imputation (Kang, 2013), i.e. filling in the missing data by using projected estimates. The results of this analysis are comparable ( $glb=0.74$ ). Hence here only the analysis and the results of the reduced data are reported. The data and analysis that support the findings are available from the corresponding author upon reasonable request.

In Table 3 various metrics for the items are presented. For almost all questions some students obtained a perfect score and some students scored zero. The ratio of the average score to the maximum available score is a measure of the item difficulty. These values are low for all questions (from 9% to 43%) indicating that the test was difficult for the

**Table 3.** An overview of psychometrics for every test item.

Question	N	Max available	Mean	SD	min obtained	max obtained	Item Difficulty	Item Discrimination	% partial credit
Q1	162	4	0.7	1.0	0	4	0.17	0.10	36
Q2	162	4	0.3	0.64	0	3	0.10	0.30	21
Q3	162	4	0.87	0.92	0	4	0.22	0.13	57
Q4	162	6	2.57	1.6	0	6	0.43	0.19	94
Q5	162	2	0.22	0.53	0	2	0.11	0.09	16
Q6	162	2	0.28	0.63	0	2	0.14	0.03	19
Q7	162	2	0.7	0.89	0	2	0.35	0.21	41
Q8	162	1	0.37	0.48	0	1	0.37	0.36	37
Q9	160	5	0.67	1.1	0	4	0.17	0.34	31
Q10	162	2	0.62	0.70	0	2	0.31	0.13	49
Q11	156	2	0.29	0.55	0	2	0.15	0.24	25
Q12	147	2	0.18	0.51	0	2	0.09	0.06	13
Q13	136	4	0.8	0.99	0	4	0.20	0.27	48
Q14	134	4	0.6	0.85	0	3	0.20	0.10	38

**Table 4.** Resulting descriptive statistics given for every subdomain Halpern distinguishes.

	N	Max	$\mu$	$\sigma$	%
AA	155	25	4.01	2.77	16 ± 11
HT	133	29	4.58	2.87	16 ± 10
VR	146	18	2.34	1.87	13 ± 10
LU	147	8	1.78	1.40	22 ± 17
PSDM	134	14	3.80	2.17	27 ± 15
Total	133	44	8.88	4.32	20 ± 10

N, number of participants;  $\mu$ , average number of points;  $\sigma$ , standard deviation from the mean; Max, maximum amount of points that can be awarded for the indicated subdomain and %,  $(\mu \pm \sigma) / \text{Max}$ . AA stands for argument analysis; HT, hypothesis testing; VR, verbal reasoning; LU, likelihood and uncertainty; PSDM, problem solving and decision making. Some questions probe several elements of CT, see Table 2, hence the total is smaller than the sum of the 5 elements of CT.

students. This is confirmed in the low overall scores ( $20 \pm 10\%$ ) and low scores for all five subdomains of CT, see Table 4. The item discrimination values, determined using the `tab_itemscale` function of the `sjPlot` package, are also low (often lower than 30%) though never negative. Finally, it was noted that many students gave partial answers.

### Discussion and conclusion of study 1

We designed and validated a test to assess secondary school students' CT skills in E&M, by mimicking the structure and format of the HCTA and the CTEM.

Validation in a large group showed that the internal consistency of the test is acceptable, certainly given that CT is a complex cognitive competence that is comprised of several interlinked subskills and that this is exploratory research (Nunnally, 1978). However, the test showed to be too long to be complete in 45 minutes for a fraction of students. In future studies more time is to be allotted (at least one hour) for the administration of the test. Item difficulty and discrimination scores indicate that the test should be interpreted as a first iteration, but that additional development is still needed.

Test administration was done in a convenience sample and as such the test results should be interpreted with care and no statement can be generalised to all Flemish students. Still, it is remarkable that the results are very poor as only a few students scored more than 50% and on average the students scored  $20 \pm 10\%$ .

Several questions arise, answers to which might explain this poor result:

- (1) Is the test too difficult for students? We think this is unlikely as in the round table discussion with experienced teachers they indicated that the questions in the test should be attainable for the students, see the section on the round table discussion, and all questions were answered by at least one student.

In future research this explanation could be checked by administering this test in conjunction with a non-CT test of the same E&M content.

