Contemporary Educational Psychology 58 (2019) 310–322

Contents lists available at ScienceDirect

Contemporary Educational Psychology

journal homepage: www.elsevier.com/locate/cedpsych

Training higher education teachers' critical thinking and attitudes towards teaching it

Eva M. Janssen[a,*], Tim Mainhard[a], Renate S.M. Buisman[b], Peter P.J.L. Verkoeijen[c,d], Anita E.G. Heijltjes[d], Lara M. van Peppen[c], Tamara van Gog[a]

[a] Department of Education, Utrecht University, Heidelberglaan 1, 3584 CS Utrecht, the Netherlands
[b] Centre for Child and Family Studies, Leiden University, P.O. Box 9555, 2300 RB Leiden, the Netherlands
[c] Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, Burgemeester Oudlaan 50, 3062 PA Rotterdam, the Netherlands
[d] Learning and Innovation Center, Avans University of Applied Sciences, Hogeschoollaan 1, 4818 CR Breda, the Netherlands

ARTICLE INFO

Keywords:
Critical thinking
Heuristics and biases
Teaching and teacher education
Instructional design
Higher education

ABSTRACT

Teachers play a crucial role in attaining a major objective of higher education: fostering students' critical thinking (CT). Yet, little is known about how to foster teachers' own CT-skills and attitudes towards teaching CT. In a quasi-experimental study ($N = 54$), we investigated whether a three-session teacher training on (teaching) CT ($n = 32$) positively affected higher education teachers' CT-skills and their attitudes towards teaching CT compared to a control condition ($n = 22$). The training consisted of explicit instruction on common reasoning biases combined with assignments focused on the teaching practice. Results showed that the training improved teachers' performance on trained but not on novel CT-tasks. Also teachers' ability to detect biases in a written student product improved; however, despite a small improvement, they still had difficulties in correctly explaining those biases. Possibly due to ceiling effects the training did not affect perceived relevance of teaching CT. Finally, perceived competence in teaching CT decreased temporarily after the first training session but this negative effect disappeared after the final third session. Future research should investigate ways to promote teachers' ability to transfer trained skills to other CT-tasks, their ability provide feedback on students' reasoning (i.e., bias explanation), and their attitudes towards teaching CT.

1. Introduction

One of the major ambitions of higher education is to foster students' critical thinking (CT)[1] in order to prepare them for functioning in a complex and rapidly changing society. Indeed, CT-skills have been associated with higher levels of employment, a more sound financial situation, and stronger civic engagement (Arum, Cho, Kim, & Roksa, 2012; Toplak, West, & Stanovich, 2017). CT-skills, however, do not develop automatically as a 'by-product' of higher education. Results of two large-scale longitudinal studies ($N = 2322$ and $N = 2212$) in the United States showed that students' CT-skills hardly improved over the college years (Arum & Roksa, 2011; Pascarella, Blaich, Martin, & Hanson, 2011). This is perhaps not surprising given that CT-skills are rarely explicitly taught (Jones, 2007), whereas research has shown that students need explicit instruction to improve their CT-skills (Abrami et al., 2015; Heijltjes, Van Gog, Leppink, & Paas, 2014). Because teachers are responsible for providing this explicit instruction, they play an important role in students' acquisition of CT-skills. Even though reviews on teaching CT have highlighted the crucial role of the teacher in this process (Abrami et al., 2015, 2008; Pithers & Soden, 2000; Ritchhart & Perkins, 2005) there is a paucity of research focusing on teachers' CT-skills. Moreover, the limited research that is available suggested that higher education teachers may not have a concrete understanding of what CT encompasses and how they can teach it (Choy & Cheah, 2009; Stedman & Adams, 2012). A prerequisite for being able to provide instruction and guidance to students on a subject is that teachers themselves possess the required skill, and that their attitude towards teaching it is positive: they need to perceive it as a highly relevant to teach and identify themselves as self-competent in teaching it (Eccles & Wigfield, 2002; Klassen & Tze, 2014; Watt & Richardson, 2007; Zee & Koomen, 2016). In the present study, we examined whether these preconditions can be positively affected through a

* Corresponding author.
E-mail addresses: e.m.janssen@uu.nl (E.M. Janssen), m.t.mainhard@uu.nl (T. Mainhard), r.s.m.buisman@fsw.leidenuniv.nl (R.S.M. Buisman), p.p.j.l.verkoeijen@essb.eur.nl, ppjl.verkoeijen@avans.nl (P.P.J.L. Verkoeijen), aeg.heijltjes@avans.nl (A.E.G. Heijltjes), vanpeppen@essb.eur.nl (L.M. van Peppen), t.vangog@uu.nl (T. van Gog).

[1] CT = critical thinking.

https://doi.org/10.1016/j.cedpsych.2019.03.007

Available online 21 March 2019

training for higher education teachers (i.e., postsecondary teachers) consisting of explicit instruction on common reasoning biases combined with the opportunity for practice, along with assignments focused on their teaching practice.

1.1. Critical thinking and cognitive biases

In the CT literature, scholars have viewed the ability to evaluate evidence and arguments independently of one's prior beliefs and opinions as an important aspect of CT (Baron, 2008; Ennis, 1987; Perkins, Tishman, Ritchhart, Donis, & Andrade, 2000; Sternberg, 2001). This is also illustrated in tests that measure CT, in which an important component consists of assessing the ability to avoid reasoning that is too biased by prior opinion and prior belief (Ennis, Millman, & Tomko, 1985; Facione, 1990; Watson & Glaser, 1980; West, Toplak, & Stanovich, 2008). In addition, cognitive theorists have analyzed critical thinking in terms of rational thinking concepts and the philosophy of rational thought (Kuhn, 2005; Siegel, 1988). Biases violate the normative rules of rationality, as set, for instance, by logic or probability (Stanovich, West, & Toplak, 2016; Tversky & Kahneman, 1974). For example, due to the base-rate neglect bias, most people are more concerned about the risks of terrorism than about statistically larger risks that they confront in daily life (Sunstein, 2003). Although biases are inherent to human cognition and often innocent, in some situations they lead to decisions that have serious consequences. For example, when a judge misinterprets statistical evidence (Thompson & Schumann, 1987) or when a doctor makes a biased decision in medical diagnosis (Schmidt et al., 2014). In the present study we focus on this essential aspect of CT: the ability to avoid bias in reasoning and decision-making (i.e., rational thinking).

To assess biases in thinking, researchers have designed heuristics-and-biases tasks (Tversky & Kahneman, 1974), consider for example the following (Frey, Johnson, & De Neys, 2017):

> In a study 1000 people were tested. Among the participants there were 5 sixteen-year-olds and 995 forty-year-olds. Lisa is a randomly chosen participant of the study. Lisa likes to listen to techno and electro music. She often wears tight sweaters and jeans. She loves to dance and has a small nose piercing.
>
> What is most likely?
> Lisa is sixteen
> Lisa is forty

Because the description is very representative of a sixteen-year-old, a majority of the university students who were given this problem incorrectly indicated that Lisa is most likely sixteen. As explained by dual processing theories (Evans, 2008; Kahneman & Frederick, 2005) one needs to replace a heuristic (Type 1) response "this description fits with the image of adolescents, Lisa is probably sixteen" with a more effortful logical (Type 2) response "there are also forty-year-olds that listen to techno wearing tight jeans, and since the study sample consisted of 995 forty-year-olds compared to only 5 sixteen-year-olds, it is most likely that Lisa is forty". Stanovich et al. (2016) argued that this shifting from a heuristic to a normative response requires a disposition towards rational thinking (e.g., actively open-minded thinking) and sufficient working memory capacity, but also what they referred to as 'mindware', that is, knowledge and skills needed for correct reasoning (e.g., of logic/probability).

1.2. Teaching critical thinking

Research on the effectiveness of CT-instruction has mainly focused on students. The results of meta-analyses showed that the most effective type of CT-intervention for student outcomes was a combination of authentic instruction, dialogue, and mentoring (Abrami et al., 2015) and that the most effective pedagogical grounding of the CT-intervention was achieved when instructors received special advanced training in preparation for teaching CT-skills (Abrami et al., 2008). Regarding the specific skill to avoid biases in reasoning, experimental studies with students have shown that explicit instruction about cognitive biases (through a video or a text) combined with the opportunity for task practice improved students' performance on tasks addressing these same biases compared to a control condition, both at an immediate posttest (Heijltjes, Van Gog, & Paas, 2014; Heijltjes, Van Gog, Leppink, & Paas, 2014, 2015) and on a delayed posttest two weeks later (Van Peppen et al., 2018). A problem with all CT-interventions (both on avoiding bias and on CT-skills in general), however, is that the effects hardly seem to transfer across tasks or contexts (Heijltjes et al., 2014; Heijltjes, Van Gog, Leppink, & Paas, 2014, 2015; Kenyon & Beaulac, 2014; Ritchhart & Perkins, 2005). A lack of transfer is problematic because the ultimate goal of CT-teaching is that students can apply the learned CT-skills in contexts outside the school context. Scholars argued that, to achieve transfer, thinking skills need to be taught with many different types of (real-world) examples and corrective feedback, while highlighting the underlying principles, so that students learn to recognize when and what thinking skill is needed in a particular situation (Halpern, 1998; Ritchhart & Perkins, 2005). Hence, for students to achieve transfer, their teachers need to be very skilled in (teaching) CT, and able to bridge across various tasks and contexts.

