Contemporary Educational Psychology 60 (2020) 101844

Contents lists available at ScienceDirect

Contemporary Educational Psychology

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

'Consider the Opposite' – Effects of elaborative feedback and correct answer feedback on reducing confirmation bias – A pre-registered study

Suzan van Brussel[a,*], Miranda Timmermans[a], Peter Verkoeijen[b,c], Fred Paas[c,d]

[a] Teacher Education Institution, Avans University of Applied Sciences, Breda, the Netherlands
[b] Learning and Innovation Centre, Avans University of Applied Sciences, Breda, the Netherlands
[c] Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, Rotterdam, the Netherlands
[d] School of Education/Early Start, University of Wollongong, Wollongong, Australia

ARTICLE INFO

Keywords:
Elaborative feedback
Correct answer feedback
Confirmation bias
Instructional video
Learning
Transfer

ABSTRACT

Unbiased reasoning is considered an essential critical thinking skill that students need to possess to face the future challenges in their work and life. Confirmation bias, which is the tendency to selectively attend to information that is consistent with held beliefs, presents a significant thread to unbiased reasoning. An effective strategy to reduce confirmation bias is the 'consider-the-opposite'-strategy (COS). The central question of this pre-registered study was whether providing elaborative, worked example feedback after COS practice would lead to a better performance on previously practised and transfer tasks than correct-answer feedback. Participants were 132 university students who took a confirmation bias pre-test, watched an instructional video on COS afterwards and next received either worked example feedback or correct answer feedback on practice tasks, practised only, watched the instruction only or received no treatment. Finally, all participants took a learning test assessing their skill to avoid confirmation bias, and a transfer test assessing whether they could apply this acquired skill to problems containing other biases. Results revealed no differences on the learning test between both feedback conditions, but students who received feedback scored significantly higher on the confirmation bias problems than students who did not receive feedback. We carried out our pre-registered analysis plan, but due to the low reliability of particularly the pre-test, we carried out an additional exploratory analysis on subsets of post-test items and a subset of transfer test items. Results on learning revealed the same pattern as the planned analyses. However, we found no differences between any of the conditions on transfer.

1. Introduction

Fostering students' critical thinking (CT), supporting the application of their CT skills, and achieving transfer of these skills to situations outside of their learning context has become more and more a central objective of education because it is considered to be crucial for decision making, leadership, judgment, and professional success (e.g., Barnett & Ceci, 2002; Beaulac & Kenyon, 2014; Catapano, Tormala, & Rucker, 2019; Halpern, 1998; Hattie, Biggs, & Purdie, 1996; Johnson & Hamby, 2015; Lobato, 2006; Quatrociocchi, Scala, & Sunstein, 2016; Yang & Chou, 2008). However, there is no 'one size fits all' regarding the definition of CT. Many definitions have been proposed but the broadest definition has probably been compiled by the Delphi Panel (Facione, 1990). The panel, which consisted of 46 philosophical experts, characterized CT as purposeful, self-regulatory judgment, which results in six CT skills, namely interpretation, analysis, evaluation, inference, explanation and self-regulation, and 19 dispositions (Facione, 1990). In this study we focus on the evaluation skill, specifically directed towards the critical assessment of information during decision making and judgment. To critically assess information, one needs to apply the disposition of open mindedness regarding alternatives and opinions (Abrami et al., 2015, p. 306 for an overview of CT skills and dispositions).

According to Evans' dual processing theory (2003), to critically assess information people should use reflective Type 2 thinking processes instead of automatic Type 1 thinking processes. The dual processing theory indicates that thinking errors occur because of rapid Type 1 reasoning instead of the desired, reflective Type 2 reasoning. The latter reasoning processes are slower and draw more heavily on working memory capacity than Type 1 reasoning (Kahneman, 2011). Excessive reliance on Type 1 thinking processes prevents reflection and may lead to cognitive biases. Cognitive biases are the result of applying

* Corresponding author at: Avans University of Applied Sciences, PO Box 90116, 4800 RA Breda, the Netherlands.
E-mail address: sjm.vanbrussel@avans.nl (S. van Brussel).

https://doi.org/10.1016/j.cedpsych.2020.101844

Available online 05 February 2020

mental shortcuts (i.e., heuristics) during judgment and decision making, and they appear when individuals draw inferences or adopt beliefs where the evidence for doing so in a rationally sound manner is either insufficient or absent (Haselton, Nettle, & Andrews, 2015). An example of a cognitive bias is the confirmation bias. The confirmation bias is the tendency of people to selectively attend to information that is consistent with held beliefs and to fail to appropriately consider alternatives to his/her held beliefs or judgments (e.g., Jonas, Schulz-Hardt, Frey, & Thelen, 2001; Nickerson, 1998; Quatrociocchi et al., 2016; Reeves, 2014; Schwind, Buder, Cress, & Hesse, 2012).

The confirmation bias can be observed amongst others in rule testing situations. A rule follows a logical structure: if we assume that A is true then B should follow. According to general principles of rule testing two conditions should be met. First, if A is instantiated then the consequence B should follow. This is a confirmatory test. The second requirement, is checking whether A does not occur when the consequence B is not present. This is a dis-confirmatory test. Consistent with the confirmation bias, it has been shown that in rule testing situations, people perform the confirmatory test but not the dis-confirmatory one. Cognitive psychologist Peter Wason (1960) was one of the first to coin the term confirmation bias in the context of rule testing. In Wason's four-card selection task, people are presented with four cards with a number on one side and a colour on the other. For two cards, the numbers are displayed (e.g., 3 and 6) and for two the colour (red and blue). Subsequently, they have to test a given rule, for example if a card has an even number on one side then its opposite side is red, by choosing which of four cards they need to turn over. Typically, people perform a confirmatory test (i.e., turning the card with the even number), but not the dis-confirmatory test (i.e., turning the card that is not red). So, people showed the tendency to confirm rather than to falsify the rule, which is an expression of the confirmation bias. In the literature, Wason's four-card selection tasks are often used to measure confirmation bias.

Next to rule testing, the confirmation bias occurs also in situations in which people need to select information to come to a decision or make a judgment. Quatrociocchi et al. (2016), showed that e.g., Facebook users have a tendency to search for information that supports their preference on a subject, and to reject disproving information that undermines their preference. In his overview article, Nickerson (1998, pp. 192–194) appoints examples of the confirmation bias in real-life situations, e.g., when diagnosing in medicine, reasoning in science and during judicial reasoning. An example of the latter is when a judge needs to come to a conclusion, he/she needs an open mind regarding the presented evidence. Therefore, judges need to disconnect the process of gathering information from that of drawing conclusions from the evidence, and they need to refrain from forming a personal belief about what happened during the crime and interpret the evidence as such. The confirmation bias appears when new evidence is interpreted based on that personal belief, because judges, as well as people in general, are more likely to remember information that is consistent with their own beliefs than inconsistent information (Nickerson, 1998).

To refrain from biased reasoning, decision making, and judgment caused by the confirmation bias, it is important that people learn the CT skills that support reducing this bias. It is generally assumed that learning to think critically does not develop and transfer automatically as a by-product of learning, instead it requires explicit instruction with practice (Abrami et al., 2008, 2015; Bangert-Drowns & Bankert, 1990; Heijltjes, 2014; Marin & Halpern, 2011; Markovits & Brunet, 2012; Mehta & Al-Mahrooqi, 2015; Niu, Behar-Horenstein, & Garvan, 2013). With respect to reducing confirmation bias, an instructional strategy that has been shown to be effective is encouraging learners to generate counter explanations. This stimulates learners to consider alternative explanations, which leads to a more balanced and objective evaluation of the evidence that is needed for decision making or judgment (Hirt & Markman, 1995). An example of such a strategy is 'consider-the-opposite' (COS; see e.g., Adame, 2016; Hirt & Markman, 1995; Lord, Lepper, & Preston, 1984; Mussweiler, Strack, & Pfeiffer, 2000). The central question of COS is: "What are some reasons that my initial judgment might be wrong?" (Larrick, 2004, p. 323). According to Larrick (2004), and Danielson and Sinatra (2017), COS works because this strategy encourages people to direct their attention to opposite evidence that would not otherwise be considered. It is known that presenting opposing information prompts learners to reconsider their prior position. For example, from the relational reasoning literature (i.e., the ability to differentiate meaningful patterns to enhance learning) it is known that reasoning about opposition (e.g., antithetical reasoning) facilitates understanding argumentation, persuasion, and conceptual change through the use of refutational texts and graphics (Danielson & Sinatra, 2017; Grossnickle, Dumas, Alexander, & Baggetta, 2016).

