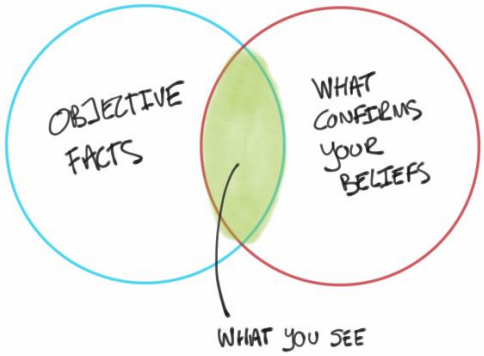
MASTER'S THESIS
Learning in Complex Systems
November 2019

Confirmation Bias: Prevalence and Debiasing Techniques



Elise Holth

OsloMet – Oslo Metropolitan University

Faculty of Health Sciences

Department of Behavioral Sciences

Acknowledgments

First, I want to thank my supervisor, Jan Wright, for useful feedback and ideas. Second, I would also like to thank the participants in the experiment for providing me with data. Finally, I want to thank my mother and father for support, interesting discussion, help with Visual Basic, and for the rewards to my participants.

Note: The illustration on the front page was retrieved from https://fs.blog/2017/05/confirmation-bias/

Abstract

‘Confirmation bias’ is a phenomenon that has been recognized by philosophers of science through centuries. The phenomenon is commonly defined as a tendency to favor confirmatory evidence in support of already stated hypotheses or response patterns summarized as “beliefs”, rather than disconfirmatory evidence. Such behavior involves missing potentially important truths. The phenomenon has been discussed in relation to numerous disciplines and other areas, such as law enforcement, politics, social media, psychology, and science. It has been suggested that the errors that exemplify the bias is to blame in cases from superstition to innocent convictions and even the event of Trump becoming the president of the United States. The research literature with respect to experiments demonstrating the phenomenon as well as experiments testing different debiasing techniques is reviewed, with a focus on the latter. In addition, a behavior-analytic approach is proposed with respect to an account of the behavior considered as showing the bias and to an explanatory model of such behavior. Moreover, two types of training are suggested and tested as debiasing techniques, namely training with explanatory feedback and training with correct/incorrect feedback. The extent to which skills generalized to novel exemplars was measured. Two participants disqualified for further participation as a result of demonstating to already have the skills of interest established in their repertoire, whereas considerable improvement was demonstrated in 13 out of the remaining 20 participants. Eleven participants also showed generalized skills.

Sammendrag

«Confirmation bias», på norsk av og til kalt «bekreftelsesfellen», er et fenomen som har blitt beskrevet av filosofer gjennom flere århundrer. Fenomenet blir vanligvis definert som en tendens til å foretrekke bekreftende, fremfor avkreftende bevis for en hypotese eller et responsmønster som oppsummeres av merkelappen «oppfatninger». Denne typen atferd innebærer at en går glipp av potensielt viktige sannheter. Fenomenet har vært diskutert innen en rekke disipliner og andre områder, som for eksempel advokat- og politibransjene, politikk, sosiale medier, psykologi og vitenskap. Enkelte hevder at denne bekreftelsestendensen har skylda for alt fra overtro til domfelling av uskyldinge personer, og til og med at Trump ble USAs president. Forskningslitteraturen gjennomgås med hensyn til eksperimenter som demonstrerer fenomenet, i tillegg til eksperimenter som innebærer testing av ulike måter å unngå at folk gjør den typen feil. Sistnevnte er mest vektlagt. En atferdsanalytisk tilnærming til fenomenet og forklaringer på det presenteres. Videre beskrives et eksperiment der to potensielle prosedyrer for å redusere feil som «confirmation bias» eksemplifiserer sammenlignes: trening med forklarende tilbakemeldinger og trening med tilbakemeldinger i form av «riktig»/«galt». Eksperimentet innebærer og måling av grad av generalisering av ferdigheter til nye eksemplarer. To deltakere ble diskvalifisert fordi de allerede hadde slike ferdigheter etablert i sitt atferdsrepertoar, mens en betydelig prestasjonsforbedring ble demonstrert hos 13 av de 20 kvalifiserte deltakerne. Elleve av disse viste i tillegg generaliserte ferdigheter.