In sum, teachers play an important role in students' CT-skills acquisition. In order to be able to teach CT, teachers first need to possess CT-skills themselves so that they can provide explicit instruction and integrate CT in their lessons (e.g., through dialogues, and feedback on students' reasoning). Based on the previous experiments on avoiding bias (Heijltjes, Van Gog, Leppink, & Paas, 2015; Heijltjes et al., 2014; Van Peppen et al., 2018), we expect explicit instruction of cognitive biases to improve teachers' performance on heuristics-and-biases tasks. However, as the previous experiments used student samples, it is an open question whether this indeed is the case. Moreover, for teachers to be able to engage their students in dialogue about biases and provide adequate feedback to their students, they also need to be able to detect biases in their students' reasoning and to explain errors students make. Finally, for teachers to apply CT-skills (acquired during the training) in practice, they need to have positive attitudes regarding CT-teaching. More specifically, in line with expectancy value theory (Eccles & Wigfield, 2002), research showed that teachers with a positive attitude towards the relevance of teaching a particular subject (i.e., high task value) and confidence in one's ability (i.e., high expectancy of success) were more likely to engage in effective teaching (Choy & Cheah, 2009; Klassen & Tze, 2014; Paul, Elder, & Bartell, 1997; Van Aalderen-Smeets & Walma van der Molen, 2013, 2015). Although this has not yet been investigated in the domain of CT-teaching, an experimental study in the domain of science teaching showed that interventions can positively affect teaching attitudes: a training focused on changing primary school teachers' professional attitudes had a large effect on perceived relevance and self-efficacy beliefs regarding science teaching as well as on self-reported teaching behavior (Van Aalderen-Smeets & Walma van der Molen, 2015).

Thus, next to investigating whether training affects teachers' CT-skills as measured with heuristics-and-biases tasks, we also investigated whether teachers' ability to recognize and explain students' reasoning errors and their attitudes towards teaching CT improved.

1.3. The present study and hypotheses

The goal of the present study was to gain insight in how to equip higher education teachers with the skills and attitudes necessary for teaching CT. We explored the potential impact of a training consisting of three sessions of three hours each. The first session provided explicit instruction on cognitive biases and task practice (e.g., Heijltjes et al., 2015). The other two sessions focused on strengthening teachers' attitudes and skills towards teaching CT through discussing the relevance of teaching CT, providing extra opportunity for practice, designing a domain-specific CT-task, and discussing ways to integrate CT during teaching. Fig. 1 displays an overview of the study design: we measured teachers' CT-skills and teaching attitudes before the start of the training, after its first session, and after its final third session (and at similar points in time in an untreated control condition). CT-skills were

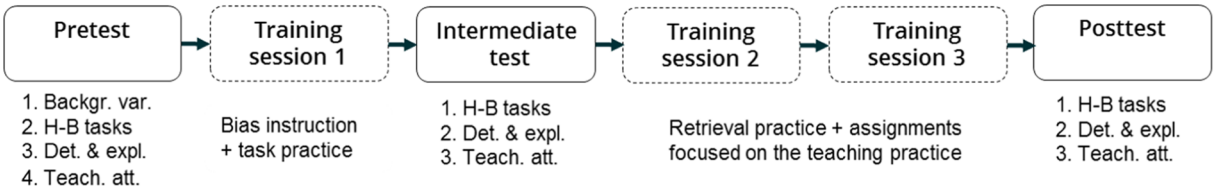


Fig. 1. Overview of the study design. The control condition did not take part in the training sessions (dashed lines), only in the measurement occasions (solid lines). Backgr. var. = background variables; H-B tasks = heuristics-and-biases tasks; Det. & Expl. = Detecting and explaining reasoning biases in a student product; Teach. att. = Attitudes towards teaching critical thinking.

measured as (1) performance on heuristics-and-biases tasks, and (2) ability to detect and explain reasoning biases in a written student product (vignette).

First, we hypothesized that the explicit instruction on cognitive biases and the opportunity for task practice that the training provided, would positively affect teachers' performance on instructed tasks (i.e., learning; Hypothesis 1). We explored whether the training would lead to better performance on not-instructed tasks (transfer); although findings on effects of training on transfer were mixed so far, the generative learning activities provided in the additional sessions (e.g., designing a CT-task) may enhance transfer (Fiorella & Mayer, 2016). Second, we expected that the training would positively affect teachers' ability to detect (hypothesis 2a) and to explain (hypothesis 2b) reasoning biases in a student product. Third, regarding teaching attitudes, we expected (based on findings from science teaching, see previous section) that CT-instruction in combination with specific attention to the teaching practice would positively affect teachers' perceived relevance of (hypothesis 3a) and perceived competence in (3b) teaching CT.

To gain further insight in how the training affected teachers' task approach, we explored self-reported mental effort investment in relation to teachers' CT-task performance. Previous studies consistently found that instruction resulted in a more efficient way of solving the learning tasks, as indicated by a an improved task performance from pretest to posttest on instructed tasks without investing more mental effort in solving the tasks (Heijltjes et al., 2014, 2015; Van Peppen et al., 2018). For transfer tasks, one study found no effects of instruction on invested mental effort (Heijltjes et al., 2014), whereas two other studies showed that – after instruction – invested effort increased, yet without a performance improvement (Heijltjes et al., 2015; Van Peppen et al., 2018). This may indicate that training stimulated students to replace a heuristic (Type 1) response with a more effortful (Type 2) response to the new, not-instructed tasks (which would arguably be an important outcome of training), but that their mindware was not sufficiently automatized to correctly perform the task.

2. Method

2.1. Participants and design

Participants were 56 teachers from a Dutch University of Applied Sciences[2]. Two participants were excluded (one was not proficient in Dutch and one was a policy maker instead of a teacher), leaving a final sample of 54 teachers (63.0% female; age: $M = 45.0$ years, $SD = 9.4$; teaching experience: $M = 8.4$ years, $SD = 6.7$). The study had a quasi-experimental design, with a training condition ($n = 32$) and a control condition ($n = 22$). The teachers in the training condition voluntarily signed up for participation in the CT-training that consisted of three sessions of approximately three hours, spread over six weeks (the second session was three weeks after the first, the third session was two weeks after the second). Participation was recommended and endorsed by their faculty management. The training was given in three separate groups. Group A ($n = 14$) consisted of teachers from one department, teachers in Group B ($n = 11$) and C ($n = 7$) came from different departments. The teachers in the control condition volunteered to take the tests and did not receive any training. The study consisted of six phases: (1) pretest; (2) first training session (3); intermediate test; (4) second training session; (5) third training session; (6) posttest. Teachers in the control condition only engaged in phase 1, 3, and 6 at approximately similar time intervals (see Fig. 1).

2.2. Materials

2.2.1. Pretest, intermediate test, and posttest

The tests were administered as an online survey with a forced response-format using Qualtrics Survey Software (Qualtrics, Provo, UT; http://www.qualtrics.com). The pretest addressed four topics in a fixed order: background variables, attitudes towards teaching CT, bias detection and explanation in a student product, and heuristics-and-biases tasks. The intermediate test and posttest addressed the following outcome measures: teaching attitudes, bias detection and explanation, and heuristics-and-biases tasks.

2.2.1.1. Background variables. Because we were not able to randomly assign participants to conditions, we collected information on some background variables to check the comparability of the subsamples: gender, age (years), teaching experience (years), teaching domain, CT-experience, relevant experience, and thinking dispositions. Response categories for *teaching domain* were (1) technology/ICT; (2) economics/HRM/business administration; (3) social studies; (4) legal studies; other, namely_. These were merged into three broad domain-categories which paralleled the sections of the University of Applied Sciences: technology, economics, and society. For *CT-experience*, teachers answered the question "For how many hours (estimation) have you been actively involved with the theme 'thinking errors' in the past two years (think for example of following a workshop, teaching lessons, or developing lesson materials)?" Answering categories were: 0 h, 1–20 h, 21–40 h, 41–60 h, 61–80 h, 81–100 h, or > 100 h. For *relevant experience,* teachers answered four yes/no questions that asked whether they had taught statistics, logic or programming and whether their job included the assessment of written student products.

Finally, we measured teachers' rational *thinking dispositions* with Dutch translations (Heijltjes et al., 2014) of two questionnaires: the 18-item short form of the Need For Cognition scale (NFC; Cacioppo, Petty, & Feng Kao, 1984) and the 41-item Actively Open-minded Thinking scale (AOT; Stanovich & West, 2007). NFC intends to measure tendency to engage in and enjoy thinking and the AOT high-level epistemic goals and the tendency to reflect on rules of inference. Participants rated their agreement to the 59 statements in total on a six-point rating scale ranging from (1) strongly disagree to (6) strongly agree. Scores on the items were averaged for NFC and AOT separately (after reverse scoring items that were formulated negatively) and could therefore range from 1 to 6. In the current study, the NFC had a Cronbach's alpha of 0.79 and the AOT of 0.78.

2.2.1.2. Heuristics-and-biases tasks. In line with previous research on the ability to avoid bias in reasoning and decision-making, we used

[2] The Dutch education system distinguishes between higher education at an academic university (i.e., MSc at a research university) or non-academic university (i.e., university of applied sciences).

several heuristics-and biases tasks as measures of critical thinking (Heijltjes et al., 2014; Tversky & Kahneman, 1974; West et al., 2008). Each test contained fifteen common heuristics-and-biases tasks (example-items of all task categories are available in Appendix A). All tasks at pretest, intermediate, and posttest were designed as structurally equivalent tasks but with different surface features. Two categories reflected *learning tasks* that were instructed and practiced during the training: (1) logic reasoning tasks in which belief bias played a role (7x), these tasks consisted of syllogisms that examined the tendency to be influenced by the believability of a conclusion when evaluating the logical validity of arguments and had a multiple-choice format with two answer options and one correct answer (adapted from Evans, 2002); and (2) tasks that assessed base-rate neglect in probability estimation (3x), which measure the tendency to overrate individual-case evidence (e.g., from personal experience, a single case, or prior beliefs) and to underrate statistical information. These tasks had multiple-choice formats with two, four, or six answer options where only a specific combination of selected answers was correct for the latter two (adapted from Fong, Krantz, & Nisbett, 1986; Stanovich & West, 2000; Stanovich et al., 2016; Tversky & Kahneman, 1974). We included a higher number of syllogisms because the chance of guessing the answer correctly was higher than for the base-rate tasks. Learning-task performance was computed as a percentage score to which both task categories contributed equally. Cronbach's alpha for the learning tasks was 0.37, and 0.43 on the intermediate test and posttest respectively.