In their study, Lord et al. (1984) induced COS through either explicit instructions to consider opposites or through stimulus materials that made opposite possibilities more salient. Their experiment was based on the assumption that biased assimilation of new evidence occurs when opposite possibilities are overlooked. Participants who either supported or opposed capital punishment were asked to read two summaries of bogus research articles which were pro or contra the death penalty including the applied methodology for each study. After reading, participants had to indicate to what extent their attitudes and beliefs towards capital punishment had changed and to rate how convincing each study seemed as evidence on the issue. In the control condition, participants were asked to be as objective and unbiased as possible in evaluating the studies. In the COS condition, participants were asked whether they would have made the same high or low evaluation had exactly the same study produced results on the opposing side of the issue. By this, the authors reminded the participants that a different experimental design might have brought different supporting cognitions to mind. Results showed that participants in the COS condition displayed less attitude polarization on the immediate post-test and the authors concluded that biases in social judgment can be corrected only by a change in strategy.

In the study by Lord et al. (1984), post-test results showed that participants reduced their confirmation bias by considering opposites immediately after the intervention. However, performance on the longer term is also an important educational goal. COS learning effects on the longer term were shown by Morewedge et al. (2015). In their study, participants either played an interactive, educational game or watched an instructional video with the aim to reduce confirmation bias. The instructional video consisted of defining several biases, examples and suggestions for mitigating strategies such as considering alternatives. In the educational game, each player was asked to find a missing neighbour who has to clear his name of any criminal activity with help of the player. During the game, an expert explained each bias and gave examples, and players made judgments designed to test the degree to which they demonstrate confirmation bias. Participants assessed their degree of bias during each game level and they performed practice judgments of confirmation bias on which they received immediate feedback. Both training methods reduced confirmation bias immediately and moreover, the effect maintained after eight weeks. However, the game debiased participants more effectively than the instructional video during a post-test and a follow-up test (Morewedge et al., 2015). Morewedge et al. (2015) provided immediate feedback to participants on their performance on practice tasks and their level of self-assessed confirmation bias in the game condition but not in the instructional video condition. In their study, it was not clear whether the educational game, the feedback or the combination of both led to the learning effects in Morewedge's study (2015), because the feedback was a confounding part of the instruction.

However, it is reasonable to assume that frequent and continuous feedback produces greater understanding, performance and application of the learned knowledge (Hattie & Timperley, 2007; Hattie, 2012; Vollmeyer & Rheinberg, 2005). Feedback is conceptualized as information provided by an agent (e.g., teacher, peer, book, parent, self, experiences) to modify one's willingness to invest effort, thinking or behaviour in order to improve learning (Hattie & Timperley, 2007; Shute, 2008). Based on a synthesis of 1200 meta-analyses relating to influences on student achievement, Hattie (2015) showed that almost every instructional intervention has a learning effect but feedback has a large effect on student achievement (combined $d = 0.73$). The type of feedback however, generates different effects on learning (e.g., Butler, Godbole, & Marsh, 2013; Roelle, Rahimkhani-Sagvand, & Berthold, 2017; Shute, 2008).

Regarding feedback type, one distinction that can be made is between correct answer feedback and elaborative feedback. Correct answer feedback implies that learners only receive the correct answer after they provided their own response. Alternatively, a learner might receive more elaborative feedback for example consisting of the correct answer as well as the solution steps that lead up to the answer. The first type of feedback can be considered as feedback at the task level, whereas the latter type can be considered as feedback at the task level and at the process level. A review article by Bangert-Drowns, Kulik, Kulik, & Morgan (1991) has shown that the effect of correct answer feedback on learning is small at best, whereas elaborative feedback, in the form of explaining the answer, produces large effects on learning (but see a study of Van der Kleij (2012), for contradicting results).

These findings on learning may be explained by a particular effect from the Cognitive Load Theory (CLT) literature (Paas, Renkl, & Sweller, 2003; Sweller, 1988; Sweller, Van Merrienboer, & Paas, 1998), namely the worked example effect. This effect entails that studying worked examples during instruction produces better learning with less investment of time and mental effort than solving conventional problems because worked examples support learners to construct cognitive schemas of how to solve problems, which can guide their learning during subsequent practice and problem solving (Cooper & Sweller, 1987; Paas, 1992; Sweller & Cooper, 1985; Sweller et al., 1998; Van Gog & Rummel, 2010; Ward & Sweller, 1990). Worked examples can be regarded as a form of elaborative feedback because a step-by-step explanation and demonstration of a strategy to arrive at the correct solutions is presented (Hattie, 2009; Shute, 2008; Sweller et al., 1998; Van der Kleij, Eggen, Timmers, & Veldkamp, 2012). In other words, worked examples focus on the processes that lead to the correct answer. In contrast, conventional problems are problems for which learners need to find the solutions themselves, and which are generally less effective regarding investment of time and mental effort (Cooper & Sweller, 1987).

Thus, based on the literature presented above, it is reasonable to assume that adding feedback to an instructional design will mitigate the confirmation bias, particularly when the feedback contains worked examples and when the final test contains task, which are similar to those practiced during instruction (i.e., when the final test taps onto learning). To the best of our knowledge, there are currently no confirmation bias studies that tested this assumption. However, doing so is important to unearth optimal instructional design principles that help to reduce confirmation bias and to subsequently enhance judgment and decision making. The first question in the present study is therefore whether elaborative feedback with worked examples will reduce confirmation bias more than correct answer feedback on tasks that have been encountered during previous practice. Furthermore, an important question in the literature on critical thinking instruction is which design features enhance transfer of learned skills and knowledge to related, but not practiced, new tasks. A recent study by Butler et al. (2013) suggest that the type of feedback might be important with that respect. In their first experiment, Butler and colleagues tested the hypothesis that the effect of elaborative (worked example) feedback compared to correct answer feedback depends on how learning is assessed. Participants who read a text passage with critical concepts were required to answer questions about the text and either received correct answer feedback,[1] elaborative feedback,[2] or no feedback on the initial test about the text passages. The final test contained questions that were repeated verbatim from the initial test (assessing learning) and additional new inference questions which measured transfer. When retention was assessed with repeated questions on the final test, elaborative feedback produced equivalent performance relative to correct answer feedback. The key finding by Butler et al. (2013) however, was that subjects who received elaborative feedback on the initial test outperformed the other participants when understanding was assessed by asking participants to transfer their knowledge to a new context. In the literature on critical thinking instruction, the differential effect of elaborative (worked example) feedback and correct answer feedback on reducing the confirmation bias on transfer test has not been examined. Therefore, the second question addressed in the present study is whether a difference between the two feedback types on transfer can be observed in confirmation bias instruction (through applying COS).

1.1. The present study[3]

The aim of the present study was to determine what information a feedback message should contain in order to be effective for reducing confirmation bias and transfer of this effect to other bias tasks. Based on previous empirical work (Butler et al., 2013; Van der Kleij et al., 2012) and theoretical considerations (Hattie & Timperley, 2007; Shute, 2008) it seemed reasonable to expect that providing feedback on practice tasks will increase the impact of a debiasing intervention on the reduction of confirmation bias and that providing elaborative feedback in the form of worked examples will be more effective than merely corrective feedback. The experiment was conducted to address this hypothesis.