Table of contents

List of Tables and Figures

The Prevalence of Confirmation Bias: A Review

Elise Holth

Oslo Metropolitan University

Abstract

‘Confirmation bias’ is a phenomenon in which people tend to favor confirmation rather than disconfirmation with respect to an already formulated hypothesis or response pattern summarized as “belief”. The phenomenon appears to be applicable across disciplines. Intelligence analysis, law enforcement, politics, science, and superstition are some of the areas to which its relevance has been suggested. The phenomenon exhibits potential danger in relation to multiple situations. The present review summarizes studies of particular relevance to variables that influence the likelihood of confirmation bias, and particularly studies concerned with debiasing techniques. In addition, some gaps in the literature are identified and discussed. A behavior-analytic point of view is provided, in addition to some further suggested steps towards overcoming the problem.

Key words: confirmation bias, debiasing, confirmation, disconfirmation, behavior analysis, decision maker, task characteristics

The Prevalence of Confirmation Bias: A Review

The term ‘confirmation bias’ usually refers to the tendency to search for evidence supporting an already stated hypothesis, or to the tendency to interpret evidence in a favoring manner to the hypothesis (e.g., Nickerson, 1998). The phenomenon has been recognized by numerous of logicans and philosophers of science over the years, and is currently a subject of widespread interest. Illustrative of the current focus, a Google search using the keywords “confirmation bias”, had approximately 1,910,000 hits, as of Oct. 6th, 2019, compared with only approximately 184,000 hits for “differential reinforcement”, 101,000 hits for “contingencies of reinforcement”, and 55,700 hits for “stimulus equivalence”, all of which have been important in behavior analysis over a long period. In addition, Nobel Memorial Prize laureate in Economic Science, psychologist Daniel Kahneman (2011), discussed the phenomenon of confirmation bias in his bestseller book titled “Thinking Fast and Slow”.

In Nickerson’s (1998) review of the confirmation bias, he assessed the problem to be one of the most widespread in relation to human reasoning. He suggested that a solution might reduce many types of misapprehensions. Further in his article he gave examples from fields such as medicine, science, and the justice system, among others. An illustrative example of people who are concerned only with exemplars supporting their theory, is the case of the creationists (Hoksnes, 2010, “Confirmation bias: the worst of all biases”). They believe that the Earth was created 6,000 years ago, have total confidence in the Grand Canyon as proof of Noah’s flood and dismiss all scientific knowledge otherwise relevant to how the canyon originated. Another conspiration example discussed by Hoksnes, is the search to prove that Francis Bacon wrote all of Shakespear’s plays. On top of that, some have claimed that Bacon hid all sorts of codes in the texts. The plays contain an enormous amount of letters. Obviously, some combinations of letters have been repeated in other contexts without that

being hard-core evidence of any direct connection between them. Ironically, in 1620, the british philosopher Francis Bacon (n.d.), noted the following:

> The human understanding when it has once adopted an opinion (either as being the received opinion or as being agreeable to itself) draws all things else to support and agree with it. And though there be a greater number and weight of instances to be found on the other side, yet these it either neglects and despices, or else by some distinction sets aside and rejects, in order that by this great and pernicious predetermination the authority of its former conclusions may remain inviolate … it is the peculiar and perpetual error of the human intellect to be more moved and excited by affirmatives than by negatives … . (pp. 9-10).

Following Bacon, a large number of writers have discussed approaches to confirm and falsify hypotheses. The 18th-century philosopher David Hume has been considered the first to examine relations between confirmation and truth of the generality of a phenomenon (1955, originally published in 1748, as referred in Mynatt, Doherty, & Tweney, 1977). Hume held that while there will always be a finite number of scientific observations of a phenomenon, there will be endless numbers of predicated consequences of a universal statement. Therefore, according to Hume, it would be impossible to formulate universal laws in the name of science. On the other hand, in order to falsify a hypothesis in the form "If P then Q" – P symbolizing the hypothesized theory and Q symbolizing the occurrence of a predicted phenomenon – one has only to observe one incident in which P does not include Q.