Additionally, two task categories reflected *transfer tasks* (i.e., not addressed during the training): (1) tasks that assessed confirmation bias in logic reasoning (3x), which were Wason selection tasks that measure the tendency to verify logic rules rather than to falsify them, using a multiple-choice format with four answer options in which only a specific combination of two selected answers was correct (adapted from Evans, 2002; Gigerenzer & Hug, 1992) and (2) tasks that assess covariation detection in probability estimation (2x), which measure the tendency to base estimations on already experienced evidence and disregard presented evidence (multiple-choice format with two answer options and one correct answer; adapted from Heijltjes et al., 2014; Stanovich & West, 2000; Wasserman, Dorner, & Kao, 1990). Transfer-task performance was computed as a percentage score to which both task types contributed equally. Cronbach's alpha for the transfer tasks was 0.26 and 0.61, on the intermediate test, and delayed posttest, respectively.

After each task, participants reported their invested mental effort on a 9-point rating scale ranging from (1) very, very low effort to (9) very, very high effort (Paas & Van Merriënboer, 1993; Paas, 1992).

2.2.1.3. Detecting and explaining reasoning biases. We constructed three vignettes – one for each measurement occasion – to assess teachers' recognition of reasoning biases in a written student product. Each vignette (about 600 words) was in the form of a summary of a bachelor thesis and contained five biases that were also addressed in the heuristics-and-biases tasks: belief bias (2x) and confirmation bias (1x) in logic reasoning and base-rate neglect (1x) and a covariation detection problem (1x) in probability estimation. Teachers were instructed to read the text carefully, to indicate all reasoning biases in the text, and to provide an explanation. Maximum score for each bias was two points, one point for detecting a reasoning bias (*bias detection*) and one point for a correct explanation (*bias explanation*). We pilot-tested one vignette and designed the other two as structurally equivalent but with the biases in a different order and different surface features. Bias detection and explanation were scored by two raters who coded 25% of the data. We found, respectively, substantial and almost perfect agreement (Landis & Koch, 1977), ICC 0.63 and ICC = 0.83. The remainder of the data was scored by one rater (the first author). The final scores for bias detection and explanation were sum-scores (range: 0–5 each).

2.2.1.4. Attitudes towards teaching. Attitudes towards teaching CT were measured with a questionnaire that originally consisted of 16 items, of which we selected 6 pretest-items (for details, see Appendix B) that addressed teachers' perceived relevance of (3 items) and perceived competence in (3 items) teaching CT. The three items for relevance perception were a slightly adapted Dutch translation of items from Stedman and Adams (2012), who measured teachers' perceptions of CT instruction. An example-item is "Learning outcomes will improve from critical thinking during educational activities." The three competence perception items were constructed by the authors, an example item is "I can explain clearly to my students how they are drawing incorrect conclusions from the available information." Participants rated their agreement to the statements on a six-point rating scale ranging from (1) strongly disagree to (6) strongly agree. Averaged scores on the perceived relevance and perceived competence scales could therefore range from 1 to 6. Per measurement occasion Cronbach's alpha for Perceived relevance was 0.71, 0.75, and 0.80 and for Perceived competence 80, 0.77, and 0.70, respectively. A CFA showed that a two-factor model fitted the data well for each measurement occasion, χ^2s (8) ≤ 9.81, *ps* ≥ 0.279, CFIs ≥ 0.97, TLIs ≥ 0.95, RMSEAs ≤ 0.07, SRMRs ≤ 0.06.

2.3. Training and procedure

The CT-training consisted of three sessions of approximately three hours, spread over six weeks. All sessions were given by the first author together with the fourth or fifth author. Two weeks before the start of the first training session of Group A (Group B started two days later; Group C two weeks later), all participants – both in the training and control condition – received a request via email to complete the pretest.

The *first session* provided explicit instruction combined with practice on CT-skills. The session consisted of four parts: (1) a general introduction on CT, heuristics and biases; (2) explicit instruction on belief bias in syllogistic reasoning and base-rate neglect in probability estimation (i.e., we presented multiple statements to which the teachers could respond whether or not the conclusions were correct according to them; hereafter we explained the correct answers using worked-examples); (3) the opportunity for practice on syllogistic reasoning problems; and (4) an intermediate test. During the last hour of the first training session, teachers in the training condition completed the intermediate test on their laptop in the same room. That same week, teachers in the control condition received a request via email to complete the intermediate test.

The *second and third session* focused on strengthening teachers' attitudes and skills towards teaching CT through providing extra opportunity for practice, discussing the relevance of teaching CT, designing a domain-specific CT-task, and discussing ways to integrate CT during teaching. The second session consisted of (1) retrieval practice of the tasks from the first session; (2) an introduction on CT in education highlighting empirical findings on effective teaching strategies and learning outcomes from CT-instruction; and (3) a small-group assignment to design a CT task in the teachers' own teaching domain. The third session consisted of (1) discussion of the designed tasks; (2) discussing ways to integrate CT during teaching; (3) posttest; (4) evaluation. One hour of the third training session was again reserved for a posttest and teachers in the control condition again received a request via email to complete the posttest that same week.

2.4. Data analysis

To examine how the training affected teachers' CT-skills and variables related to teaching it, we employed multilevel analyses for each outcome measure. Multilevel-analysis handled the 11% posttest drop-out (see below) without excluding cases list-wise (Hox, 2010). We used the 'lme4' package in R (Bates, Mächler, Bolker, & Walker, 2015; R Development Core & Team, 2008) and employed likelihood ratios to

evaluate fit between a multilevel model and the data. For every outcome measure, we started with an intercept-only or so-called empty model (M1) including no predictors, that decomposed the variance of the outcome measure in two separate components: variance attributed to random error and to differences between the measurement occasions (level 1; σ^2_e) and variance attributed to stable differences between the teachers (level 2; σ^2_{u0}). We used this model to calculate the proportion of variance located at the teacher level, the intraclass correlation (ICC) = $\sigma^2_{u0}/(\sigma^2_{u0} + \sigma^2_e)$. Next, we tested an unconditional growth model with Occasion as fixed predictor (dummy-coded with the pretest as the reference category; M2), that tested a main effect of Occasion. Third[3], we added a cross-level interaction between Occasion and Condition (with the control condition as the reference category; M3) which tested our hypotheses that teachers' progress on the outcome variables would differ from pretest to intermediate test and pretest to posttest, depending on being in the training or control condition. To compute effect sizes (R^2_2), we used M2 as baseline model and calculated how much of the variance between teachers' scores could be explained by the predictors in the final model, as advised Hox (2010), $R^2_2 = (\sigma^2_{u0(M2)} - \sigma^2_{u0(M3)})/\sigma^2_{u0(M2)}$, for which we considered 0.01, 0.09, and 0.25 a small, medium, and large effect, respectively (Cohen, 1988).

2.4.1. Assumptions

Some variables did not meet the required assumptions for multilevel analysis. First, we identified one univariate outlier for relevance perception (at intermediate test) with a standardized score < -3.29. We ran the analyses both without and with winsorizing this outlier (i.e., subtracting the difference between the two next lowest values of the next lowest value with standardized value > -3.29; Tabachnick & Fidell, 2014), yielding similar results (results from the original data are reported). We did not identify any multivariate outliers using Mahalanobis' distance. Second, the assumption of univariate normality was violated for bias detection (moderate negative skewness) at the posttest and for bias explanation (moderate positive skewness) at the pretest and the intermediate test. When included in the multilevel models, we applied square root transformations on these variables (Tabachnick & Fidell, 2014). As bias detection was negatively skewed, the transformation was on the reflected variables (i.e., the subtracting each score from the largest score plus 1) that we re-reflected after transformation to enhance interpretation. After inspecting Q-Q plots (i.e., a probability plot of the standardized residuals against the values that would be expected under normality) for each multilevel model, we concluded that multivariate normality was met.

2.4.2. Absence and missing data

All 54 teachers filled out the pretest and the intermediate test and a final sample of 49 teachers completed the posttest. Within the control condition, three teachers dropped out after the intermediate test (although one did fill out the posttest teaching-attitudes questionnaire). Within the training condition, all 32 teachers attended the first session and completed the pretest and intermediate test, and 30 teachers completed the posttest. Eight teachers were unable to fully attend (one of) the next sessions (one could not attend session 2; six could not fully attend session 3; and one could not fully attend both sessions). Five of the seven teachers who did not (fully) attend session 3 were willing to complete the posttest online at home. We ran analyses both with and without the data of those (partly) absent teachers to assess how absence affected the results (dissimilar results are reported).