- (2) Was the test taken at a poorly chosen time? The students might not be able to remember the subject matter they were tested on given that the content was taught 4–12 months earlier. To reduce this factor the students were given a formulary with all necessary equations, yet this might not have been enough. We cannot rule out this option since no analogue test was administered which only tested understanding of the E&M subject matter.

However, students in study 2, see further, obtained similar results despite having the test administered at the end of a year of instruction (without a 4 month delay). This suggests that this explanation is also unlikely.

- (3) Are the students not adequately prepared to think critically? This might express itself in two ways. Students might have learned to think critically but may have, over the summer holidays, lost this skill. Or, alternatively, students might not have learned how to think critically in E&M at all. Either way this explanation is worrisome for obvious reasons. It means that one of the major educational goals is not reached.

In conclusion of this first study, a DSCT assessment was developed for secondary education on the topic of electricity and magnetism. The test is added in attachment for future use, where the authors see the following possibilities:

- As a start for future development of DSCT tests, be it in physics or as inspiration for other fields.
- As a way to have an initial quantitative measure of changes in the effect of instruction on the DSCT skills of the students (see study 2).
- As a way to investigate the transferability of CT skills.
- As a way teachers can evaluate the level of CT in their class and evaluate whether or not more attention to CT is needed.

## Study 2: stimulating CT through an intervention

Given the real possibility of the third explanation in the discussion of study 1, i.e. that students are not adequately prepared to think critically, an intervention study was set up with the goal of improving the DSCT skills of students (in E&M).

We therefore designed, implemented and assessed a learning environment that supports critical thinking within physics lessons, based on the theoretical framework presented at the beginning of this paper. The following research questions were asked in the context of grade 11 (ages 17–18) secondary education physics education:

- RQ1) What is the effect of the designed intervention on DSCT skills of the students? The hypothesis is that the intervention has a positive influence on DSCT skills of students.

RQ2) What is the effect of the designed intervention on DGCT skills of the students? The hypothesis is that as students grow in their DSCT, they are able to transfer some of these skills to domains outside of E&M (or physics), and so that also the DGCT skills should improve (as measured by an adaptation of the HCTA).

RQ3) What is the effect of the designed intervention on content knowledge of the students? It is expected that students are, throughout the intervention, triggered to think critically and that therefore they will gain more insight in the subject matter. We expect a better performance on a physics content test.

## ***Design of the study***

### ***Learning community and lesson design***

Five experienced physics teachers participated in a yearlong professional learning community (PLC) where they co-designed a learning environment aimed to support CT in physics. More specifically, they co-designed 6 lesson packages. One lesson package consisted of one to three consecutive lessons covering one specific topic in E&M:

- developing a model for electric charge,
- Coulomb's law,
- defining the electric field and the field lines representation,
- resistance and Ohm's law,
- introduction to dc-circuits,
- the magnetic field surrounding a current carrying wire or inside a coil, and electro-magnetic induction.

All lesson packages were designed in line with the first principles of instruction (Merrill, 2002). That means, they all started from a real world problem and included phases of activation, demonstration, application and integration. Additionally, the lessons were infused with elements that invite students to think critically (Ennis, 1989). This means that the five aspects of CT and their interpretation were explicitly taught to the students. At different occasions during the six lesson packages the applied CT skills were made explicit and discussed with students. More details are given below. This lesson design is in line with Tiruneh et al. (2018), Bensley et al. (2010) and Abrami et al. (2015).

The first lesson package was designed by the authors. The other five lesson packages were co-designed by the participating teachers. For every topic, one teacher took the lead in designing the lessons. This proposal was then reviewed by the other teachers and researchers, and comments and alternatives were formulated. After a group discussion, the lesson package was edited and finalised. Field notes were made by the authors throughout the year during the learning community meetings.

### ***Example of an intervention lesson***

To illustrate how CT was introduced into the lessons an example is presented. The lesson started with two movie clips showing the difference in sound an electric guitar makes when it is, or is not, plugged into an amplifier ( $2 \times 30$  sec). This sets the context of the lesson wherein the goal is to understand how the electric guitar is able

to pick up the vibrations of the strings and turn them into an electric signal that can be amplified.

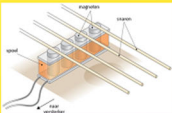
After these clips, and still as a part of the introduction, a quote that the teachers found online is presented to the students, see Figure 3(a). The students were asked to, in groups of two and for five minutes, analyse the quote using the elements of CT, see Figure 3(b). This was followed by a 5 minute whole class discussion.