Five conditions were tested: (1) an instructional COS video, practice tasks with elaborative feedback in the form of worked examples (WE), (2) an instructional COS video, practice tasks with correct answer feedback (CA), (3) an instructional COS video and practising without feedback (PO), (4) an instructional COS video, without practice (VO) and (5) testing only (NT). Conditions 4 and 5 were created as control conditions to determine, respectively, learning effects of instruction-only and effects of pre-testing without an intervention. It was explored whether the form of feedback has an effect on self-reported mental effort during practise in the WE, CA and PO condition. Learning and transfer outcomes of bias tasks on which COS can be applied were measured to answer the research questions. Therefore, test outcomes on isomorphic test and practice confirmation bias tasks (i.e., same structure, different cover story) and transfer tasks (i.e., tasks on other biases) were used.

1.2. Hypotheses

In our design, all participants took an isomorphic pre-test and post-test containing items that were addressed during instruction and that were practised in the WE, CA, and PO condition. The pre-post performance gain score was assumed to reflect learning. On the post-test, we also measured transfer in all five conditions. Based on the literature presented above, we hypothesized that the performance gain in the WE

[1] A statement of the correct answer.

[2] The correct answer and in addition text passages that explained the concept but did not provide the answer to the inference question (transfer).

[3] The present study was pre-registered at the Open Science Framework: hypotheses, planned data-collection, planned methods, and planned analyses can be found on *< website hidden in order to ensure anonymity >* (McNutt, 2014; Simmons et al., 2011).

condition would be larger than in the four other conditions (Hypothesis 1a). Furthermore, it was hypothesized that performance gain in the CA condition would be larger in the PO, VO and NT condition, since feedback is more effective for learning compared to no feedback at all (Hypothesis 1b). Moreover, it was hypothesized that performance in the PO condition would be larger than the VO and NT condition because instruction is a key factor in enhancing CT skills (Hypothesis 1c). Finally, it was hypothesized that performance in the VO condition would be larger than in the NT condition, partly because participants were exposed to the instructional video (Hypothesis 1d).

Regarding transfer, we hypothesized that performance in the elaborative feedback (worked examples - WE) condition would be larger than in the four other conditions (Hypothesis 2a) because elaborative feedback enables learners to better comprehend the concepts and thus facilitates the application of that knowledge to new contexts. Finally, it was hypothesized that performance scores in the correct answer feedback condition (CA) would be larger than performance in the conditions without feedback on post-test transfer performance because correct answer feedback is more effective on transfer than no feedback at all (Hypothesis 2b). For the remaining conditions (i.e., PO, VO and NT) we expected the same ordering on transfer performance as on learning (hypothesis 2c).

In addition, mental effort investment was used to exploratively assess possible differences between types of feedback (WE and CA), and no feedback (PO). Therefore, mental effort during practise was taken into account as an index of the cognitive load (Paas, 1992). Mental effort is defined as 'the total amount of controlled cognitive processing in which a subject is engaged' (Paas & Van Merriënboer, 1993, p. 738).

2. Method

2.1. Participants and design

Participants were 132 first year psychology students from a Dutch university ($n = 107$) and second year student teachers from a Dutch university of applied sciences ($n = 25$), who were randomly assigned to one of only five conditions ($M_{age} = 20.73$, $SD = 3.46$, 114 women). All subjects gave informed consent prior to participating in the present experiment. With regard to their prior education, 90% reported the required secondary school level to enter higher education or university in the Netherlands, and 10% reported other prior education, such as a completed study at a university of applied sciences. At the time of the experiment CT had not yet been taught in the curriculum. For the learning measure, the experiment used a 5 (Condition: WE, CA, PO, VO and NT) × 2 (Test Moment: pre vs post) mixed design with repeated measures on the second factor. For the transfer measure, the experiment was a single-factor (WE, CA, PO, VO and NT) between-subjects design. Fig. 1 presents a schematic overview of the five conditions.

We used G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) to determine the required sample size for a standard sensitivity of the test procedure (i.e., power) of 0.80 under a fixed α level of 0.05, a correlation between repeated measures of 0.08 based on an earlier conducted and similar experiment by the same authors, and an estimated medium effect size of $f = 0.23$ (cf. Cohen's $d = 0.90$ by Butler et al., 2013). From the G*Power calculation a total sample size of 110 was required. For the transfer tasks, which had the same requirements as the confirmation bias post-test tasks to achieve a power of 0.80, 235 participants were needed. Therefore, the aim was to test as many participants as possible within the available execution time with $n = 110$ as the lower limit.

2.2. Materials

All materials were delivered through the online Qualtrics platform and published on the Open Science Framework. The material was presented in Dutch.

2.3. Instructional video

See Fig. 2 for a QR-code that leads to the instruction on YouTube (in Dutch). In the 4.05 min instructional video a teacher explained how confirmation bias is related to CT and the importance of reducing confirmation bias and the necessary CT skills. The teacher then explained how confirmation bias emerges in the information searching process related to a hypothesis testing task from the pre-test. Next, a step-by-step explanation of COS (Lord et al., 1984) as a strategy to reduce confirmation bias during the information searching process was given. On the screen the steps were shown as text. The first step of COS that was mentioned was to ask oneself the question: "What are possible reasons that my initial judgment might be wrong?" as used in the original study by Lord et al. (1984). Next, one had to "consider the opposite" which meant that one must consider contradictive information. The last step in COS is that one has to critically assess the contradictive information and weigh both answers before making a final decision.

2.4. Test tasks

All tasks on the pre-test and post-test were designed by the first author. The tasks were based on widely accepted tasks regarding

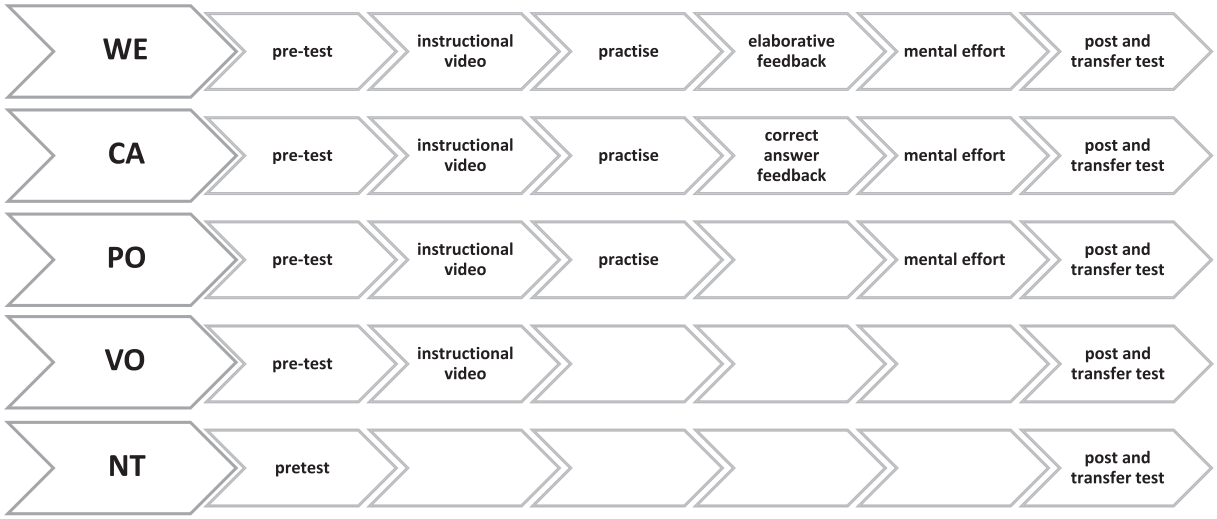

Fig. 1. Schematic overview of the five conditions.

Fig. 2. QR-code of the instructional video (in Dutch).

measuring confirmation bias and included four-card selection tasks (Wason, 1968) and hypothesis testing tasks (Jonas et al., 2001). The pre-test as well as the post-test consisted of four four-card selection tasks and two hypothesis testing tasks.