3. Results

To check the comparability of the experimental conditions, we first tested for significant differences on the background variables (Table 1) and on the outcome variables at the pretest. The training condition consisted of significantly more females and teachers from the economical teaching domain than the control condition (all teachers in training Group A were females from the same department). There were no other significant differences and the conditions did not differ on the pretest outcome variables, $ts \leq 1.63$, $ps \geq 0.109$. Table 2 displays the pretest, intermediate test, and posttest data per condition (Tables C1 and C2 in Appendix C additionally display performance on the heuristics-and-biases and the vignettes for each task category separately).

Table 1
Comparison of participant characteristics between conditions.

Chi-square tests to compare distribution	Control %	Training %	$\chi^2(1)$	p
Gender – female	46	75	4.88	0.027
Teaching domain[a]	–	–	15.32	< 0.001
Experience critical thinking[b]	–	–	1.51	0.220
Experience teaching statistics – yes	14	19	0.25	0.620
Experience teaching logic – yes	9	13	0.15	0.695
Experience teaching programming – yes	14	9	0.24	0.624
Experience student product assessment – yes	91	100	3.02	0.082
Independent t-tests to compare means	***M* (*SD*)**	***M* (*SD*)**	***t*(52)**	***p***
Age (years)	46.4 (10.6)	44.0 (8.5)	0.9	0.373
Teaching experience (years)	7.0 (6.8)	9.5 (6.5)	−1.4	0.182
Need for Cognition (range: 1–6)	4.5 (0.5)	4.5 (0.5)	−0.4	0.705
Actively Open-minded Thinking (range: 1–6)	4.7 (0.3)	4.8 (0.3)	−0.6	0.580

Note. $n = 22$ and $n = 32$ for the control and training condition, respectively.
[a]Domains were Technology ($n = 10$), Economics ($n = 29$), and Society ($n = 15$) with 2 degrees of freedom in Chi-square test instead of 1.
[b]Consisted of 6 ordinal answer categories, therefore we used a chi square test for trend for comparison.

[3] Initially, we intended to run a model with Time as a random predictor before adding the cross-level interaction, but we lacked power to test for random slope variance, which is a common issue with categorical predictors. As recommended by LaHuis and Ferguson (2009), we checked further for cross-level interaction effects because fixed effects are estimated with more precision.

Table 2
Means (M) and Standard Deviations (SD) of the Outcome Variables at Pretest, Intermediate Test, and Posttest per Condition.

	Control $n = 22$[a]		Training $n = 32$[b]	
	M	*SD*	*M*	*SD*
Learning tasks (range: 0–100%)				
Pretest	53.7	21.1	61.5	14.4
Intermediate test	57.9	16.8	74.7	13.3
Posttest	52.9	20.1	73.1	17.8
Transfer tasks (range: 0–100%)				
Pretest	48.9	13.2	42.7	21.8
Intermediate test	54.2	16.4	51.0	21.1
Posttest	53.5	18.9	43.3	29.1
Mental effort learning tasks (range: 1–9)				
Pretest	3.9	0.9	4.2	1.2
Intermediate test	4.2	1.0	4.8	1.2
Posttest	3.9	1.0	4.3	1.2
Mental effort transfer tasks (range: 1–9)				
Pretest	5.3	1.0	5.0	1.6
Intermediate test	5.0	1.6	5.5	1.4
Posttest	4.3	1.4	5.6	1.6
Bias detection (range: 0–5)				
Pretest	2.6	1.5	2.8	1.1
Intermediate test	1.6	1.3	2.4	1.2
Posttest	2.8	1.5	4.6	0.5
Bias explanation (range: 0–5)				
Pretest	1.3	1.5	1.1	1.0
Intermediate test	0.3	0.6	0.7	0.9
Posttest	0.6	0.7	1.8	1.3
Perceived relevance (range: 1–6)				
Pretest	5.0	0.5	5.2	0.5
Intermediate test	5.0	0.7	5.1	0.5
Posttest	4.8	0.8	5.1	0.6
Perceived competence (range: 1–6)				
Pretest	4.3	0.7	4.0	0.8
Intermediate test	4.2	0.6	3.6	0.9
Posttest	4.1	0.7	4.0	0.6

[a] At posttest $n = 19$. [b] At posttest $n = 30$.

3.1. Heuristics-and-biases tasks

Average pretest performance on the learning and transfer tasks was, respectively, 58.3% ($SD = 17.7$) and 45.2% ($SD = 18.9$). Table 3 summarizes the results of the multilevel analyses that tested whether the training improved teachers' performance on learning (Hypothesis 1) and transfer.

3.1.1. Learning tasks

The ICC (M1) for learning-task performance revealed that 45% of the variance in performance on the instructed heuristics-and-biases tasks could be attributed to stable differences between teachers (level 2). Both adding Occasion as a predictor (M2) and, subsequently, adding a cross-level interaction between Occasion and Condition (M3) improved model fit significantly, M2: $\chi^2(2) = 14.05$, $p < .001$; M3: $\chi^2(3) = 21.65$, $p < .001$. The results revealed that, as compared to the control condition, teachers in the training condition did not improve significantly more on learning tasks from the pretest to the intermediate test, but did improve significantly more from pretest to posttest, Pre-Intermediate × Condition: $t(149) = 1.82$, $p = .070$; Pre-Post × Condition: $t(149) = 2.61$, $p = .010$. Thus, the training positively affected teachers' learning-task performance but this effect was apparent only after all three sessions (posttest). R_2^2 indicated that the predictors in the final model (M3) explained 28% of the variance between teachers' learning-task performance, which is a large overall effect. When excluding the teachers who were (partially) absent during the second or the third session, training effects were also statistically significant after the first session, Pre-Intermediate × Condition: $t(131) = 2{,}13$, $p = .035$; Pre-Post × Condition: $t(131) = 2.63$, $p = .010$.

3.1.1.1. Mental effort. As compared to the pretest, teachers in both conditions invested more mental effort in the learning tasks at the intermediate test ($p < .001$) but not at the posttest (see Table C3, Appendix C).

3.1.2. Transfer tasks

The ICC for transfer-task performance indicated that 37% of the variance in performance on the not-instructed heuristics-and-biases

Table 3
Effect of critical thinking training on teachers' performance on learning and transfer (heuristics-and-biases) tasks.

	M1: intercept only	M2: effect occasion	M3: effect of training
Learning tasks			
Fixed part	Coefficient (s.e.)	Coefficient (s.e.)	Coefficient (s.e.)
Intercept	63.62 (2.05)**	58.33 (2.50)**	53.68 (3.55)**
Pre-Intermediate		9.52 (2.50)**	4.22 (3.78)
Pre-Post		6.44 (2.59)*	−1.65 (3.97)
Condition			7.85 (4.61)
Pre-Intermediate × Condition			8.95 (4.91)
Pre-Post × Condition			13.28 (5.10)*
Random part			
Occasion and error variance (σ^2_e)	193.5 (13.91)	169.0 (13.00)	157.0 (12.53)
Teacher variance (σ^2_{u0})	160.1 (12.65)	167.8 (12.96)	120.3 (10.97)
Deviance	1338.2	1324.1	1302.5
Transfer tasks			
Fixed Part	Coefficient (s.e.)	Coefficient (s.e.)	Coefficient (s.e.)
Intercept	48.36 (2.23)**	45.22 (2.88)**	48.86 (4.45)**
Pre-Intermediate		7.10 (3.19)*	5.30 (4.97)
Pre-Post		2.23 (3.29)	4.75 (5.22)
Condition			−6.16 (5.78)
Pre-Intermediate × Condition			3.03 (6.46)
Pre-Post × Condition			−4.02 (6.70)
Random part			
Occasion and error variance (σ^2_e)	288.4 (16.98)	274.9 (16.58)	271.9 (16.49)
Teacher variance (σ^2_{u0})	167.9 (12.96)	172.0 (13.12)	163.5 (12.79)
Deviance	1388.2	1383.2	1380.1

Note. Occasion is dummy-coded in two dummies (Intermediate test, Posttest) with Pretest as reference category. Condition coded 0 = control, 1 = training.
*$p < .05$. **$p < .001$.

tasks was due to stable differences between teachers. However, neither adding Occasion (M2) nor a cross-level interaction between Occasion and Condition (M3) could explain these differences, M2: $\chi^2(2) = 5.04$, $p = .080$; M3: $\chi^2(5) = 8.11$, $p = .150$.

3.1.2.1. Mental effort. As compared to the pretest, teachers in the training condition invested more effort in the transfer tasks than teachers in the control condition at the intermediate test ($p = .018$) and at the posttest ($p < .001$; see Table C3, Appendix C).

3.2. Detecting and explaining reasoning biases

At the pretest, teachers detected 2.7 ($SD = 1.3$) out of five biases in a student product and correctly explained 1.2 ($SD = 1.3$) biases. Both condition-averages decreased from pretest to intermediate test, whereas only the training condition increased from pretest to posttest on bias detection and explanation (Table 2). At all tests, the average number of correctly explained biases remained substantially behind the number of detected biases. Table 4 summarizes the analyses that tested whether the training significantly improved teachers' detection and explanation of reasoning biases (hypothesis 2a and 2b).

Table 4
Effect of critical thinking training on teachers' ability to detect and explain reasoning biases in a student product.