More recently, the philosopher Karl Popper discussed Hume's problem of induction in his work on the method of falsification (Mynatt et al., 1977). Mynatt et al. argued that the form "If P then Q" essentially makes up the most of scientific hypotheses. Evidence confirming the truth of Q does not bring us closer to either confirming or disconfirming the truth of P. Nevertheless, the falsification of Q will in any case be tantamount to the

falsification of P. On the basis of this idea, Popper concluded that in the search of general scientific laws, a method of falsification should be the matter of interest. According to Popper, scientific laws can never be conclusively verifiable, becauce there is always the possibility that a later observation may conflict with the stated law. Logically, however, a single observation in conflict with a stated law is sufficient to falsify it. In practice, the issue of falsifiability is much more complex, because any observation carries the possibility of error. Hence, a single observation of a phenomenon that is in conflict with a stated law may never be sufficient to actually reject the law (Thornton, 2018).

In everyday-life, people often make judgements in favor of preliminary expectations or hypotheses. It has been proposed that people often concentrate on cases that are likely to confirm precurrent assumptions, at the expence of cases that are likely to disconfirm them (Klayman & Ha, 1987). People may selectively search for evidence by asking questions in which the answer is likely to be *yes*. An illustrative example may be when someone tries to make the predictions from a horoscope fit real-life events.

As opposed to Popper's principle of *falsifiability*, even scientists frequently search for data consistent with current beliefs (e.g., Wason, 1960; Greenwald, Pratkanis, Leippe, & Baumgardner, 1986; Nickerson, 1998; Kahneman, 2011). Psychologist, Peter Cathcart Wason (1960), was the first to use the term *confirmation bias* and to address the phenomenon experimentally. He conducted a number of experiments in the field that demonstrated the phenomenon in relation to laboratory hypothesis testing (e.g., Wason 1960,1968; Wason & Shapiro, 1971).

In addition, people may interpret information by favoring preexisting views. For example, interrogators, police officers, jurors, and verdicts may selectively color evidence and search for confirmation for their earliest assumptions regarding criminal cases, as opposed to considering alternative hypotheses. Lidén, Gräns, and Juslin (2018) pointed out that a

wrongful conviction must be the absolute worst case scenario in criminal cases. In addition, errors that exemplify the confirmation bias made by interrogators and others in the system may lead to false confessions, trauma, and/or fabrication of memories (Lilienfeld & Landfield, 2008). Contemporary researchers have investigated whether police officers' assumptions of guilt during interrogations were stronger with reference to apprehended suspects, than non-apprehended suspects (Lidén et al., 2018). Their conclusion was that their data indicated that suspects in custody often are considered guilty even though they have the right, by law, to remain assumed innocent until proven otherwise. The authors pointed out that techniques to reduce errors exemplifying confirmation bias are crucial in situations such as interrogations.

An example of a recent invention that may foster confirmation bias is algorithmic editing in social media. Such algorithms make sure that people are presented with information that they probably agree with, meaning that it is related to prior searches and even to their social media-friends' searches. Baer (2016) argued that such personalizing of appearing information is beneficial for online shopping, for example. However, he suggested that it is a threat to the very democracy of the United States by neglecting many important aspects of the society. The phenomenon of such algorithms has to do with the confirmation bias in that people get access to more information confirming their already existing views and less disconfirmative information at the same time. Furthermore, selective remembering is a way of making errors that examplify the confirmation bias. For example, Kahneman (2011) pointed out that what people remember when asked if a specific person is kind probably differ from what they remember if asked if the same person is mean.

Seeing that the confirmation bias is a recurrent phenomenon in many guises, attempts to teach people how not to make such mistakes could be important to prevent error in numerous situations. The present paper is concerned with different areas in which the concept

of ‘confirmation bias’ has been addressed. Additionally, a behavior-analytic interpretation of confirmation bias is briefly discussed. Such an interpretation is relevant to the question of how the phenomenon can be harnessed or prevented.

Method

Wason’s (1960, 1968) research was the starting point for the present paper. Further, the references included were identified primarily using the electronic database PSYCINFO/Ovid. The search words were *confirmation bias*, with limits *all journals* and *English language only*. This search was last run on Sept. 1st, 2019 (See *Figure 1*). The articles identified through this search were screened on the basis of their titles and abstract, and those which included questionnaires or interviews, participants under the age of 18, or had inappropriate aims (e.g., reviews or not operationalized target behavior) were excluded. Furtermore, the articles remaining were scanned on the basis of full-text assessment, and those that fit into the exclusion criteria were omitted, while the remaining articles are considered in the present paper. In addition, reference lists of these articles were scanned. Finally, the behavior-analytic sources were identified based on the section concerned with bias in our main textbook on Learning (Catania, 1998).