For the four-card selection tasks, participants had to choose which of the four given cards they needed to turn over in order to test a given rule. In general, participants have the tendency to select two affirmative cards, however, in order to test the rule they need to turn over an affirmative card as well as a non-affirmative card. The tendency to only select the affirmative cards can be seen as an expression of confirmation bias (Nickerson, 1998). The Appendix shows and example of this task with both types of feedback. The rule is: "If the card shows an even number on one face, than its opposite is red". The four cards that are shown are "Red", "Yellow", "3", and "8". The only possible way to falsify the rule is by turning an affirmative card ("8" in the example) and by turning a non-affirmative card ("not red" = "yellow" in the example). These cards are worth checking to test the rule. However, participants usually choose the cards that confirm the rule rather than cards that disconfirm, which is an expression of confirmation bias.

The hypothesis testing tasks based on Jonas et al. (2001), showed how the biased search for information is related to someone's position on a statement. In this task, participants were asked whether or not they agreed with a statement (e.g., "Nowadays, children have less perseverance than children from the past") and were asked to motivate their answer to stimulate them to think seriously about their statement. After this, participants read eight short summaries of articles about this subject. Four articles were strongly in favour and four were strongly opposed to the statement. Based on the given summaries, participants had to choose four articles which they wanted to read to gather more information about the statement. After selecting, the task was finished. Note that participants did not actually receive the articles. Confirmation bias is expressed by participants if they choose information that agrees with their initial point of view. If their choice is more balanced (choosing e.g. two articles in favour and two articles contra), their confirmation bias is mitigated.

2.5. Practice tasks

These tasks were similar to the test tasks. Participants practised four tasks: three four-card selection tasks and one hypotheses testing task. After each practice task, participants reported experienced mental effort using the 9-point Mental Effort Rating Scale developed by Paas (1992).

2.6. Transfer test tasks

The six transfer tasks provoked other biases than confirmation bias although a person could apply COS as well to come to the correct answer. The tasks are discussed below and they are based on sample items from the Comprehensive Assessment of Rational Thinking (CART) by Stanovich, West, and Toplak (2016). Also, the structure of the transfer tasks deviated from the instructed and practised tasks. Participants had to motivate their answer after each transfer task.

We used a task based on the research by Snyder and Swann (1978) on hypothesis testing which differed in structure from the practised tasks. In this task, participants were provided with information about extravert behaviour. They were told they had to choose 12 out of the 27 provided questions (11 on extraversion, 11 on introversion and five neutral questions) to interview a child to find out whether the child is an extravert. Confirmation bias is at risk here, because people have the tendency to choose questions that confirm the extraversion (e.g., "What would you do to enliven a boring party?") and avoid to ask questions that must prove the opposite (e.g., "Which things do you dislike about a noisy party?"). Two transfer tasks were based on conjunction fallacy tasks used by Tversky and Kahneman (1983) on the role of heuristics in decision making and the conflict between heuristics and logic. We used the "Linda"-conjunction fallacy task in which the probability of two events occurring together can never be larger than the probability of any of these events occurring alone. Three transfer tasks were included based on tasks by Stanovich et al. (2016). Two of these tasks presented problems that evoke a direct, intuitive but incorrect answer where reflection is more appropriate to solve the problem. Participants intuitively select an answer, which they assume is correct. However, in these tasks, the intuitive answer is typically incorrect and getting to the correct answer requires at least that people take a step back and think about why their initial answer might be incorrect. This is a form of considering the opposite. An example of such a task is: "The number of bacteria in a bin doubles every hour. If it takes 32 h to completely fill the container, how much time do the bacteria need to fill half the bin?". The intuitive answer is 16 but the right answer is 31. Finally, a base rate fallacy task was used in which probability is a problem (cf. the "Linda"-problem). This fallacy refers to the phenomenon whereby people ignore or undervalue probability, typically in less informative but more intuitively appealing information about a case (Kahneman & Tversky, 1973). For this, we translated a case by Stanovich et al. (2016, p. 333) in Dutch. In this case a professor must choose between two textbooks for a new course. Participants are being told how she thoughtfully comes to a choice, however at the end one less relevant and reliable piece of information for the other textbook is given. Participants are asked to advise the professor which one of the two textbooks she should definitely or probably must use.

Compared to the practice tasks, the items on the transfer test generally require a near transfer according to Barnett and Ceci's taxonomy (2002). On the transfer tasks, participants had to apply COS to solve the problem. Furthermore, because the entire procedure took place in a single session without focus on a specific knowledge domain, the training tasks and the transfer test items show considerable overlap in physical and temporal context. The modality overlap was also substantial between the practice tasks and the transfer test items as both contained short, written cases.

2.7. Procedure

The experiment was conducted at a Dutch university and a Dutch university of applied sciences in a computer room with individual cubicles. By participating, university students earned course credits and students from the university of applied sciences participated voluntarily. The experimenter provided a short, general verbal instruction about the nature of the experiment and explained some rules (e.g., mobile phones must be switched off and it is not permitted to leave the lab during the experiment). Participants were asked to read the procedure and by clicking to continue, they agreed to participate in the experiment. Prior to the experiment, participants were randomly assigned to one of the five conditions. Headphones were provided and participants started by clicking on a link that lead to the Qualtrics platform (Qualtrics. , 2017).

Participants were presented with each element of the allocated condition on a separate page on the computer screen. They could not

Table 1
Mean scores and SD on pre-test, post-test, mental effort rating and practice time (min) per condition.

	WE (n = 25) *M (SD)*	CA (n = 26) *M (SD)*	PO (n = 26) *M (SD)*	VO (n = 28) *M (SD)*	NT (n = 27) *M (SD)*
Pretest	3.56 (0.25)	3.62 (1.27)	3.37 (0.97)	3.15 (1.54)	3.56 (0.64)
Posttest	6.04 (1.46)	5.81 (1.27)	3.89 (1.09)	4.11 (1.67)	4.00 (1.36)
Mental Effort	4.54 (1.83)	4.23 (1.56)	4.27 (1.37)		
Practice Time	4.39 (1.76)	5.06 (2.06)	4.80 (1.51)		

switch to previous tasks, nor were they allowed to continue until the current task was completed. The experiment ended with a short demographic questionnaire (age, gender and prior education). After completing the test, participants were thanked and left the room. Total time was automatically logged by Qualtrics, and during the practice phase, all elements were logged separately.

As can be seen in Fig. 1, participants in the WE condition received elaborative feedback after completing each practice task. The elaborative feedback consisted of a worked example. They had to read the text which consisted of a step-by-step elaboration of the task. In the CA condition, only the correct answer was provided as feedback after each practice tasks. The Appendix shows a four-card selection practice task with examples of both feedback types. The PO group practised after the instruction without receiving feedback. Participants in the VO conditions watched the instructional video after the pre-test and started making the post-test immediately afterwards. In the NT condition, participants only took the pre-test and the post-test.

2.8. Data analysis

All participants could score a maximum of eight points on the pre-test. On the post-test participants could score a maximum of fourteen points: eight for the learning tasks and six for the transfer tasks. The data of all participants (n = 132) were included in the analyses. Results are reported with and without outliers

For each correct answer on a four-card selection task participants received one point. The answer was correct when participants chose one card which confirmed the rule and one that disconfirmed the answer. For the hypothesis testing tasks based on Jonas et al. (2001), participants received one or two points, depending on the information which they selected: participants who selected four pieces of information that were either all pro or contra the statement, they received no points. If they selected two pieces of information in favour and two pieces of information that disagreed with their statement, participants received two points because that showed a tendency for falsification. If they selected three pieces of information in favour and one not in favour, they earned one point because that meant they still showed too much confirmation bias. The pre-test with six items had a Cronbach's α = 0.16, so the proportion of systematic variance in the pre-test sum score was low. This was most probably due to the low level of relevant prior task knowledge, resulting in response guessing and consequently a high degree of non-systematic variance in the pre-test sum scores. Furthermore, the post-test with six items had a Cronbach's α = 0.58, which is still low for the purpose of the presents study, namely detecting differences between group means. We carried out our pre-registered analysis plan, but due to the low reliability of particularly the pre-test, we carried out an additional exploratory analysis on the post-test items. Because the reliability of the post-test total score was sub-optimal, we examined Cronbach's α for two subsets of post-test items, i.e., the two hypothesis testing tasks and the four-card selection tasks, for which the item responses can be expected to be correlated as the items in each subset are isomorphic. The two hypothesis testing tasks had a negative correlation with Cronbach's α = −0.17. The other subset of the four-card selection tasks had an acceptable Cronbach's α = 0.76. Hence, for the post-test, we conducted an exploratory analysis on the sum score for the latter subset.