	M1: intercept only	M2: effect occasion	M3: effect of training
Bias detection (sq. root)			
Fixed part	Coefficient (s.e.)	Coefficient (s.e.)	Coefficient (s.e.)
Intercept	1.73 (0.04)**	1.67 (0.05)**	1.64 (0.07)**
Pre-Intermediate		−0.18 (0.06)*	−0.27 (0.08)*
Pre-Post		0.40 (0.06)**	0.09 (0.09)
Condition			0.05 (0.09)
Pre-Intermediate × Condition			0.16 (0.11)
Pre-Post × Condition			0.50 (0.11)**
Random part			
Occasion and error variance (σ^2_e)	0.17 (0.4)	0.09 (0.30)	0.08 (0.27)
Teacher variance (σ^2_{u0})	0.02 (0.12)	0.05 (0.22)	0.03 (0.19)
Deviance	183.8	116.4	85.8
Bias explanation (sq. root)			
Fixed Part	Coefficient (s.e.)	Coefficient (s.e.)	Coefficient (s.e.)
Intercept	0.74 (0.06)**	0.84 (0.09)**	0.83 (0.13)**
Pre-Intermediate		−0.39 (0.11)**	−0.53 (0.16)*
Pre-Post		0.13 (0.11)	−0.31 (0.17)
Condition			0.02 (0.17)
Pre-Intermediate × Condition			0.25 (0.21)
Pre-Post × Condition			0.72 (0.21)**
Random part			
Occasion and error variance (σ^2_e)	0.39 (0.62)	0.31 (0.56)	0.28 (0.53)
Teacher variance (σ^2_{u0})	0.07 (0.26)	0.10 (0.31)	0.09 (0.30)
Deviance	319.6	298.0	280.1

Note. Occasion is dummy-coded in two dummies (Intermediate test, Posttest) with Pretest as reference category. Condition coded 0 = control, 1 = training. *$p < .05$. **$p < .001$.

3.2.1. Bias detection

The ICC for bias detection revealed that only 8% of the variance in teachers' ability to detect biases was due to stable differences between teachers. Both Occasion (M2) and the Occasion × Condition interaction (M3) improved the model fit significantly, M2: $\chi^2(2) = 67.40$, $p < .001$; M3: $\chi^2(3) = 30.61$, $p < .001$. As expected, the training positively affected teachers' ability to detect biases, but only after three sessions, Pre-Intermediate × Condition: $t(149) = 1.45$, $p = .152$; Pre-Post × Condition: $t(149) = 4.44$, $p < .001$. R^2_2 indicated that the predictors in this final model explained 25% of the variability in teachers' ability to detect biases in student products, which is a large effect.

3.2.2. Bias explanation

15% of the variance in teachers' ability to explain reasoning biases was due to stable differences between teachers (ICC). Both adding Occasion and the Occasion × Condition interaction improved the model fit significantly, M2: $\chi^2(2) = 21.65$, $p < .001$; M3: $\chi^2(3) = 17.93$, $p < .001$, showing that –as expected– the training positively affected teachers' ability to explain biases correctly, but (again) only after three sessions, Pre-Intermediate × Condition: $t(149) = 1.20$, $p = .232$; Pre-Post × Condition: $t(149) = 3.38$, $p < .001$. The predictors in this model explained 10% of the variability in bias explanation (a medium effect).

3.3. Attitudes towards teaching

Average perceived relevance and perceived competence ratings on the teaching attitudes questionnaire were already quite high at pretest, at 5.1 ($SD = 0.5$) and 4.1 ($SD = 0.8$) on a six-point scale. Table 5 summarizes the results of the analyses that tested whether the training positively affected teachers' relevance perception (hypothesis 3a) and competence perception (hypothesis 3b).

Table 5
Effect of critical thinking training on teachers' perceived relevance of and perceived competence in teaching critical thinking.

	M1: intercept only	M2: effect occasion	M3: effect of training
Perceived relevance			
Fixed part	Coefficient (s.e.)	Coefficient (s.e.)	Coefficient (s.e.)
Intercept	5.07 (0.07)**	5.11 (0.08)**	4.97 (0.13)**
Pre-Intermediate		−0.05 (0.08)	0.06 (0.12)
Pre-Post		−0.09 (0.08)	−0.12 (0.13)
Condition			0.24 (0.16)
Pre-Intermediate × Condition			−0.19 (0.16)
Pre-Post × Condition			0.04 (0.16)
Random part			
Occasion and error variance (σ^2_e)	0.17 (0.41)	0.17 (0.41)	0.17 (0.41)
Teacher variance (σ^2_{u0})	0.20 (0.44)	0.20 (0.44)	0.19 (0.43)
Deviance	249.0	247.8	243.7
Perceived competence			
Fixed Part	Coefficient (s.e.)	Coefficient (s.e.)	Coefficient (s.e.)
Intercept	4.02 (0.09)**	4.12 (0.10)**	4.33 (0.15)**
Pre-Intermediate		−0.27 (0.10)**	−0.09 (0.14)
Pre-Post		−0.02 (0.10)	−0.15 (0.15)
Condition			−0.35 (0.20)
Pre-Intermediate × Condition			−0.30 (0.19)
Pre-Post × Condition			0.22 (0.19)
Random part			
Occasion and error variance (σ^2_e)	0.27 (0.52)	0.25 (0.50)	0.23 (0.48)
Teacher variance (σ^2_{u0})	0.31 (0.56)	0.32 (0.57)	0.29 (0.54)
Deviance	321.7	312.2	299.9

Note. Occasion is dummy-coded in two dummies (Intermediate test, Posttest) with Pretest as reference category. Condition coded 0 = control, 1 = training. *p < .05. **p < .001.

3.3.1. Perceived relevance

53% of the variance in teachers' perceived relevance of teaching CT was due to stable differences between teachers. Unexpectedly, neither adding Occasion (M2) nor the Occasion × Condition interaction (M3) further improved the model, M2: $\chi^2(2) = 1.25$, $p = .535$; M3: $\chi^2(5) = 5.33$, $p = .377$, indicating that the training did not affect

teachers' perceived relevance of teaching CT.

3.3.2. Perceived competence

Fifty-three percent of the variance in teachers' perceived competence in teaching CT was due to stable individual differences (M1). Adding Occasion (M2) improved the model fit significantly, $\chi^2(2) = 9.48$, $p = .009$. Although M3 showed an improved fit, $\chi^2(3) = 12.33$, $p = .006$, neither of the added Occasion × Condition interactions in this model were individually statistically significant. The interactions pointed towards a temporary decrease in teachers' perceived competence from pretest to intermediate test and a slight increase from pretest tot posttest, Pre-Intermediate × Condition: $t(150) = -1.62$, $p = .107$; Pre-Post × Condition: $t(150) = 1.13$, $p = .261$. The predictors explained 10% of the variability in teachers' perceived competence (a medium effect).

4. Discussion

The aim of this study was to gain insight into how higher education teachers' skills and attitudes related to teaching CT could be fostered, which is a first step towards better preparing and supporting teachers for their crucial role in fostering students' CT-skill acquisition. We focused on an essential CT-skill: the ability to avoid bias in reasoning and decision-making, as measured with heuristics-and-biases tasks. The training focused on improving teachers' own performance on those tasks, their ability to detect and explain biases in student products, and their attitudes towards teaching CT.

4.1. Heuristics-and-biases tasks

Because there were no previous experimental studies available on the effects of CT-training for teachers, we first tested whether we would replicate a main finding in student populations, that explicit bias instruction combined with task practice improves performance on learning (i.e., instructed) tasks but not on transfer (i.e., not-instructed but related) tasks (e.g., Heijltjes et al., 2014). Indeed, we replicated this finding with teachers and found a large effect on teachers' performance on learning tasks (hypothesis 1). However, in contrast to prior single-session studies with students, our learning effect was only significant after three training sessions when considering the entire sample (note that the training condition did improve significantly at intermediate test when excluding absent teachers from the analyses). The explorative analyses indicated that the improved performance on learning tasks after three training sessions was attained with a similar amount of mental effort investment as prior to training. This is consistent with previous research with students (Heijltjes et al., 2014, 2015; Van Peppen et al., 2018) and points to an acquired efficiency when dealing with the tasks (Hoffman & Schraw, 2010; Paas & Van Merriënboer, 1993; Van Gog & Paas, 2008).

The lack of an effect on transfer task performance is in line with prior research with student populations (e.g., Heijltjes et al., 2014). Although we speculated that the generative learning activities (e.g., designing a CT-task) offered in the second and third session could positively affect transfer, we found no evidence that this was the case. We should note that the low reliability of the transfer tasks dramatically reduced the power to detect intervention effects (Kanyongo, Brook, Kyei-Blankson, & Gocmen, 2007). In our study, this low reliability can at least partly be explained by, respectively, low variance due to floor and ceiling effects at all tests for the Wason selection tasks and the covariation detection tasks that we used to measure transfer (see Table C1, Appendix C). Nevertheless, the measurement of CT is still a major challenge (Ku, 2009; Liu, Frankel, & Roohr, 2014). Not only did other studies using heuristics-and-biases tasks report low reliabilities (Aczel, Bago, Szollosi, Foldes, & Lukacs, 2015; Bruine de Bruin, Parker, & Fischhoff, 2007; West et al., 2008), but multiple studies also reported low reliabilities and/or poor construct validity on widely used standardized CT tests (e.g., California Critical Thinking Skills Test and Watson–Glaser Critical Thinking Appraisal; Bernard et al., 2008; Bernard et al., 2008; Bondy, Koenigseder, Ishee, & Williams, 2001; Jacobs, 1999; Leppa, 1997; Loo & Thorpe, 1999). We do have two other indications that at least some transfer took place. First, the explorative analyses showed an increase in invested mental effort in the transfer tasks on the intermediate test and the posttest compared to the pretest, which may indicate that the trained teachers did detect a conflict between their heuristic response and the normative response and invested more effort in attempting to resolve this conflict, but (mostly) without success. Following this, Stanovich (2018) would explain the incorrect performance as an override failure due to insufficiently automatized mindware (i.e., requisite knowledge for correct reasoning), that is, subjects possess enough mindware to detect a conflict but not enough mindware to trigger the normatively correct response. More importantly, we found some evidence of transfer to other contexts, as the training improved teachers' ability to detect and explain similar biases in student products (see also Table C2, Appendix C).