Everyday Superstition

Superstitious beliefs, either in relation to astrology, psychics, or religion, are seemingly well maintained by a few predictions appearing to be true or almost true, even though a majority of such predictions appear to be inaccurate. Additionally, people seem to fail to recognize the possibility that others, who do not claim to have predictive powers, could have made equally valid (or invalid) predictions (Nickerson, 1998). Gardner (1957, as referred in Nickerson, 1998) described the hunt for pyramide “truths”, largely in accord with the Shakespear/Bacon example. The advanced constructs contain a huge amount of numbers in relation to lengths, widths, heights, quantity of stone blocks, and so on, so that only ones

patience set the boundries for the revealation of historical or scientific numbers that coincide with them. On the basis of countless observations, many different numbers will appear. Some of them will probably coincide with numbers that are also relevant in other settings. Most of the numbers, which do not coincide with the theme of interest are simply ignored. These examples illustrate and suggest the generality of people's hunt for confirmation favoring a desired truth, while disregarding all disconfirmatory information, even when found in overweight.

Bacon (n.d.) was very much ahead of his time when he stated that men often focus on the few events that fulfill their beliefs and see through information that indicates the opposite, even though it is occurring in overweight. He even mentioned the problem in relation to sciences, where people also seek confirmation for an already existing conclusion and ignore disconfirmative data despite its relevance.

Experimental Studies

Wason conducted some of the earliest experiments in the area (e.g., Wason, 1960, 1968, 1968; Wason & Shapiro, 1971). His (1960) first study involved presenting participants with a simple contextual task, in which the experimenter listed three numbers and told the participants that the numbers conformed to a simple rule that he had formulated, concerned with the relationship between them. The participants' task was to state the rule as correctly as possible. In order to do so, they were asked to come up with three other numbers compatible with the rule. They were given feedback on whether their suggested numbers did or did not conform to the rule. This procedure was repeated until the participants wrote down a suggestion about a rule, and again (for up to five times) if they got it wrong. Out of the 29 participants, six reached the correct rule on their first try, whereas the rest reached one or more incorrect rules or did not reach a rule at all. Wason's interpretation of these results was

that most participants were only seeking confirming evidence instead of eliminating possibilities by seeking both confirming and disconfirming evidence.

Later on, Wason (1966, as referred in Wason, 1968 and in Nickerson, 1998) invented what he called the selection task. In this selection task participants were given the information: "There is a letter on one side of each card, and a number on the other side", and asked which one(s) of four cards that had to be turned over in order to decide whether a hypothesis, such as: "If there is a vowel on one side, there will be an even number on the other side", were true. The hypothesis was an extended version of a conditional argument of the form "If P, then Q". They were exposed to four cards with A (P), D ($\overline{P}$; symbolizing not P), 4 (Q), and 7 ($\bar{Q}$; symbolizing not Q) on them, respectively. Wason's results, as well as later replications (e.g., Hughes, 1966, as referred in Wason, 1968) have shown repeatedly that a majority of participants turned over the P-card, only. The second most common reaction was to turn over both the P-card and the Q-card, while very few were concerned with the $\overline{P}$-card or the $\bar{Q}$-card. Wason and Shapiro (1971) found that the rate of solving such problems increased when participants were given thematic problems, rather than abstract ones. In order to solve the problem, the participant had to turn over the P-card, to make sure there was a Q on the other side, and the $\bar{Q}$-card, to make sure there was not a P on the other side. The hypothesis did not involve that there has to be a P on the other side of a Q, therefore turning over the Q-card is unnecessary. Further, no symbol on the other side of the last card, with $\overline{P}$ on the visible side, could falsify the hypothesis.