Then the lesson continued with experiments showing different aspects related to electromagnetic induction with guided inquiry worksheets. These guided experiments lead to Lenz's law. The lesson ended by using Lenz's law to explain how the electric signal is generated in the electric guitar.

### Quasi-experimental design

To study the effectiveness of the designed learning environment, a pretest – post-test quasi-experimental research design was implemented. The experimental and control group consisted of convenience samples. The five teachers of the PLC implemented the designed learning environment in their classes (experimental group). The experimental group consisted of 6 classes in 4 schools for a total of 114 students. Four teachers, who were not part of the PLC, taught their E&M course 'as usual' (control group). The control group consisted of 4 classes in 3 schools for a total of 83 students. All classes were selected from the same school system (general education preparing students for higher education), and from study tracks with a large math and/or science component. Table 5 gives an overview of the characteristics of both groups.

To further check comparability of both groups and learning gains a physics concept test and a DGCT test were administered at the start of the intervention year. The DSCT test for the first study was not administered as a pretest because the students had not yet received the necessary E&M content. Without having acquired that E&M knowledge, participants would have been doomed to fail. We wanted to avoid such a



"from the internet..."

An electromagnetic element consists, roughly, of one or more **magnets**, and a **spool**.  
Your strings are made of a **ferromagnetic** alloy consisting of iron and nickel. The magnet lines like going through ferromagnetic material, because it conducts better than the air around it.  
Now you have to imagine that a string **moves** up and down (in reality the string makes of course much more complicated jumps) and if the string is up, the magnetic fieldlines will be pulled up as if it were. If the string is below, the magnetic field is less stretched.  
So in the frequency wherein the string makes one vibration (up, down), the **magnetic field inside the spool** also went "up and down" as in more magnet lines to less magnet lines.

At secondary school you learned that **if the magnetic field inside a spool changes, a current will run in the spool**. That current has the same frequency as the string has above the element.

**Critical questions:**

- .Vague terms/wording?
- .Faulty reasoning?
- .Extrapolation/generalization?
- .Cause-effect?
- .Essential steps?
- .Need extra information?

(a)
(b)

**Figure 3.** Two slides (a and b) from the supporting powerpoint presentation that teachers used to introduce CT in a lesson on EM induction. A fragment of an online discussion forum was presented to the students. The quote is translated from Dutch to English in an attempt to keep the spirit of the quote the same. Hence any vocabulary, grammatical or physical errors are from the original text. E.g. 'magneetlijnen' is translated as 'magnet lines' rather than translating it to 'magnetic field lines'.

**Table 5.** overview of the different metrics describing the experimental and control group.

	Experimental	Control
N	114	83
Age	15.8 ± 0.5	15.7 ± 0.5
Sex	52 M, 62 F	39 M, 40 F, 4 unknown
#h physics '15-'16	1.4 ± 0.5	1.7 ± 0.5
#h math '15-'16	5.1 ± 0.5	5.0 ± 0.1
#h physics '16-'17	2 ± 0	2.4 ± 0.7
#h math '16-'17	6.1 ± 1.3	6.4 ± 0.8

Note that one 'hour' of physics or math is in reality 50 minutes. The academic years are indicated, showing that students also before the intervention year showed similar values on the different metrics.

negative experience. The physics concept test consisted of three open ended questions adapted from McDermott and Shaffer (2002) covering topics that were treated in the preceding physics course (which was on thermal physics and buoyancy). The score of this test is interpreted as a proxy for the cognitive abilities in physics of the students. The DGCT test is a reduced version of the Dutch translation of the HCTA (Evens et al., 2014).

To study the effect of the intervention several tests were administered to both groups after the intervention:

- to measure their DSCT the test presented in the first study of this paper was administered,
- to measure the effect on DGCT the same DGCT test was administered after the intervention as was administered before,
- to measure the effect on content knowledge, physics questions for the December and June exams were designed and administered.

Table 6 gives an overview of the quasi-experimental research design. No classroom observations were done for either of the groups.