For the six transfer tasks, one point was assigned if one provided the correct answer and the correct explanation. If a participant gave the correct answer but an incorrect explanation or an incorrect explanation but a correct explanation then 0.5 points were assigned. If a participant gave an incorrect answer as well as an incorrect explanation then no points were assigned. The transfer test had a low reliability, Cronbach's α = 0.24. Thus, the total-test scores cannot be meaningfully interpreted. Hence, and therefore, contrary to what was mentioned in the pre-registrational part of the study, explorative analyses per item and on subset of items were conducted on the items from the transfer test.

Qualtrics automatically scored the responses and total time spent on the intervention. The scores were transferred to SPSS for the statistical analyses. The mental effort scores and the time measurement during the practice phase in the WE, CA and PO conditions, were computed.

3. Results

The data of all participants (n = 132) were included in the analyses. Results are reported with and without outliers (if necessary), and for planned as well as explorative analyses. In all analyses below, a significance level of 0.05 was used. Partial eta-squared (η_p^2) is reported as a measure of effect size for the ANOVAs for which 0.01 is considered small, 0.06 medium, and 0.14 large. See Table 1 for means and *SD* of pre-test (eight points maximum), post-test without transfer scores (eight points maximum), self-reported mental effort means (nine points maximum) and time spent on practising. In the WE, CA, PO, and VO condition participants watched the 4.05 min instructional video. The means and *SD* of the explorative transfer results with a maximum of six points are presented in Table 2.

Table 2
Mean scores and SD of subsets of transfer test items.

	WE (n = 25) *M (SD)*	CA (n = 26) *M (SD)*	PO (n = 26) *M (SD)*	VO (n = 28) *M (SD)*	NT (n = 27) *M (SD)*
3 Item subset*	1.36 (1.07)	1.65 (1.06)	1.81 (0.94)	1.75 (1.08)	2.11 (1.09)
Base rate	0.56 (0.51)	0.52 (0.50)	0.50 (0.51)	0.50 (0.51)	0.74 (0.45)
Linda	0.40 (0.50)	0.19 (0.40)	0.27 (0.45)	0.46 (0.51)	0.30 (0.47)
Extra-vert	0.64 (0.49)	0.69 (0.47)	0.58 (0.50)	0.57 (0.50)	0.37 (0.49)

* Two reflection vs. intuition tasks by Stanovich et al. (2016) combined with a conjunction fallacy task (Tversky & Kahneman, 1983).

3.1. Planned analyses

3.1.1. Performance on learning

A 5 (Condition: WE, CA, PO, VO and NT) × 2 (Test Moment: pre vs post) mixed ANOVA on the learning scores with repeated measures on the second factor was performed. We found a main effect of Test Moment on learning, $F(1, 127) = 99.66$, $p < .001$, $\eta_p^2 = 0.44$. However, and most importantly, there was a Test Moment × Condition interaction effect on learning, $F(4, 127) = 10.29$, $p < .001$, $\eta_p^2 = 0.25$. As a follow up of this significant interaction, a planned Helmert contrast with NT as baseline was performed. As expected, this contrast showed that the mean learning gain in the NT condition was significantly lower with a contrast estimate of -1.09 ($SE = 0.39$), $p < .05$, than the combined mean learning from the subsequent four conditions. Furthermore, with a contrast estimate of -0.81 ($SE = 34$), $p < .05$, the mean learning gain in the VO condition as expected was significantly lower than the combined mean learning gain in the PO, CA and WE condition. Also as expected, the mean learning gain in the PO condition was significant lower than the combined mean in the two feedback conditions with a contrast estimate of -1.80 ($SE = 0.37$), $p < .05$. Contrary to what was expected, the two feedback conditions WE and CA did not differ in mean learning gain with a contrast estimate of -0.29 ($SE = 0.43$), $p > .05$.

Five paired t-tests were conducted to evaluate the differences within the groups from pre-test to post-test performance. There was a statistically significant increase from pre-test to post-test for the WE group ($M_{dif} = 2.48$, $SD = 1.46$), $t(24) = -6.96$, $p < .001$, CA group ($M_{dif} = 2.19$, $SD = 1.27$), $t(25) = -7.466$, $p < .001$, PO group ($M_{dif} = 0.54$, $SD = 1.09$), $t(25) = -2.34$, $p = .028$, and VO group ($M_{dif} = 0.93$, $SD = 1.65$), $t(27) = -3.06$, $p = .005$. There was no statistically significant increase for the NT condition ($M_{dif} = 0.44$, $SD = 1.36$), $t(26) = -1.56$, $p = .130$. Outliers were detected in the CA group (5 high extremes) and PO group (1 high extreme and 2 low extremes) but results were similar with outliers.

3.1.2. Mental effort scores

To gain more insight into the mean mental effort per group as a function of type of feedback, or no feedback, mental effort scores were collected after each of the four practised items and the feedback in WE and CA and after the four practised items in PO. Mean results per condition are shown in Table 1. The mean of the self-rated mental effort scores over all three conditions was 4.35 ($SD = 1.59$). The mental effort rating scale by Paas (1992) indicates this score as rather low mental effort. For this, a mixed ANOVA was performed. We found no significant effects of condition on mental effort, $F(6, 148) = 0.736$, $p = .621$, $\eta_p^2 = 0.029$.

3.1.3. Practice time and feedback

Practice time and time spent on reading feedback (WE and CA) and practice time (PO) was used for exploratory reasons only. We found no significant difference between the time participants spent on practising and reading the feedback in the WE, CA, and PO conditions: $F(1, 49) = 1.53$, $p > .05$, $\eta_p^2 = 0.03$.

3.2. Explorative analyses

3.2.1. Explorative analyses on learning (post-test performance)

As mentioned above, beyond our pre-registration, we conducted analyses on a subset of similar post-test tasks, namely the four-card selection post-test tasks. For this subset, we found a significant effect on post-test performance between conditions, $F(4, 127) = 17{,}32$, $p < 0.05$, $\eta_p^2 = 0.35$. A Helmert contrast showed the same pattern as the pre-registered mixed ANOVA[4]. That is, the mean post-test performance in the NT condition was significantly lower with a contrast estimate of -0.21 ($SE = 0.07$), $p < .05$, than the combined mean post-test performance from the subsequent four conditions. Furthermore, with a contrast estimate of -0.24 ($SE = 0.07$), $p < .05$, the mean post-test performance in the VO condition was significantly lower than the combined mean post-test performance in the PO, CA and WE condition. Also, the mean post-test performance in the PO condition was significant lower than the combined mean post-test performance condition in the two feedback conditions with a contrast estimate of -0.50 ($SE = 0.07$), $p < .05$. The two feedback conditions WE and CA did not differ in mean post-test performance with a contrast estimate of -0.02 ($SE = 0.09$), $p > .05$.