Another experimental study on psychology students' preference for confirmation versus disconfirmation during hypothesis testing was conducted by Mynatt et al. (1977). The experiment was carried out to investigate whether participants were more concerned with observations that might confirm their preexisting hypotheses or if they would test alternative hypotheses. Among other things, the experimenters designed an environment in which the

participants were presented with events and objects similar to real-world exemplars. Three figures (a disc, a triangle, and a square) appeared on a screen and participants were tought how to fire small lighted dots from a corner on the screen, and were then asked to state a hypothesis concerning the dots' endpoint on the screen. Furthermore, they were presented with a screen showing five of the same kind of figures (a disc, two triangles, and two squares) as on the first screen. The instruction to hypothesize was repeated and participants were informed that this hypothesis could be identical to their first one, or they could change it. The experimenters had designed the two initial screens in a way thought to increase the likelihood of participants focusing on triangles in their hypotheses. Out of the initial 45 participants, 20 of them focused their hypotheses on triangles, and the experimenters selected them for further study. Random allocation was applied to spread participants over the three different instructional conditions. Participants in the three groups were informed that the job of a scientist was to confirm hypotheses (Group 1), disconfirm hypotheses (Group 2), or to test hypotheses (Group 3). The two first instructions each included a historical example of such a confirmation or disconfirmation, respectively. The next step in the experiment involved for participants to look at ten sets of screen picures presented to them, consecutively. For each set, they were instructed to pick one picture to use as a starting point for firing the small dots to test their hypothesis. Each of these sets consisted of one which gave the opportunity to add confirmation to their hypothesis, and one which gave the opportunity to falsify it. In spite of nearly identical distribution of number of confirmatory choices made by participants exposed to each of the three conditions, results revealed that while 73 % of the participants associated with the disconfirmation or the testing conditions ended up with hypotheses which were either correct or partially correct, only 40 % of the participants presented with the confirmation condition were. The experimenters emphasized one set of screens in particular, where 11 out of the 20 participants picked the disconfirming test, and 10 of them ended up with a correct or

nearly correct hypothesis. In comparison, only 4 of the other 9 participants ended up with a correct or nearly correct hypothesis. Mynatt et al. (1977) argued that these results revealed more evidence that people generally fail to consider alternative hypotheses. They drew a parallel to Wason's (e.g., 1960, 1968) findings, and suggested that the confirmation bias does not only apply to abstract tasks, but also to those that are more concrete in nature. Additionally, the results of this experiment suggested that people are able to make use of disconfirmation once presented with it, even though also suggesting a pervasive bias with respect to picking confirmation over disconfirmation.

As an illustrative example of the confirmation bias as a phenomenon, Greenwald et al. (1986) described the Wyatt-Campbell and Bruner-Potter studies from the 50's and 60's, where participants were asked to watch a blurry picture as the experimenter made it appear less and less blurry. At first, identifying the motive was impossible. However, at last, the picture appeared just slightly blurry, and participants were asked to tell what the motive was. The results showed that about 25 % of their guesses were correct. Moreover, their studies suggested that participants could identify those slighty blurry motives correctly in about 75 % of the cases when not previously exposed to the gradual focusing.

More than three centuries before these studies were conducted, Bacon (n.d.) noted the following: "But with far more subtlety does this mischief insinuate itself into philosophy and the sciences; in which the first conclusion colors and brings into conformity with itself all that come after, though far sounder and better." (p. 10).

Overcoming the Bias

Wason and Shapiro (1971) carried out an experiment to search for differences in participant performance between selection tasks with different task characteristics. Their experiment was designed to find out whether participants' selection task performance was affected by presenting familiar, rather than abstract, relations and concepts. One group of

participants was presented with classical selection tasks with abstract concepts and arbitrary relationships between them. In contrast, participants in a second group were presented with a rule concerning journeys made by the experimentor, containing a mode of transport, a destination, and the name of a day of the week. Two of the four cards that were presented had the name of a mode of transportation on them, whereas two had the name of a destination on them. The content of each card differed from all the others. Additionally, the name of a different day of the week appeared on the visible side of each card. Wason and Shapiro's results showed differences in performance related to different task characteristics. While only 12.5 % of tasks of the abstract form were solved, 62.5 % of the thematic tasks were solved.