## Results

### Quantitative results

To answer the RQ's, the test data were statistically analysed. The data and analysis that support the findings are available from the corresponding author upon reasonable

**Table 6.** Timeline of the research design. DGCT and DSCT refers to domain general/specific critical thinking.

Timing	Action	Experimental group	Control group
Beginning of the year	Concept test	X	X
	DGCT test	X	X
During the first semester	Regular lessons	N-3	N
	intervention lessons	3	
December-exams	Physics content test	X	X
During the second semester	Regular lessons	N-3	N
	intervention lessons	3	
Before the June-exams	DGCT test	X	X
	DSCT test	X	X
June-exams	Physics content test	X	X

An 'X' indicates that the experimental or control group was administered the test or received lessons packages. 'N-3' indicates that all lessons packages except 3 were regular lessons.

**Table 7.** An overview of the scores (average  $\pm$  standard deviation) of both groups on the different tests that were taken, as well as the  $p$ -value reported from the Shapiro-Wilk test for normality.

Test	Experimental		Control	
	mean $\pm$ sd	$p$ -value Shapiro-Wilk test for normality	mean $\pm$ sd	$p$ -value Shapiro-Wilk test for normality
Concept	3.79 $\pm$ 1.53	<.001	3.93 $\pm$ 1.55	0.0014
DGCT Pre	30.39 $\pm$ 4.97	0.002	30.73 $\pm$ 3.77	0.079
Physics content December	4.95 $\pm$ 1.69	0.091	5.85 $\pm$ 2.30	0.14
Physics content June	8.66 $\pm$ 2.26	<.001	7.75 $\pm$ 2.26	0.017
DGCT Post	33.03 $\pm$ 4.10	0.17	33.27 $\pm$ 4.27	0.058
DSCT	10.69 $\pm$ 4.84	0.018	8.03 $\pm$ 3.57	0.14

request. The results are listed in Table 7. For most tests normality could not be assumed, as is evident from the  $p$ -values of the Shapiro–Wilk test. Comparing the control group and the experimental group with regard to the concept test and the pre-DGCT test through a Wilcoxon test showed no significant differences ( $p = .41$  and  $p = .59$ , respectively). Based on these results there is initially no significant difference between both groups, neither in the physics concept test nor in the DGCT test. Based on these tests and on the metrics reported in Table 5, we therefore assume that the control group and experimental group are initially comparable on all relevant variables.

There was not attrition, however, there are large gaps in the data as unfortunately some of the data was lost by the teachers, see Table 8. The subjects with missing values were removed through pairwise deletion. As data were predominantly missing in groups, i.e. a class group did not complete a test, this is assumed to be missing at random (MAR).

To answer RQ1 we performed a linear mixed effects analysis of the relationship between the score on the DSCT test and the intervention (Bates et al., 2015). The intervention was entered as the fixed effect, the class group of the students was a random effect on the intercept and slope of the linear model.  $P$ -values reported were obtained by likelihood ratio tests of the full model with the effect in question against the model without the effect in question (Winter, 2013). Gender was inspected as a fixed effect but was omitted as it did not have a significant effect on the linear model ( $\chi^2=1.5669$ ,  $p=.2107$ ). The intervention had a significant effect on the DSCT test scores of the students ( $\chi^2 = 4.6134$ ,  $p=.03172$ ). The following linear model described the data  $DSCT = 8 +$

**Table 8.** The dataset is incomplete. This table gives an overview of which data was obtained (X).

Teacher	Concept	DGCT pre	Physics content December	DGCT post	Physics content June	DSCT
Exp 1			X	X	X	X
Exp 2	X	X	X		X	X
Exp 3	X	X	X	X	X	
Exp 4	X	X	X	X	X	X
Exp 5	X	X	X	X	X	X
Contr 1	X	X	X	X		X
Contr 2	X	X	X	X	X	X
Contr 3	X	X		X		X
Contr 4	X	X	X	X	X	

Boxes without an X indicate that the data was not analysed (either the teacher did not administer the test or the data was lost by the teacher).



2.4\*intervention. However, the residuals of this model were not normally distributed. A Shapiro Wilk test for normality resulted in  $W = 0.97679$  with a  $p$ -value = .03218. Hence a subsequent fitting was conducted using the MCMCglmm package (Hadfield, 2010), in R. This Markov Chain Monte Carlo generalised linear mixed model uses the same fixed and random effects. From this the same linear model was obtained:  $DSCT = 8 \pm 1.3 + 2.4 \pm 1.6$ \*intervention. This model, and its error bars, describes 95% of all Markov chain Monte Carlo simulations. The intervention was found significant with  $p < .01$ .