4.2. Detecting and explaining reasoning biases

Because improving teachers' own CT-skills is not sufficient for teaching it, we also examined whether the training positively affected other variables, specifically related to teaching CT. In line with our hypotheses (2a and 2b), we found that the trained teachers detected significantly more biases in the student product (vignettes) and were better able to explain the biases correctly than teachers in the control condition. Yet again, these effects were only visible after three training sessions. The results of the intermediate test are hard to interpret, because performance in both the training and the control condition dropped substantially from pretest to intermediate test. This indicates a potential limitation of our study: despite their structurally equivalent features, this may indicate that the vignettes varied in difficulty level. Notably, despite the fact that the ability to explain biases significantly improved after three sessions, performance was still relatively low: the descriptive statistics revealed that while the trained teachers detected on average all biases hidden in the student products, they could explain only half of them correctly. From a teaching perspective this is a crucial shortcoming as providing adequate feedback on students' reasoning, is essential for students' CT-skills acquisition (Abrami et al., 2015).

4.3. Attitudes towards teaching

In contrast to our hypothesis (3a) and to the findings of a training study in the domain of science teaching (Van Aalderen-Smeets & Walma van der Molen, 2015), we found no effect of the training on teachers' perceived relevance of teaching CT. The most likely explanation is a ceiling effect, as all participating teachers already perceived the teaching of CT as highly relevant prior to the training. Because participation was voluntary, the participating teachers may have been positively biased regarding the relevance of teaching CT compared to the general teacher population.

Also in contrast to our hypothesis (3b) and Van Aalderen-Smeets and Walma van der Molen (2015), we found no positive effect of the training on perceived competence. Interestingly, perceived competence of teachers in the training session even seemed to drop temporarily from pretest to intermediate test. Possibly, the first training session made teachers more aware of their knowledge gaps and thereby temporarily decreased their perceived competence in teaching CT. In addition, it may also be possible that, in order to improve teachers' attitudes, the focus on attitudes should be stronger than was the case in our training. We intended to foster attitudes with the second and third training session through discussing the relevance of teaching CT and focusing on ways to integrate CT during teaching, but this may not have been sufficient and may require more time. In comparison, the training in the study by Van Aalderen-Smeets and Walma van der Molen (2015) was fully focused on improving teachers' attitudes towards teaching science and more intensive (i.e., six meetings of three hours, spread over six months).

4.4. Limitations and future research

Two potential limitations of this study arise from the fact that this study was conducted in the context of an authentic professional development course for teachers. First, we were not able to randomly assign teachers to the training and the control condition and, therefore, cannot draw strong causal conclusions on the effects of the training, although our pretest data did not give strong reasons to question the comparability of both conditions (i.e., no significant condition differences on the outcome measures and on most of the background variables). The training condition did consist of significantly more females and teachers from the economical teaching domain than the control condition (all teachers in training Group A were females from the same department), and we cannot fully rule out that this may have affected the outcome measures. However, neither gender nor teaching domain correlated with improvement on the outcome measures, indicating that there was no confound.

Second, because participation in the training and the study was voluntary, our sample might not be representative for the overall population of higher education teachers. As mentioned earlier, this likely affected our findings regarding effects of training on teaching attitudes. Future research should attempt to address perceived relevance of and perceived competence in teaching CT in more representative teacher samples, because according to expectancy-value theory, these variables are important predictors of whether teachers will actually teach CT-skills in their classroom. Regarding the measures of the effects of training on teachers' CT-skills, this was presumably less problematic as we see no reason (given similar findings from prior research with students) to expect a different data pattern in a more varied sample. However, as our sample size was rather small and because this is the first experiment focusing on teachers' CT-skills, future research should point out whether our findings are replicated. Furthermore, what could possibly be a result of self-selection bias, is that the results of this study (and a previous survey study; Authors, submitted) suggest that higher education teachers perform relatively well (compared to students) on the heuristics-and-biases tasks even prior to any intervention.

Another potential limitation is that we cannot draw conclusions on what specific aspects of the training were most effective for which outcomes. Because there were no experimental studies available on training CT-skills of (higher education) teachers, our first step was to explore whether CT-training would have the potential to affect teachers' CT-skills and attitudes towards teaching CT. Future research on teacher training should address this question and our findings suggest that such research should not only focus on what is most effective for establishing improvement in teachers' performance on CT-tasks, but particularly on how to improve their performance on tasks that more closely related to the teaching practice, like the detection and explanation measures we introduced in this study. These measures can be seen as direct indicators of an essential condition necessary for teaching CT, namely providing adequate feedback so that students know how to improve their reasoning. As explaining why particular arguments were invalid was especially hard for teachers, we suggest that (research on) teacher CT-training should particularly focus on how to further improve this ability.

Finally, because we found clear effects on our outcome measures only after three training sessions, and even found a temporary drop in teachers' competence perception after one session, it seems important to investigate issues concerning the amount of training time needed and spacing and repetition of practice opportunities. Spacing and repetition might also be effective means to foster deep learning and, thereby, increase transfer to other tasks and contexts, and increase the ability to explain students' reasoning biases, which are both very important in teaching CT. Especially given the constraints of time available for professional development in practice, insight into efficient training programs that nevertheless provide teachers with enough time to acquire the complex skill of CT and to gain confidence in their ability to teach it are needed.

In a future study, it would be desirable to replicate the findings using a full experimental design with a larger and more representative teacher sample. In addition, it would also be worthwhile to test specific interventions that focus on impacting one of the outcome measures addressed in our study (e.g., bias explanation or teaching attitudes).

4.5. Conclusion

To the best of our knowledge this was the first study that explored the potential effects of CT-training on teachers' CT-skills and attitudes. Where previous research largely focused on how to improve students' CT-skills, we focused on teachers' CT-skills. In line with previous studies with students, our study provides evidence for the trainability of teachers' CT-skills, but it also shows that the skills and attitudes needed for teaching CT do not improve automatically. Our findings highlight the importance of supporting teachers in their challenging but crucial role of fostering students' critical thinking skills and ask for further research into the best ways to promote teachers' ability to transfer trained skills to other CT-tasks and their ability to explain students' reasoning and to foster their attitudes towards teaching CT.

Acknowledgement

This research was funded by the Netherlands Organization for Scientific Research (project number 409-15-203). The authors would like to thank the participating teachers for their time and effort and Steven Raaijmakers for his assistance with data analysis.

Appendix A. Example items Heuristics-and-biases tasks

Below, we translated an example item of each task category administered at the pretest, intermediate test, and posttest. Examples of the other items can be found in the references in our method section or through contacting the first author.

A.1. Learning tasks

Belief bias in logic reasoning (cf. Evans, 2002)

Premise 1: No lawyers are straightforward
Premise 2: Some crooks are straightforward
Conclusion: Some lawyers are no crooks

Given that both premises are true,

a. the conclusion follows logically from the premises
b. the conclusion does not follow logically from the premises

Correct answer: b

The important thing to notice here is that one does not confuse the believability of the conclusion with the logical validity of the conclusion. For more information see Evans (2002, p. 983).

Base-rate neglect in probability estimation (cf. Stanovich et al., 2016*)*

Imagine you started a new web shop and you need to gain brand awareness among your target group. You are considering paid advertising through social media. What information would you want to have in order to estimate the probability that your web shop will gain brand awareness among your target group given that you use paid advertising through social media? Below are four pieces of information that may or may not be relevant for determining the probability. Please indicate all of the pieces of information and only those pieces of information that are necessary to determine the probability.

a. % web shops that used paid advertisement through social media, of all web shops that gained brand awareness among their target group
b. % of web shops that gained brand awareness among their target group
c. % of web shops that did not gain brand awareness among their target group
d. % web shops that used paid advertisement through social media, of all web shops that did not gain brand awareness among their target group

Correct answer: a + b + d or a + c + d (we counted a + b + c + d also as correct). The important thing to notice here is that one should not overrate the first piece of information (a) as this percentage is uninformative for the probability estimation as long as you do not know how many *unsuccessful* web shops used paid advertisement as well (d). For more information, see Stanovich & West (Stanovich & West, 2000, p. 654) and Stanovich et al. (2016, p. 337)

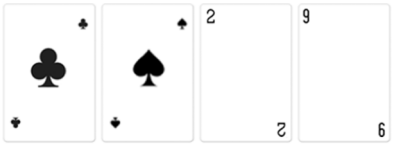
A.2. Transfer tasks

Confirmation bias in logic reasoning (cf. Evans, 2002*)*

Each of the four cards below have an image on one side and a number on the other side. The following rule applies to the cards:

'If there is a clover on one side, then there is a 9 on the other side'.

Choose those cards and only those cards that need to be turned over in order to decide whether the rule is true or false.



Correct answer: b and c.

The important thing to notice here is that one should not only confirm the logical rule (card b *must* have a 9 on the other side) but also look for falsification of the rule (card c *cannot* have a clover on the other side). For more information, see Evans (2002, p. 984) and Gigerenzer and Hug (1992).