McKenzie (2006) carried out two experiments to investigate differences in the occurence of confirmation bias with respect to abstract versus thematic or concrete materials. In the first experiment, participants were exposed to one of three conditions: abstract and statistics, concrete, or concrete and statistics. The first condition (abstract and statistics) involved presentation with a text on some imaginary species, Gloms and Fizos, living in an imaginary world. Participants in this group were informed that there were equal numbers of each species and instructed to guess whether each one out of four of them belonged to one or the other species, based upon the creatures yes-/no-responses to one of two questions each (related to what they wear and play). Participants were then presented with these responses, consecutively, together with statistics consisting of information about the percentage of each species which would have responded yes/no to each question. Two out of the four creatures responded to one question while two responded to the other. Additionally, two responded positively while two responded negatively, resulting in four different combinations. The diagnosticity of yes-/no-responses to the questions was different in that, for example, a yes-response to the question about play implied different probabilities of that particular creature being a Glom or a Fizo. Participants exposed to the second condition were instructed to guess

the gender of four peers based on peers' responses to some questions related to their height. The peers' reponses were presented in the same manner as the creatures' in the first condition. Participants were informed that there were equal numbers of each gender. In addition to the instuction used in the second condition, participants in the third condition were presented with some statistics with respect to the peers' responses (% of yes/no responses associated with male/female peers). All participants rated their guesses on an 11-point scale, ranging from 50, representing "blind guess", through 55, 60, 65, and so on to 100, representing "completely certain". The main finding of this experiment was that participants exposed to the concrete and concrete and statistics condition showed more sensitivity to diagnosticity than those exposed to the abstract and statistics conditon. This means that responses of participants in the abstract and statistics group were less affected by rarity information. McKenzie argued that such insensitivity to differential diagnosticity is not tantamount with confirmation bias. However, the combination with a preference for extreme values for their hypothesis rather than alternative hypothesis, demonstrates confirmation bias. Even in lack of statistical information, participants presented with concrete materials were more capable of guessing than those presented with abstract materials.

In the second experiment, participants were exposed to either the abstract and statistics condition or the concrete condition. Participants in the abstract and statistics group were presented with the information used in the first experiment with a few alterations with respect to the values in percentage for yes/no responses of each species: the numbers that differed were higher in experiment two. These participants were instructed to rate their confidence of each guess on an 21-point scale, ranging from 0, representing "certain not [hypothesis]", to 100, representing "certain is [hypothesis]". 50 on the scale still represented "blind guess". Participants in the concrete group were presented with information identical to that used in the first experiment, except that one of the questions for two out of the four asked peers was

altered to produce more extreme values. The differential diagnosticity that appeard when individuals responded differently to a question, affected participants exposed to concrete materials much more than participants exposed to abstract materials. Hence, the results replicated those of the first experiment. Accordning to McKenzie, this insensitivity to diagosticity observed together with an extremety bias for participants presented with abstract materials were evidence of a confirmation bias. The researcher also stated that because perfect sensitivity was not demonstrated by participants exposed to concrete materials, the possibility of confirmation bias among them was not excluded. In spite of this fact, he concluded that confirmation bias was more likely to occur among abstract than concrete materials.

Cook and Smallman (2008) considered the different attempts throughout the literature both to explain the confirmation bias and to search for debiasing techniques. They argued that in one type of explanation, researchers focused their attention on mental processes inside the decision maker's brain, while the other type of explanation was concerned with the stimuli in the decision maker's environment. Likewise, the first method of debiasing was supposed to manipulate mental processes in order for the participants to solve a given task correctly. They examplified this approach by applying training methods and informing partcipants about the bias, among other suggestions. According to the authors, the second debiasing technique involved manipulation of the decision makers' environment in order for their mental processes to concur with a given task. However, Cook and Smallman stressed the inadequate amount of applied research in the relation to methods for overcoming the confirmation bias.

Moreover, Cook and Smallman (2008) studied confirmation bias in relation to intelligence analysis. To ensure situations as close to reality as possible, they used naval trainee reservists' and analysts' as participants and made the experimental conditions similar to conditions familiar to the participants from their normal work situation. They investigated evidence analysis performance under controlled circumstances, and tested two suggested