Scores of the experimental group were higher across all items of the DSCT test, with a significantly higher score for questions 1, 10 and 11 ( $p$ -values were respectively .0041, .029 and .034). These items are not part of one CT-construct that can be reliably assessed. This might be related to the non-result of the exploratory factor analysis in Study 1.

To answer RQ2 again general linear mixed models were used. As repeated measures are analysed, time is added as a fixed effect in addition to the intervention. The class group remained the random effect on the intercept and slope. From this analysis the scores of both the control and experimental group were significantly higher at the end of the year ( $p < .001$ ). There was, however, no significant impact of the intervention on the DGCT ( $p = .83$ ).

To answer RQ3 again general linear mixed models were used to model the relationship between the intervention and the content knowledge. For this the score on the conceptual test in September, the physics content test in December and the physics content test in June were used as repeated measures. From this analysis again no significant impact of the intervention was found ( $p = .32$ ).

### **Qualitative results**

While this study does not have an explicit qualitative element, we think it is worthwhile to share the following.

Teachers indicated that making CT skills explicit was difficult. They referred both to pinpointing the exact point where CT skills are required as well as correctly formulating the CT-skill that is being trained. The teachers indicated that they improved throughout the intervention.

The teachers of the learning community also reported that students were more attentive during the CT-lessons. Students also explicitly asked to work using the 'CT-approach'. One teacher quoted one of her students: 'Can we do this as 'Critical Thinking'? That way we'll better understand it.'. However, teachers also mentioned that students did not transfer the obtained CT skills and active thinking attitude to other lessons of their physics course or to other courses.

Participating teachers perceived the learning community as useful, insightful (partially because they were 'forced' to spend time and attention to the lesson-design) and not as a 'waste of time' despite the substantial time investment they made.

### **Conclusion and discussion of study 2**

With regard to RQ1 we conclude that the DSCT skills, in E&M, improved for the students who learned E&M in the designed learning environment. Several aspects of the design might have contributed to create a successful intervention. The intervention

lasted an entire year giving both teachers and students the chance to absorb the CT skills. The lessons were designed based on Merrill's First Principles of Instruction, creating a strong learning environment. And finally, CT was made explicit during the lessons, allowing teachers to demonstrate the CT skills and students to consciously learn and apply the CT skills. This is in line with the suggestions made by Abrami et al. (2015) where he states that instruction that incorporates dialogue, authentic instruction and mentorship may lead to more effective CT learning. With regard to the DGCT skills (RQ2) we found that while for both groups the average score was higher after the intervention, there was no difference between the control and experimental group, i.e. the intervention had no effect on the DGCT skills. The increased score of both groups might be attributed to a learning effect of the test (the test was identical in both pre and post measurement), to the year of secondary education all students received, or to the increased maturity of the students.

Combining the higher score of the experimental group on DSCT and the lack of difference on DGCT suggests that indeed there is a difference between DG and DSCT. It also suggests that improvements in DSCT do not necessarily constitute an improvement in DGCT. This adds to literature where it has been shown that, inversely, improving DGCT also does not necessarily implies an improvement in DSCT (Halpern, 1998). Transfer remains a tantalising issue. It may be expected that additional attention to DSCT in other subjects will improve the CT skills in those domains. The question that remains, and which should be subject of future research, asks whether improving students' DSCT skills in a wide range of scholarly subjects will improve the students' DGCT skills and whether this will lead to thinking critically in everyday life as citizens.

With regard to RQ3, the initial conclusion is that there is no impact of the intervention on the content knowledge of the students. However, it is worth noting that the physics content test scores of the experimental group were, compared to the control group, equal in September, lower in December and higher in June. This might suggest that the intervention does have an effect, but that it is nonlinear and is hence not picked up by the linear models. One might imagine that introducing a novel teaching approach is initially confusing for the students and/or the teachers, leading to lower test scores. As the year progresses the students and/or teachers settle into the new approach and improve teaching, learning and understanding, resulting in higher test scores. This is not a new idea (e.g. Vosniadou, 2009). Combining this with the reports of the teachers, describing their students as more attentive during the CT-lessons, a carefully positive answer can be given to RQ3.