3.2.2. Explorative analyses on transfer

See Table 2 for the descriptive statistics: the mean scores and SD's of the subsets of transfer test items we used for the explorative analyses. Cronbach's α for the six transfer test items was very low ($\alpha = 0.24$), so we conducted some explorative analyses. Based on the correlation matrix of the transfer items, we tried to identify clusters of items with positive correlations. A subset of three items appeared to form a cluster (Cronbach's $\alpha = 0.54$). These items were the two reflection vs. intuition tasks by Stanovich et al. (2016) combined with one of the conjunction fallacy tasks (Tversky & Kahneman, 1983). An explorative ANOVA on the sum scores of these three items failed to reveal a significant effect of Condition and the effect size was small to medium ($\eta_p^2 = 0.05$). The three other items were analysed separately. Again, there were no significant results and the effect sizes of these tasks were small to medium.[5]

4. Conclusion and discussion

The aim of the experiment was to examine the effect of elaborative feedback in the form of worked examples on reducing confirmation bias and transfer compared to correct answer feedback or no feedback. We hypothesized that providing elaborative feedback would be more effective than providing correct answer feedback, or no feedback. The results showed that both feedback types were beneficial for learning how to avoid confirmation bias in hypothesis testing tasks compared to practising without feedback, or no practising. This finding is in line with the findings by Butler et al. (2013), and corroborates the effects of feedback on learning that were reported in other domains.

The differential effect of feedback type on learning might be due to various reasons. The first reason might be related to the remarkable finding that we did not find any differences between the WE and the CA condition in time spent on practice and reading the feedback. This finding might have emerged because participants in the WE condition did not engage in a deep-level processing of the worked out solutions. Participants might have focused on *what* the correct solution steps were but might not have focussed on *why* these solution steps were correct (e.g., Renkl & Atkinson, 2010). As a result, the benefit of WE feedback over CA feedback might have not been effectuated. Finding an explanation for the deviation of the results from the expectations might however be an example of hindsight bias (e.g., *p*-hacking and subsequent inflations of false positive rate (Simmons, Nelson, & Simonsohn, 2011) and HARKing: hypothesizing after the results are known (Kerr, 1998)). To increase the credibility of our study and to prevent ourselves from our own cognitive biases, we pre-registered our research plan.

Secondly, the order in which the information was presented in the hypothesis testing tasks might also have influenced the WE-CA comparison. Concerning the order of presenting information, in their study,

[4] We also conducted an exploratory principle component analysis (PCA) on the six items of the post-test but the results did not deviate from the One-way ANOVA.

[5] We also conducted an exploratory principle component analysis (PCA) on the six items of the transfer test but the results did not deviate from the exploratory One-way ANOVA.

Jonas et al. (2001) presented the information to participants in either a sequential or simultaneous order. Sequentially presenting information caused a stronger preference for supporting information (i.e., confirmation bias) than simultaneous presenting. Our participants received the information on Jonas et al.'s hypothesis testing task on the post-test simultaneously, which might have resulted in lower levels of confirmation bias overall. In addition, it might be possible that a ceiling effect occurred because most participants probably found out that choosing balanced information is the principle, which made it hard to find differences between conditions.

According to Nickerson (1998), in general people do not perform well on Wason four-card selection tasks because its abstractness does not relate to everyday hypothesis testing. Many studies come to the same conclusion (e.g., Kellen & Klauer, 2019; Nickerson, 1998; Ragni, Kola, & Johnson-Laird, 2018; Van Peppen et al., 2018). In addition, the complexity of the four-card selection task lies in the fact that people over and over select the card(s) that confirm(s) the rule and seldomly select a card that might disconfirm the rule (Ragni et al., 2018; Wason, 1968). In other words, people often do not consider the opposite spontaneously but use heuristics that allow for fast information processing and that lead to thinking errors and biased reasoning when instead more reflective thinking must be applied (see Evans, 2003 on Type 1 and 2 processing). During the practice phase, participants in the WE and CA condition might have learned what the correct answers are to this kind of tasks given them an advantage on the post-test over the other conditions. However, due to the task complexity, participants in the WE condition might not have been able to fully understand the underlying rules of logic of the Wason four-card selection tasks they received in their feedback. As a result, this extra information could not be used to give them an advantage over the participants in the CA condition on the post test.

Participants did not receive customized feedback but general feedback, which differed in the degree of elaboration (the correct answer only or elaborative feedback). This might have caused superficial studying when a participant givens a correct answer, because it is then conceivable that a participant does not read the solution thoroughly. For future research, providing customized feedback on the given answer might be more effective: e.g., when a participant gives the correct answer, one could only give the elaboration on the solution steps without an explicit referral to the correct answer; this might prevent a superficial processing of the correct answer only. By contrast, in case of an incorrect answer more background information in combination with elaboration on the solution might enhance deeper learning. Also, exploring the learning effects of other types of feedback on confirmation bias tasks than the ones used in the present study might be considered in future research. Hattie and Timperley (2007, p. 88) for instance mention feed up ("Where am I going?") and feed forward ("Where to next?") next to feedback ("How am I doing?").

For the transfer outcome measure, we did not find any effects of the conditions in our experiment. This was most probably due to the low reliability of the transfer test. Hence, other transfer test items must be created if researchers plan a follow-up study along the lines of the present study. The number of test items based on the three transfer test items that showed a somewhat acceptable (but still low) reliability, can be extended for a transfer test. Interesting to mention is that the explorative ANOVA of the transfer test tasks based on Snyder and Swann (1978) revealed the same pattern as the learning phase although these findings were not significant. Although the difference was not significant, this pattern could indicate that participants who received feedback (WE and CA) were able to apply the learned knowledge to tasks that differ only slightly on structure compared to the learning tasks, because the Snyder and Swann-based task most closely resembled the hypothesis testing tasks from the instruction and learning phase. However, and also not significant, the reversed learning pattern was found on the base rate fallacy problem based on Stanovich et al. (2016), because participants who received no treatment (NT) scored the highest and participants who received worked example feedback (WE) scored the lowest on this task. Instruction, practise and feedback did probably not lead to the relevant knowledge structure to achieve transfer. Therefore, applying COS might have hindered participants in solving this task. However, as reliability was very low and results were not significant, the interpretation of these patterns must be exercised with great caution but it is interesting for further research because transfer effects are difficult to establish in unbiased reasoning studies (e.g., Heijltjes, 2014; Van Brussel, Timmermans, Verkoeijen, & Paas, 2018; Van Peppen et al., 2018).

Apart from transfer, we did not find any effects of our conditions on mental effort and time spent on practising. These results might be explained by Evans' dual processing theory (2003). If our participants applied a heuristic, Type 1 thinking is sufficient for learning. Since our results on transfer were not significant, one might assume that participants struggled to apply the learned strategy to new contexts.

4.1. Limitations

The study has some limitations. First, participants were not asked to reason about their answers on the tasks. Research shows positive learning and transfer effects of e.g. self-explanation on tasks through which participants are asked to reason on their anser (Chi, Leeuw, Chiu, & Lavancher, 1994; Fonseca & Chi, 2013). However, other research did not show effect of self-explanation prompts on learning and transfer (Van Peppen et al., 2018). For further research, it is interesting to gain more insight in the reasoning process in order to enhance learning and transfer performance, taking e.g., insights from the relational reasoning literature into account that focus on the support of analogies, anomalies, antinomies, and antitheses to enhance reasoning (Danielson & Sinatra, 2017; Grossnickle et al., 2016).

Also, elaborative feedback may be more important in drawing inferences, or applying of rules in more complex situations (Bangert-Drowns et al., 1991). In the present study, participants completed rather abstract tasks in a short, on-off intervention at the university. For future research, more complex tasks which deal with (simulated) real-life situations are interesting to use to find effect of feedback type on learning, for example in-class debates or small group discussions with immediate process feedback by the teacher. Also, it is unclear whether learning effects persist in the longer term. However, this is an interesting question for future research to address because a central educational goal is to achieve learning effects in the longer term.

The tasks used in our study are used often in confirmation bias studies. However, measuring proneness to confirmation bias with a self-report scale is suggested by Rassin (2008). He states that "the concept of confirmation is multifaceted and therefore by definition not internally reliable" (Rassin, 2008, p. 92). Data from the Rassin's Confirmation Inventory (CI) self-report might be a useful to collect in addition to performance measurements in future research to interpret pre-test post-test performance on the four-card selection tasks and hypothesis testing tasks that were used in the current study. However, this measurement needs more validation because the psychometric data of the CI are limited.