Covariation detection problem in probability estimation (cf. Heijltjes et al., 2014*)*

You are an entrepreneur and your company is on the brink of bankruptcy. Your neighbor tells you about Corporate Fixer: a company that specializes in solving business problems. 'They do fan-tas-tic work', he says, 'the company of a good friend of mine became extremely successful after their help!' You visit their website and find out that the services of Corporate Fixer are quite pricey. You are prepared to pay the price, provided that you have a better chance of solving your business problems with their help than without any help. On an independent comparison website, you see that (a) 188 companies received help from corporate fixer and solved their business problems, (b) 95 companies did not receive help and solved their problems, (c) 90 companies received help without solving their business problems, and (d) 25 companies did not receive help and did not solve their problems:

	Help from Corporate fixer	No help from Corporate fixer
Business problems solved	188	95
Business problem unsolved	90	25

Based on this information, would you commission from Corporate fixer or not?

a. yes
b. no

Correct answer: b

The important thing to notice here is that one should to evaluate the information given in a 2 × 2 contingency table equally and surpress the tendency to focus on the large number in cell A. For more information, see Heijltjes et al. (2014, p. 40) and Wasserman et al. (1990).

Appendix B. Details questionnaire attitudes towards teaching critical thinking

We intended to measure attitudes towards teaching critical thinking with a 16-item questionnaire that addressed perceived relevance with 4 items and perceived competence with three underlying subscales: perceived competence regarding own critical thinking skills (4 items), the ability to recognize students' reasoning biases (4 items), and instructing critical thinking skills to students (4 items). However, because this factor structure did not fit our data well, we explored alternative factor structures in the pretest data. We found that a two-factor model with a total of 6 items fitted the data best. Important revisions were that we excluded the competence items addressing teachers' own critical thinking skills (as this did not

concern teaching) and the items that lacked a reference to a specific activity related to teaching critical thinking (e.g., "I am able to integrate critical thinking in the content I am teaching"). Two confirmative factor analyses on the intermediate test and posttest data confirmed this factor structure. Therefore, we reduced the 16-item version to a 6-item version as mentioned in the manuscript:

Perceived relevance (translated from Dutch):

1. Critical thinking during educational activities encourages students to become independent thinkers.
2. Learning outcomes will improve from critical thinking during educational activities.
3. Critical thinking allows students to better understand the course content.

Perceived competence (translated from Dutch):

4. I notice it immediately when students commit a thinking fallacy during my lessons.
5. I can explain clearly to my students how they are drawing incorrect conclusions from the available information.
6. I can explain various fallacies to my students in such a way that they understand it.

Appendix C. Supplementary results

Table C1
Means (M) and Standard Deviations (SD) of performance on heuristics-and-biases tasks per task category at pretest, intermediate test, and posttest per condition.

	Control		Training	
	M	*SD*	*M*	*SD*
Syllogisms with belief bias (range: 0–7)				
Pretest	4.5	1.4	4.5	1.0
Intermediate test	4.5	1.5	5.3	1.2
Posttest	4.2	1.2	4.9	1.2
Base-rate tasks (range: 0–3)				
Pretest	1.3	1.0	1.8	0.8
Intermediate test	1.5	0.7	2.2	0.6
Posttest	1.4	0.8	2.3	0.8
Wason selection tasks (range: 0–3)				
Pretest	0.3	0.6	0.3	0.6
Intermediate test	0.6	0.8	0.7	1.0
Posttest	0.5	0.8	0.6	1.0
Covariation detection tasks (range: 0–2)				
Pretest	1.8	0.4	1.5	0.7
Intermediate test	1.8	0.5	1.6	0.6
Posttest	1.8	0.4	1.3	0.8

Table C2
Percentage of teachers who detected and correctly explained the hidden biases in the vignettes (for each bias separately) at pretest, intermediate test, and posttest per condition.

	Bias detection		Bias explanation	
	Control	Training	Control	Training
Belief bias (denial of the antecedent)				
Pretest	68.2%	59.4%	36.4%	25.0%
Intermediate test	50.0%	50.0%	22.7%	25.0%
Posttest	52.6%	80.0%	5.3%	46.7%
Belief bias (affirmation of the consequent)				
Pretest	63.6%	84.4%	45.5%	50.0%
Intermediate test	36.4%	59.4%	4.5%	12.5%
Posttest	73.7%	93.3%	31.6%	50.0%
Confirmation bias (Wason selection task)				
Pretest	36.4%	34.4%	13.6%	6.3%
Intermediate test	13.6%	40.6%	0.0%	3.1%
Posttest	47.4%	96.7%	5.3%	16.7%
Base-rate neglect bias				
Pretest	31.8%	28.1%	9.1%	9.4%
Intermediate test	50.0%	53.1%	4.5%	12.5%
Posttest	36.8%	93.3%	0.0%	36.7%
Covariation detection problem				
Pretest	54.5%	75.0%	22.7%	18.8%
Intermediate test	9.1%	34.4%	0.0%	15.6%
Posttest	73.7%	93.3%	15.8%	33.3%

Note. Percentages represent the number of teachers that detected and explained the bias correctly.

Table C3
Effect of critical thinking training on teachers' invested mental effort in learning and transfer (heuristics-and-biases) tasks.

	M1: intercept only	M2: effect occasion	M3: effect of training
Mental effort learning tasks			
Fixed part	Coefficient (s.e.)	Coefficient (s.e.)	Coefficient (s.e.)
Intercept	4.28 (0.13)**	4.09 (0.15)**	3.91 (0.23)**
Pre-Intermediate		0.48 (0.12)**	0.28 (0.19)
Pre-Post		0.06 (0.12)	0.02 (0.20)
Condition			0.30 (0.30)
Pre-Intermediate × Condition			0.34 (0.24)
Pre-Post × Condition			0.06 (0.25)
Random part			
Occasion and error variance (σ^2_e)	0.46 (0.68)	0.39 (0.63)	0.38 (0.62)
Teacher variance (σ^2_{u0})	0.79 (0.89)	0.81 (0.90)	0.77 (0.88)
Deviance	420.3	403.2	398.2
Mental effort transfer tasks			
Fixed Part	Coefficient (s.e.)	Coefficient (s.e.)	Coefficient (s.e.)
Intercept	5.17 (0.18)**	5.12 (0.20)**	5.27 (0.31)**
Pre-Intermediate		0.20 (0.16)	−0.23 (0.23)
Pre-Post		−0.06 (0.17)	−0.99 (0.24)
Condition			−0.25 (0.40)
Pre-Intermediate × Condition			0.71 (0.30)*
Pre-Post × Condition			1.51 (0.31)**
Random part			
Occasion and error variance (σ^2_e)	0.73 (0.86)	0.72 (0.85)	0.58 (0.76)
Teacher variance (σ^2_{u0})	1.50 (1.22)	1.50 (1.23)	1.50 (1.23)
Deviance	501.4	498.7	475.9

Note. Occasion is dummy-coded in two dummies (Intermediate test, Posttest) with Pretest as reference category. Condition coded 0 = control, 1 = training. $^{*}p < .05$. $^{**}p < .001$.

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*, 275–314. https://doi.org/10.3102/0034654314551063.

Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R., & Zhang, D. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of Educational Research, 78*, 1102–1134. https://doi.org/10.3102/0034654308326084.

Aczel, B., Bago, B., Szollosi, A., Foldes, A., & Lukacs, B. (2015). Measuring individual differences in decision biases: Methodological considerations. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.01770.

Arum, R., Cho, E., Kim, J., & Roksa, J. (2012). *Documenting uncertain times: Post-graduate transitions of the Academically Adrift cohort.* New York: Social Science Research Council.

Arum, R., & Roksa, J. (2011). Limited learning on college campuses. *Society, 48*, 203–207. https://doi.org/10.1007/s12115-011-9417-8.

Baron, J. (2008). *Thinking and deciding* (4th ed.). Cambridge: Cambridge University Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48. https://doi.org/10.18637/jss.v067.i01.

Bernard, R. M., Zhang, D., Abrami, P. C., Sicoly, F., Borokhovski, E., & Surkes, M. A. (2008). Exploring the structure of the watson-glaser critical thinking appraisal: One scale or many subscales? *Thinking Skills and Creativity, 3*, 15–22. https://doi.org/10.1016/j.tsc.2007.11.001.

Bondy, K. N., Koenigseder, L. A., Ishee, J. H., & Williams, B. G. (2001). Psychometric properties of the California critical thinking tests. *Journal of Nursing Measurement, 9*, 309–328. https://doi.org/10.1891/1061-3749.9.3.309.

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology, 92*, 938–956. https://doi.org/10.1037/0022-3514.92.5.938.

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of Need for Cognition. *Journal of Personality Assessment, 48*, 306–307. https://doi.org/10.1207/s15327752jpa4803_13.

Choy, S. C., & Cheah, P. K. (2009). Teacher perceptions of critical thinking among students and its influence on higher education. *International Journal of Teaching and Learning in Higher Education, 20*, 198–206.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed., reprint). New York, NY: Psychology Press.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*, 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153.

Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In J. Baron, & R. J. Sternberg (Eds.). *Teaching thinking skills: Theory and practice* (pp. 9–26). New York, NY, US: Freeman.

Ennis, R. H., Millman, J., & Tomko, T. N. (1985). *Cornell critical thinking tests* (3rd ed.). Pacific Grove, CA, US: Midwest Publications.

Evans, J. S. B. T. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin, 128*, 978–996. https://doi.org/10.1037//0033-2909.128.6.978.

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629.

Facione, P. A. (1990). *The California critical thinking skills test.* Millbrae, CA, US: California Academic Press.

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review, 28*(4), 717–741. https://doi.org/10.1007/s10648-015-9348-9.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology, 18*, 253–292. https://doi.org/10.1016/0010-0285(86)90001-0.