All these results are in line with Tiruneh et al. (2018) who also found that an intervention, designed based on Merrill and including CT through infusion, had a positive effect on DSCT skills and content knowledge but no effect on DGCT skills.

The overall positive, or non-negative, results of the tests and reported enthusiasm of the students suggests that integrating attention to CT in physics lectures is worthwhile. The lack of transfer from one topic, physics, to DGCT may even suggest that CT should be integrated in the lessons of more topics or maybe even all. This idea is not new (e.g. Ennis, 2018).

Integrating DSCT into physics lessons might, however, be perceived by teachers as yet another thing they have to take up in their lessons, yet another thing that takes up precious time. We would argue that there are several reasons for teachers to consider

incorporating DSCT activities into their teaching. In addition to picking up important CT skills the students might be more attentive during classes and increase their content knowledge. They are, after all, forced to think very carefully about the physics content through the demonstration of the teacher and purposeful exercises. This means that the initial time investment in introducing the students to DSCT in physics and the (limited) time investment during classes throughout the year, might lead to time savings at the end of the year (or increased learning by the students). It is therefore not so that more time is needed to incorporate CT, but that teaching time should be spent differently. To be clear, we are very cautious here.

This study was exploratory and requires further work to confirm or reject the formulated conclusions. Among the limitations of this study are first and foremost the limited sample size, the convenience sampling and lack of fidelity control. Future research could improve on these aspects. Increasing sample size could be done by establishing several PLCs. Increasing fidelity control could be done through in class observations or through video-analysis. Finally, based on the limited qualitative results of this study it seems worthwhile to set up a purposeful qualitative study of both the teachers' and students' thoughts on and experience with teaching and learning CT.

## General conclusion and discussion

In this work a domain specific critical thinking test and an intervention to stimulate the domain specific critical thinking skills (DSCT) of secondary education students were developed and assessed. In doing so this work adds to the work of Tiruneh et al. (2017, 2018) by expanding this research line from higher education to secondary education.

From both studies 1 and 2 it is clear that the students that were tested showed very limited DSCT skills. One explanation might be that they are not adequately prepared, i.e. that students do not learn to think critically. We took steps in addressing this through a yearlong quasi-experimental intervention study (study 2). Results indicate that it is possible to increase students DSCT skills by infusing lessons with CT. However, such lessons did not result in increased DGCT skills.

Both studies were exploratory and have several limitations. Further work is required to confirm or reject the formulated conclusions. Suggestions for future research were formulated.

We cautiously conclude that it is possible to measure domain specific critical thinking, that students have very limited CT skills and that it is possible to increase students' domain specific critical thinking skills using carefully designed lessons infused with critical thinking.

## Acknowledgements

The authors would like to thank the teachers for their contributions to this research. Additionally they would like to thank Dawit Tiruneh for the fruitful discussions and AVL (Academisch Vormingscentrum voor Leraren, KU Leuven) for financially supporting this research (PZO-C9253-AVL/15/004 and PZO-C9254-AVL/15/005). Finally, the authors would like to thank the reviewers for their suggestions to improve the manuscript.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported by Academisch Vormingscentrum voor Leraren, KU Leuven: [Grant Number PZO-C9253-AVL/15/004,PZO-C9254-AVL/15/005].