Finally, the low Cronbach's α on the post-test and transfer test will be discussed here. Although we used confirmation bias tasks and transfer tasks for the post-test that were validated by previous research

(Jonas et al., 2001; Stanovich et al., 2016; Wason, 1968), combining two tasks in the pre-test and post-test did not lead to one unidimensional construct that measured the level of confirmation bias.

Retrospectively, this was most probably due to the low level of relevant prior task knowledge, resulting in response guessing and consequently a high degree of non-systematic variance in the pre-test sum scores. This was mainly - and unexpectedly - due to the hypothesis testing tasks. However, the other four learning tasks did provide a reliable sum score on the post-test and on these tasks exactly the same effects were found as with the less reliable pre-registered measure. However, for future research, the type of test tasks needs rethinking.

The construct problem also occurred in the transfer test. This might be a result of a priori assuming that these tasks measured a more heterogeneous construct of transferring COS to other biases than the confirmation bias. However, the unreliability is caused by the "Linda"-task, the base rate task and the "extra vert" task by Snyder and Swann (1978).

4.2. Significance of the study

Our findings contribute to the existing body of knowledge on teaching CT, especially on instructional design studies including feedback. Feedback was not considered before in confirmation bias research. The results of the present study showed that adding feedback to confirmation bias practice tasks reduces confirmation bias more than practising only but that elaborate feedback in the form of a worked example does not enhance performance more than providing correct answer feedback on practice tasks. This is partly in line with other findings in the feedback literature that show mixed results of various types of feedback and performance outcomes (e.g., Butler et al., 2013; Hattie, 2009; Hattie & Timperley, 2007; Roelle et al., 2017; Shute, 2008). Therefore, further research can focus on increasing the effects of feedback on reducing the confirmation bias through COS.

Moreover, the results from our study are relevant from a methodological perspective. In the vast majority of educational psychology studies, null hypothesis significance testing (NHST), or *p*-hacking, is used to make inferences about population parameters based on sample statistics. In NHST, a *p*-value is calculated, which is a conditional probability of observing a particular value of a test-statistic or a more extreme value given the null hypothesis is true *and* given that the researchers adhere to a predefined sampling and analysis plan. However, researchers often face considerable degrees of freedom when sampling observations and analysing their data. When these researchers' degrees of freedom are not exposed and not taking into account, this will lead to a system in which the probability of false positives becomes very high (see e.g., Nosek, 2017; Simmons et al., 2011). To prevent this inflation of the false positive rate, researchers should pre-register their hypotheses, their sampling plan and their analysis plan. In addition, pre-registration enhances transparency and accountability (Munafò et al., 2017; Nosek & Lakens, 2014; Simmons et al., 2011). Also, our pre-registration identified a problem concerning the reliability of the measurements used. Although the tasks that were used are regarded as typical confirmation bias tasks and reasoning tasks, we were not able to compose a reliable set of items. Based on the pre-registration, we could not ignore this problem but instead this calls for further methodological action to improve these measurements. Without pre-registering research plans, measurement problems will continue to be underexposed within the research community and *p*-hacking will continue to exist (Chambers & Munaf, 2013; Munafò et al., 2017; Nosek & Lakens, 2014; Simmons et al., 2011). Hence, pre-registration is very helpful in identifying measurement problems in the field. This is an important first step in addressing them. When addressing these issues, data explorations, which are perfectly well possible in pre-registered study, might point at fruitful directions that could be tested in future research. Hence, the present study could serve as an example for other studies because by doing so, we protected ourselves from biasing the results. Pre-registering is one way to actively contribute to ways to achieve more transparency and objectivity in science and in our view, it would be good, if pre-registration would become the norm for educational psychology studies in which hypotheses are tested and counteract the confirmation bias and publication bias in science.

In addition, the findings of the study may also have implications for educational practice. It is important to help students to learn the CT skills that they need to face the future challenges in their work and life. Therefore, they need to transfer what they have learned in school to real-life situations. The findings of our study demonstrate that watching an instructional video and practise tasks with feedback contributes to reducing confirmation bias. However, it is necessary that future research focusses on supporting students in acquisition of transferring these skills in order to enhance unbiased reasoning and decision making. Therefore, it is recommended that cues are given to support transfer skills.

To sum up, learning critical thinking skills such as reducing confirmation bias, is considered to be very hard, because people are resistant to change and have the tendency to cling to their initial beliefs when contradictory information is presented (Douglas, 2000). More research is warranted to prevent people from becoming more divided, but instead let them realize that they can learn to make better decisions from standing in someone else's shoes.

Ethical considerations

All subjects gave informed consent prior to participating in the present experiment.

Data availability

All data are available on the OSF page: osf.io/thn6x.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Acknowledgments

The authors would like to thank Lara van Peppen, MSc for her practical help with the preparation of the experiment in the Erasmus Behavioural Lab, Tatia Gruenbaum, MA for revising the text and Madelief van Meer and Fleur de Vries for testing and checking the test material.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix. Four-card selection practice task with an example of elaborative as well as correct answer feedback (translated from Dutch)

Number and colours

On the cards below, there is a number on one face and a colour on the opposite face. Which card(s) do you have to turn over to test the following rule:

"If the card shows an even number on one face, then its opposite face is red"

RED **YELLOW** **3** **8**

Elaborative feedback:

The rule was: "If the card shows an even number on one face, then its opposite face is red". You apply "consider the opposite" by looking for a card that confirms the rule and a card that disconfirms the rule ("the opposite").

In this case, you turn "8" over, because you have to check whether the opposite face is indeed "red" and not accidentally "yellow".

However, you also have to turn "yellow" over because the only way in which the rule can be disconfirmed is to find a card with "8" and "yellow".

"3" and "red" can be ignored, because it has not been said that there should be an even number on a red card and for odd numbers such as "3" there is no requirement at al.

Correct answer feedback:

The correct answer was: the "yellow" card, and the "8" card.

References

Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, A. C., & 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.

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*(4), 1102–1134. https://doi.org/10.3102/0034654308326084.

Adame, B. J. (2016). Training in the mitigation of anchoring bias: A test of the consider-the-opposite strategy. *Learning and Motivation, 53*, 36–48.

Bangert-Drowns, R. L., & Bankert, E. (1990). *Meta-analysis of effects of explicit instruction for critical thinking.* Boston, MA: Paper presented at the American Educational Research Association.

Bangert-Drowns, R. L., Kulik, C., Kulik, J., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213–238.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612–637. https://doi.org/10.1037//0033-2909.128.4.612.

Beaulac, G., & Kenyon, T. (2014). Critical thinking education and debiasing. *Informal Logic, 34*(4), 341–363.

Butler, A. C., Godbole, N., & Marsh, E. J. (2013). Explanation feedback is better than correct answer feedback for promoting transfer of learning. *Journal of Educational Psychology, 105*(2), 290–298.

Catapano, R., Tormala, Z. L., & Rucker, D. D. (2019). Perspective taking and self-persuasion: Why "putting yourself in their shoes" reduces openness to attitude change. *Psychological Science, 30*(2), 1–12. https://doi.org/10.1177/0956797618822697.

Chambers, C., & Munaf, M. (2013). Trust in science would be improved by study pre-registration. Retrieved from https://www.theguardian.com/science/blog/2013/jun/05/trust-in-science-study-pre-registration.

Chi, M. T. H., Leeuw, N., Chiu, M.-H., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*(3), 439–477. https://doi.org/10.1207/s15516709cog1803_3.

Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology, 79*(4), 347–362.

Danielson, R. W., & Sinatra, G. M. (2017). A relational reasoning approach to text-graphic processing. *Educational Psychology Review, 29*(1), 55–72.

Douglas, N. L. (2000). Enemies of critical thinking: Lessons from social psychology research. *Reading Psychology, 21*(2), 129–144.