Frey, D., Johnson, E. D., & De Neys, W. (2017). Individual differences in conflict detection during reasoning. *Quarterly Journal of Experimental Psychology, 1–52*. https://doi.org/10.1080/17470218.2017.1313283.

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition, 43*, 127–171. https://doi.org/10.1016/0010-0277(92)90060-U.

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains. *American Psychologist, 53*(4), 449–455. https://doi.org/10.1037//0003-066X.53.4.449.

Heijltjes, A., Van Gog, T., Leppink, J., & Paas, F. (2014). Improving critical thinking: Effects of dispositions and instructions on economics students' reasoning skills. *Learning and Instruction, 29*, 31–42. https://doi.org/10.1016/j.learninstruc.2013.07.003.

Heijltjes, A., Van Gog, T., Leppink, J., & Paas, F. (2015). Unraveling the effects of critical thinking instructions, practice, and self-explanation on students' reasoning performance. *Instructional Science, 43*, 487–506. https://doi.org/10.1007/s11251-015-9347-8.

Heijltjes, A., Van Gog, T., & Paas, F. (2014). Improving students' critical thinking: Empirical support for explicit instructions combined with practice: Critical thinking instructions. *Applied Cognitive Psychology, 28*, 518–530. https://doi.org/10.1002/acp.3025.

Hoffman, B., & Schraw, G. (2010). Conceptions of efficiency: Applications in learning and problem solving. *Educational Psychologist, 45*, 1–14. https://doi.org/10.1080/00461520903213618.

Hox, J. J. (2010). *Multilevel analysis: techniques and applications* (2nd ed). New York: Routledge, Taylor & Francis.

Jacobs, S. S. (1999). The equivalence of forms A and B of the California critical thinking skills test. *Measurement and Evaluation in Counseling and Development, 31*, 211–222.

Jones, A. (2007). Multiplicities or manna from heaven? Critical thinking and the disciplinary context. *Australian Journal of Education, 51*, 84–103. https://doi.org/10.1177/000494410705100107.

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In K. J. Holyoak, & R. G. Morrison (Eds.). *The Cambridge handbook of thinking and reasoning* (pp. 267–293). Los Angeles, California: Cambridge University Press.

Kanyongo, G. Y., Brook, G. P., Kyei-Blankson, L., & Gocmen, G. (2007). Reliability and statistical power: How measurement fallibility affects power and required sample

sizes for several parametric and nonparametric statistics. *Journal of Modern Applied Statistical Methods, 6*, 81–90. https://doi.org/10.22237/jmasm/1177992480.
Kenyon, T., & Beaulac (2014). Critical thinking education and debiasing. *Informal Logic, 34*, 341. https://doi.org/10.22329/il.v34i4.4203.
Klassen, R. M., & Tze, V. M. C. (2014). Teachers' self-efficacy, personality, and teaching effectiveness: A meta-analysis. *Educational Research Review, 59–76*. https://doi.org/10.1016/j.edurev.2014.06.001.
Ku, K. Y. L. (2009). Assessing students' critical thinking performance: Urging for measurements using multi-response format. *Thinking Skills and Creativity, 4*, 70–76. https://doi.org/10.1016/j.tsc.2009.02.001.
Kuhn, D. (2005). *Education for thinking.* Cambridge, MA: Harvard University Press.
LaHuis, D. M., & Ferguson, M. W. (2009). The accuracy of significance tests for slope variance components in multilevel random coefficient models. *Organizational Research Methods, 12*, 418–435. https://doi.org/10.1177/1094428107308984.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174. https://doi.org/10.2307/2529310.
Leppa, C. J. (1997). Standardized measures of critical thinking: Experience with the California critical thinking tests. *Nurse Education, 22*, 29–33.
Liu, O. L., Frankel, L., & Roohr, K. C. (2014). Assessing critical thinking in higher education: Current state and directions for next-generation assessment. *ETS Research Report Series, 2014*(1), 1–23. https://doi.org/10.1002/ets2.12009.
Loo, R., & Thorpe, K. (1999). A psychometric investigation ofscores on the Watson-glaser critical thinking appraisal new forms. *Educational and Psychological Measurement, 59*, 995–1003.
Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*, 429–434. https://doi.org/10.1037/0022-0663.84.4.429.
Paas, F. G. W. C., & Van Merriënboer, J. J. G. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 35*, 737–743. https://doi.org/10.1177/001872089303500412.
Pascarella, E. T., Blaich, C., Martin, G. L., & Hanson, J. M. (2011). How robust are the findings of Academically Adrift? *Change: The Magazine of Higher Learning, 43*, 20–24. https://doi.org/10.1080/00091383.2011.568898.
Paul, R. W., Elder, L., & Bartell, T. (1997). *California teacher preparation for instruction in critical thinking: Research findings and policy recommendations.* Sacramento: California Commission on Teacher Credentialing.
Perkins, D., Tishman, S., Ritchhart, R., Donis, K., & Andrade, A. (2000). Intelligence in the wild: A dispositional view of intellectual traits. *Educational Psychology Review, 12*, 269–293. https://doi.org/10.1023/A:1009031605464.
Pithers, R. T., & Soden, R. (2000). Critical thinking in education: A review. *Educational Research, 42*, 237–249. https://doi.org/10.1080/001318800440579.
R Development Core, & Team (2008). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.
Ritchhart, R., & Perkins, D. N. (2005). Learning to think: The challenges of teaching thinking. In K. J. Holyoak, & R. G. Morrison (Eds.). *The Cambridge handbook of thinking and reasoning.* Los Angeles, California: Cambridge University Press.
Schmidt, H. G., Mamede, S., van den Berge, K., van Gog, T., van Saase, J. L. C. M., & Rikers, R. M. J. P. (2014). Exposure to media information about a disease can cause doctors to misdiagnose similar-looking clinical cases. *Academic Medicine, 89*, 285–291. https://doi.org/10.1097/ACM.0000000000000107.
Siegel, H. (1988). *Educating reason: Rationality, critical thinking and education.* New York, N.Y. U.A.: Routledge.
Stanovich, K. E. (2018). Miserliness in human cognition: the interaction of detection, override and mindware. *Thinking & Reasoning, 1–22*. https://doi.org/10.1080/13546783.2018.1459314.
Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*, 645–665. https://doi.org/10.1017/S0140525X00003435.
Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning, 13*, 225–247. https://doi.org/10.1080/13546780600780796.
Stanovich, K. E., West, R. F., & Toplak, M. E. (2016). *The rationality quotient: Toward a test of rational thinking.* Cambridge, Massachusetts London, England: MIT Press.
Stedman, N. L. P., & Adams, B. L. (2012). Identifying faculty's knowledge of critical thinking concepts and perceptions of critical thinking instruction in higher education. *NACTA Journal, 56*, 9–14.
Sternberg, R. J. (2001). Why schools should teach for wisdom: The balance theory of wisdom in educational settings. *Educational Psychologist, 36*, 227–245. https://doi.org/10.1207/S15326985EP3604_2.
Sunstein, C. R. (2003). Terrorism and probability neglect. *Journal of Risk and Uncertainty, 26*, 121–136. https://doi.org/10.1023/A:1024111006336.
Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics (Pearson new international edition)* (6th ed.). Harlow: Pearson.
Thompson, W. C., & Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials: The proscecutor's fallacy and defense attorney's fallacy. *Law and Human Behavior, 11*, 167–187. https://doi.org/10.2307/1393631.
Toplak, M. E., West, R. F., & Stanovich, K. E. (2017). Real-world correlates of performance on heuristics and biases tasks in a community sample: heuristics and biases tasks and outcomes. *Journal of Behavioral Decision Making, 30*, 541–554. https://doi.org/10.1002/bdm.1973.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131. https://doi.org/10.1126/science.185.4157.1124.
Van Aalderen-Smeets, S. I., & Walma van der Molen, J. H. (2013). Measuring primary teachers' attitudes toward teaching science: Development of the Dimensions of Attitude toward Science (DAS) instrument. *International Journal of Science Education, 35*, 577–600. https://doi.org/10.1080/09500693.2012.755576.
Van Aalderen-Smeets, S. I., & Walma van der Molen, J. H. (2015). Improving primary teachers' attitudes toward science by attitude-focused professional development. *Journal of Research in Science Teaching, 52*, 710–734. https://doi.org/10.1002/tea.21218.
Van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist, 43*, 16–26. https://doi.org/10.1080/00461520701756248.
Van Peppen, L. M., Verkoeijen, P. P. J. L., Heijltjes, A. E. G., Janssen, E. M., Koopmans, D., & van Gog, T. (2018). Effects of self-explaining on learning and transfer of critical thinking skills. *Frontiers in Education, 3*. https://doi.org/10.3389/feduc.2018.00100.
Wasserman, E. A., Dorner, W. W., & Kao, S. F. (1990). Contributions of specific cell information to judgments of interevent contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 509–521. https://doi.org/10.1037/0278-7393.16.3.509.
Watson, G., & Glaser, E. M. (1980). *Watson-glaser critical thinking appraisal.* New York, NY, US: Psychological Corp.
Watt, H. M. G., & Richardson, P. W. (2007). Motivational factors influencing teaching as a career choice: Development and validation of the FIT-choice scale. *The Journal of Experimental Education, 75*(3), 167–202. https://doi.org/10.3200/JEXE.75.3.167-202.
West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology, 100*, 930–941. https://doi.org/10.1037/a0012842.
Zee, M., & Koomen, H. M. Y. (2016). Teacher self-efficacy and its effects on classroom processes, student academic adjustment, and teacher well-being: A synthesis of 40 years of research. *Review of Educational Research, 86*, 981–1015. https://doi.org/10.3102/0034654315626801.