## ORCID

Jan Sermeus  <http://orcid.org/0000-0002-5191-2590>

M. De Cock  <http://orcid.org/0000-0002-2489-1528>

J. Elen  <http://orcid.org/0000-0003-1611-5075>

## References

- Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. *Review of Educational Research*, 85(2), 275–314. <https://doi.org/10.3102/0034654314551063>
- Adams, W. K., & Wieman, C. E. (2011). Development and validation of instruments to measure learning of expert-like thinking. *International Journal of Science Education*, 33(9), 1289–1312. <https://doi.org/10.1080/09500693.2010.512369>
- Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. University of Chicago Press.
- Bailin, S. (2002). Critical thinking and science education. *Science & Education*, 11(4), 361–375. <https://doi.org/10.1023/A:1016042608621>
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Bensley, D. A., Crowe, D. S., Bernhardt, P., Buckner, C., & Allman, A. L. (2010). Teaching and assessing critical thinking skills for argument analysis in psychology. *Teaching of Psychology*, 37(2), 91–96. <https://doi.org/10.1080/00986281003626656>
- Butler, H. A., Pentoney, C., & Bong, M. P. (2017). Predicting real-world outcomes: Critical thinking ability is a better predictor of life decisions than intelligence. *Thinking Skills and Creativity*, 25, 38–46. <https://doi.org/10.1016/j.tsc.2017.06.005>
- Changwong, K., Sukkamart, A., & Sisan, B. (2018). Critical thinking skill development: Analysis of a new learning management model for Thai high schools. *Journal of International Studies*, 11(2), 37–48. <https://doi.org/10.14254/2071-8330.2018/11-2/3>
- Davies, M. (2013). Critical thinking and the disciplines reconsidered. *Higher Education Research & Development*, 32(4), 529–544. <https://doi.org/10.1080/07294360.2012.697878>
- Dunn, J. (2015). Critical thinking in Japanese secondary education: Student and teacher perspectives. *Critical Thinking and Language Learning*, 2(1), 29–39. <https://www.jaltcriticalthinking.org/ctl/about/>
- Elen, J., Jiang, L., Huyghe, S., Evers, M., Verburgh, A., Palaigeorgiou, G. (2019). In C. Dominguez, & R. Payan-Carreira (Eds.), *Promoting critical thinking in European higher education institutions: Towards an educational protocol*. UTAD.
- Ennis, R. H. (1989). Critical thinking and subject specificity: Clarification and needed research. *Educational Researcher*, 18(3), 4–10. <https://doi.org/10.3102/0013189X018003004>
- Ennis, R. H. (1993). Critical thinking assessment. *Theory Into Practice*, 32(3), 179–186. <https://doi.org/10.1080/00405849309543594>
- Ennis, R. H. (2009). *An annotated list of critical thinking tests*. [https://web.archive.org/web/20151025045645/http://faculty.education.illinois.edu/rhennis/TestListRevised11\\_27\\_09.htm](https://web.archive.org/web/20151025045645/http://faculty.education.illinois.edu/rhennis/TestListRevised11_27_09.htm)

- Ennis, R. H. (2011). *The nature of critical thinking: An outline of critical thinking dispositions and abilities*. <http://criticalthinking.net/wp-content/uploads/2018/01/The-Nature-of-Critical-Thinking.pdf>
- Ennis, R. H. (2018). Critical thinking across the curriculum: A vision. *Topoi*, 37(1), 165–184. <https://doi.org/10.1007/s11245-016-9401-4>
- Evens, M., Verburch, A., & Elen, J. (2014). The development of critical thinking in professional and academic bachelor programmes. *Higher Education Studies*, 4(2), 42–51. <https://doi.org/10.5539/hes.v4n2p42>
- Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction*. California Academic Press.
- Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, 33(2), 1–22. <https://doi.org/10.18637/jss.v033.i02>
- Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American Psychologist*, 53(4), 449–455. <https://doi.org/10.1037/0003-066X.53.4.449>
- Halpern, D. F. (2003). *Thought and knowledge: An introduction to critical thinking* (4th ed). Lawrence Erlbaum Associates Publishers.
- Halpern, D. F. (2010). *The Halpern critical thinking assessment: Manual*. Schuhfried GmbH.
- Halpern, D. F. (2014). *Critical thinking across the curriculum: A brief edition of thought & knowledge*. Routledge.
- Hatcher, D. L. (2011). Which test? Whose scores? Comparing standardized critical thinking tests. *New Directions for Institutional Research*, 2011(149), 29–39. <https://doi.org/10.1002/ir.378>
- Hieggelke, C. J., Maloney, D. P., Kanim, S. E., & O’Kuma, T. L. (2006). *E&M TIPERS: Electricity and magnetism tasks*. Pearson.
- Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 64(5), 402. <https://doi.org/10.4097/kjae.2013.64.5.402>
- Ku, K. Y. (2009). Assessing students’ critical thinking performance: Urging for measurements using multi-response format. *Thinking Skills and Creativity*, 4(1), 70–76. <https://doi.org/10.1016/j.tsc.2009.02.001>
- Lai, E. R. (2011). *Critical thinking: A literature review*. Pearson.
- McDermott, L. C., & Shaffer, P. S. (2002). *Tutorials in introductory physics: Homework*. Prentice Hall.
- McPeck, J. E. (1981). *Critical thinking and education*. Martin Robertson.
- Merrill, M. D. (2002). First principles of instruction. *Educational Technology Research and Development*, 50(3), 43–59. <https://doi.org/10.1007/BF02505024>
- Nunnally, J. C. (1978). *Psychometric theory*. McGraw-Hill Book Company.
- Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students: A third decade of research*. Vol 2. Jossey-Bass.
- Paul, R., & Elder, L. (2001). *The miniature guide to critical thinking concepts and tools*. The Foundation for Critical Thinking.
- Paul, R. W., & Binker, A. J. A. (1990). *Critical thinking: What every person needs to survive in a rapidly changing world*. Sonoma State University Press.
- Pithers, R. T., & Soden, R. (2000). Critical thinking in education: A review. *Educational Research*, 42(3), 237–249. <https://doi.org/10.1080/001318800440579>
- R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Revelle, W. (2019). *Psych: Procedures for personality and psychological research*. Northwestern University. <https://CRAN.R-project.org/package=psych> Version=1.9.12
- Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, 74(1), 145–154. <https://doi.org/10.1007/s11336-008-9102-z>
- Robinson, S. R. (2011). Teaching logic and teaching critical thinking: Revisiting McPeck. *Higher Education Research & Development*, 30(3), 275–287. <https://doi.org/10.1080/07294360.2010.500656>