Evans, J. S. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences, 7*(10), 454–459.

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction.* Millbrae, CA: The California Academic Press.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

Fonseca, B. A., & Chi, M. T. H. (2013). *Instruction based on self-explanation. Handbook of research on learning and instruction.* Routledge.

Grossnickle, E. M., Dumas, D., Alexander, P. A., & Baggetta, P. (2016). Individual differences in the process of relational reasoning. *Learning and Instruction, 42*, 141–159. https://doi.org/10.1016/j.learninstruc.2016.01.013.

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American Psychologist, 53*(4), 449–455.

Haselton, M. G., Nettle, D., & Andrews, P. W. (2015). The evolution of cognitive bias. In D. M. Buss (Ed.). *The handbook of evolutionary psychology* (pp. 724–746). Hoboken, NJ, USA: John Wiley & Sons Inc.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* London: Routledge.
Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning.* London: Routledge.
Hattie, J. (2015). The applicability of Visible learning to higher education. *Scholarship of Teaching and Learning in Psychology, 1*(1), 79–91.
Hattie, J., Biggs, J., & Purdie, N. (1996). Effects of learning skills interventions on student learning: A meta-analysis. *Review of Educational Research, 66*(2), 99–136.
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.
Heijltjes, A. (2014). *Cultivating critical thinking: The effects of instructions on economics students' reasoning. (PhD).* Rotterdam: Erasmus University Rotterdam WorldCat database.
Hirt, E. R., & Markman, K. D. (1995). Multiple explanation: A consider-an-alternative strategy for debiasing judgments. *Journal of Personality and Social Psychology, 69*(6), 1069–1086.
Johnson, R. H., & Hamby, B. (2015). A meta-level approach to the problem of defining 'critical thinking'. *Argumentation, 29*(4), 417–430. https://doi.org/10.1007/s10503-015-9356-4.
Jonas, E., Schulz-Hardt, S., Frey, D., & Thelen, N. (2001). Confirmation bias in sequential information search after preliminary decisions: An expansion of dissonance theoretical research on selective exposure to information. *Journal of Personality and Social Psychology, 80*(4), 557–571.
Kahneman, D. (2011). *Thinking, fast and slow.* Macmillan.
Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review,* (4), 237–251.
Kellen, D., & Klauer, K. C. (2019). Theories of the Wason selection task: A critical assessment of boundaries and benchmarks. *Computational Brain & Behavior.* https://doi.org/10.1007/s42113-019-00034-1.
Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196–217.
Larrick, R. P. (2004). Debiasing. In D. J. Koehler, & N. Harvey (Eds.). *Blackwell handbook of judgment and decision making* (pp. 316–338). Malden, MA: Blackwell Publishing Ltd.
Lobato, J. (2006). Alternative perspectives on the transfer of learning: History, issues, and challenges for future research. *The Journal of the Learning Sciences, 15*(4), 431–449.
Lord, C., Lepper, M., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology, 47*(6), 1231–1243.
Marin, L. M., & Halpern, D. F. (2011). Pedagogy for developing critical thinking in adolescents: Explicit instruction produces greatest gains. *Thinking Skills and Creativity, 6*(1), 1–13.
Markovits, H., & Brunet, M.-L. (2012). Priming divergent thinking promotes logical reasoning in 6- to 8-year olds: But more for high than low SES students. *Journal of Cognitive Psychology, 24*(8), 991–1001.
McNutt, M. (2014). Reproducibility. *Science, 343*(6168), 229.
Mehta, S. R., & Al-Mahrooqi, R. (2015). Can thinking be taught? Linking critical thinking and writing in an EFL context. *RELC Journal, 46*(1), 23–36.
Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing decisions: Improved decision making with a single training intervention. *Policy Insights from the Behavioral and Brain Sciences, 2*(1), 129–140.
Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., ... Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*(1), 1–8. https://doi.org/10.1038/s41562-016-0021.
Mussweiler, T., Strack, F., & Pfeiffer, T. (2000). Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin, 26*(9), 1142–1150.
Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220.
Niu, L., Behar-Horenstein, L. S., & Garvan, C. W. (2013). Do instructional interventions influence college students' critical thinking skills? A meta-analysis. *Educational Research Review, 9*, 114–128.
Nosek, B. A. (2017). Opening science. *Open: The philosophy and practices that are revolutionizing education and science* (pp. 89–99). Ubiquity Press.
Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192.
Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*(4), 429–434.
Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*(1), 1–4.
Paas, F., & Van Merriënboer, J. J. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors, 35*(4), 737–743.
Qualtrics (2017). Qualtrics [12-2017]. Provo, Utah, United States. Retrieved from http://www.qualtrics.com.
Quatrociocchi, W., Scala, A., & Sunstein, C. (2016). Echo chambers on Facebook. Retrieved from https://ssrn.com/abstract=2795110.
Ragni, M., Kola, I., & Johnson-Laird, P. N. (2018). On selecting evidence to test hypotheses: A theory of selection tasks. *Psychological Bulletin, 144*(8), 779–796. https://doi.org/10.1037/bul0000146.
Rassin, E. G. C. (2008). Individual differences in the susceptibility to confirmation bias. *Netherlands Journal of Psychology, 64*(2), 87–93. https://doi.org/10.1007/BF03076410.
Reeves, A. N. (2014). Written in black & white. Exploring confirmation bias in racialized perceptions of writing skills. Yellow Paper Series.
Renkl, A., & Atkinson, R. (2010). Learning from worked-out examples and problem solving. *Cognitive load theory* (pp. 89–108). Cambridge University Press.
Roelle, J., Rahimkhani-Sagvand, N., & Berthold, K. (2017). Detrimental effects of immediate explanation feedback. *European Journal of Psychology of Education: A Journal of Education and Development, 32*(3), 367–384.
Schwind, C., Buder, J., Cress, U., & Hesse, F. W. (2012). Preference-inconsistent recommendations: An effective approach for reducing confirmation bias and stimulating divergent thinking? *Computers & Education, 58*(2), 787–796.
Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.
Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366.
Snyder, M., & Swann, W. B. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology, 36*(11), 1202–1212.
Stanovich, K. E., West, R. F., & Toplak, M. E. (2016). *The rationality quotient: Toward a test of rational thinking.* Cambridge, Massachusetts: The MIT Press.
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285.
Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction, 2*(1), 59–89.
Sweller, J., Van Merrienboer, J. J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251–296.
Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293–315.
Van Brussel, S., Timmermans, M. C. L., Verkoeijen, P. P. J. L., & Paas, F. (2018). 'Consider the Opposite' – Effects of context in an instructional video on reducing student-teachers' confirmation bias (Manuscript submitted for publication).
Van der Kleij, F. M., Eggen, T. J. H. M., Timmers, C. F., & Veldkamp, B. P. (2012). Effects of feedback in a computer-based assessment for learning. *Computers & Education, 58*(1), 263–272. https://doi.org/10.1016/j.compedu.2011.07.020.
Van Gog, T., & Rummel, N. (2010). Example-based learning: Integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review, 22*(2), 155–174.
Van Peppen, L. M., Verkoeijen, P. P. J. L., Heijltjes, A. E. G., Janssen, E. M., Koopmans, D., & Van Gog, T. A. J. M. (2018). Effects of self-explaining on learning and transfer of critical thinking skills. *Frontiers in Education, 3*, 100. https://doi.org/10.3389/feduc.2018.00100.
Vollmeyer, R., & Rheinberg, F. (2005). A surprising effect of feedback on learning. *Learning and Instruction, 15*(6), 589–602.
Ward, M., & Sweller, J. (1990). Structuring effective worked examples. *Cognition and Instruction, 7*(1), 1–39.
Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*(3), 129–140.
Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology,* (3), 273–281.
Yang, Y.-T. C., & Chou, H.-A. (2008). Beyond critical thinking skills: Investigating the relationship between critical thinking skills and dispositions through different online instructional strategies. *British Journal of Educational Technology, 39*(4), 666–684.