- Rudd, R. D. (2007, October). Defining critical thinking. *Techniques*, 82(7), 46–49. <https://link.gale.com/apps/doc/A170157748/EAIM?u=anon~6bf6b408&sid=bookmark-EAIM&xid=31870d18>
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74(1), 107–120. <https://doi.org/10.1007/s11336-008-9101-0>
- Sugiarti, T., Kaniawati, I., & Aviyanti, L. (2017). Development of assessment instrument of critical thinking in physics at senior high school. *Journal of Physics: Conference Series*, 812(1), 012018. <https://doi.org/10.1088/1742-6596/812/1/012018>
- Thompson, C. (2011). Critical thinking across the curriculum: Process over output. *International Journal of Humanities and Social Science*, 1(9), 1–7. <https://www.ijhssnet.com/journal/index/263>
- Tiruneh, D. T., De Cock, M., Gu, X., & Elen, J. (2018). Systematic design of domain-specific instruction on near and far transfer of critical thinking skills. *International Journal of Educational Research*, 87, 1–11. <https://doi.org/10.1016/j.ijer.2017.10.005>
- Tiruneh, D. T., De Cock, M., Weldeclassie, A. G., Elen, J., & Janssen, R. (2017). Measuring critical thinking in physics: Development and validation of a critical thinking test in electricity and magnetism. *International Journal of Science and Mathematics Education*, 15(4), 663–682. <https://doi.org/10.1007/s10763-016-9723-0>
- van der Zanden, P. J., Denessen, E., Cillessen, A. H., & Meijer, P. C. (2020). Fostering critical thinking skills in secondary education to prepare students for university: Teacher perceptions and practices. *Research in Post-Compulsory Education*, 25(4), 394–419. <https://doi.org/10.1080/13596748.2020.1846313>
- Vlaamse overheid. (2010). VOET@2010: Nieuwe vakoverschrijdende eindtermen voor het secundair onderwijs [New cross curricular goals for secondary education]. <http://eindtermen.vlaanderen.be/publicaties/voet/voet2010.pdf>
- Vosniadou, S. (2009). *International handbook of research on conceptual change*. Routledge.
- Walsh, C., Quinn, K. N., Wieman, C., & Holmes, N. G. (2019). Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking. *Physical Review Physics Education Research*, 15(1), 010135. <https://doi.org/10.1103/PhysRevPhysEducRes.15.010135>
- Widhiarso, W., & Ravand, H. (2014). Estimating reliability coefficient for multidimensional measures: A pedagogical illustration. *Review of Psychology*, 21(2), 111–121. <http://psihologija.ffzg.unizg.hr/review>
- Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. <https://arxiv.org/abs/1308.5499>
- Yanti, T. D., Suana, W., Maharta, N., Herlina, K., & Distrik, I. W. (2019). Development of critical thinking instrument of electricity for senior high school students. *Journal of Physics: Conference Series*, 1157(3), 032007. doi:10.1088/1742-6596/1157/3/032007