debiasing techniques. One compared a graphical form of evidence with a simple text format, and the other compared the participant's own evaluation of evidence with a fictive analyst's evaluation. Each participants were presented with four different vignettes consecuitively, realistic in tactical essence. Each of these vignettes consisted of a hypothesis on a potential event that participants were asked to investigate, together with related information and evidence elements that they were asked to contemplate in their investigation. They were then asked to pretend to be a primary analyst in a real-world situation, and evaluate, select, and prioritize the evidence elements in a 5-min-time span for each of the four vignettes. The time limit was included to give the impression of realistic potential time preassure and participants were told to respond to the three different tasks to make sure that biased behavior was to be measured repeatedly. In the evaluation task, participants were told to rate eight different evidence elements in order from most to least relevant to the hypothesis. They also were presented with other fictive analysts' evaluations. These fictive analysts were four judges who evaluated the hypothesis' importance on a 7-point Likert scale, where 1 represented "strongly refutes", 4 represented "neutral", and 7, represented "strongly supports", before the experiment was carried out. In the graphical condition, participants evaluated elements by the same numeric ratings, used by the fictive analysts, by placing them in a graphical system, whereas they in the text condition evaluated by rating elements by typing a number between 1 and 7 next to each element on the screen. Participants were presented with the fictive analysts' placement of elements in the graphical system or number ratings for each of the elements. During the selection task, participants were told to select four, out of the eight evidence elements, which they assumed to be most important to investigate further in relation to the hypothesis. Finally, participants were told to prioritize evidence elements by ordering them from most to least important. The authors operationalized the occurences of confirmation bias differently for each step in the experiment. During the evaluation task bias

was measured by ratings of own evaluations of the relevance of elements as supportive, on average. In the selection task, bias was measured by an amount of selection of supportive-rated elements equal to or above average, and a systematic selection of more supportive elements in general. The final measure, during the rank-ordering task, was by increased rank-ordering of elements rated as more supportive, relative to those of less supportive ranking. Manipulations of neither of the two independent variables showed any significant impact on the dependent variable in the evaluation task. A significant effect was observed in the selection task in all of the conditions. Graphical presentation reduced the effect significantly. However, the effect was independent of participants' own versus fictive analysts' evaluation. The evidence elements that to the largest degree confirmed the hypothesis correlated with the largest degree of rank-ordering, suggesting a pervasive confirmation bias in relation to the rank-ordering task.

Rassin, Eerland, and Kuijpers (2010) conducted three experiments on the phenomenon of confirmation bias in relation to criminal investigations. In their first study, naïve participants read a file on a case of abuse. In this file, there was information about a man who was suspected to have beaten and kicked another man, and participants were told to determine whether the suspect was guilty or innocent, and were presented with a predetermined list of 20 further investigation options that they could select from if they found it appropriate. Ten of these investigations were concentrated on producing more evidence of the suspect's guilt, whereas the other ten were concentrated on exoneration of the suspect, or on producing evidence for alternative scenarios. In order for the experimenters to be certain that their fictious incriminating and exonerating investigations were interpreted as such by others, and to minimize the risk that others evaluated one type of investigations to be more important than the other, they carried out two pre-experimental studies. Participants in both pre-studies were presented with a short-version of the case file. In the first study, they were informed that some

of the investigations appeared to be incriminating, while others appeared to be exonerating. They were instructed to rate to what degree they thought each investigation was incriminating or exonerating on a 10-point scale, ranging from 0, symbolizing totally incriminating, to 100, symbolizing toally exonerating. Moreover, participants in the second pre-study were instructed only to rate how important they thought each investigation was on a 10-point scale, ranging from 0, equvalent to not at all, to 100, equivalent to very important. As a result of these pre-studies, three incriminating investigations were not included in further analysis, but remained present in the main experiments. In the first of their main studies, 58 % of the participants concluded that the suspect was guilty of the crime, and 45 % of the selected further investigations among them were incriminating. 42 % concluded that he was not guilty, and 57 % of them ordered investigations were incriminating. These findings demonstrated that a higher percentage of participants who presumed the suspect to be guilty, looked for further confirmation of guilt, wereas a higher percentage of participants who presumed his innocence looked for further confirmation of innocence – both illustrated that people look for information to confirm prior hypotheses.

In Rassin et al.'s second study, participants were presented with four versions of the case file used in the first experiment, together with the same 20 further investigation possibilities. In two versions of the file, the suspect confessed. In the other two versions, the suspect claimed he was innocent and witnesses stated that they were not sure whether he was guilty or not. Moreover, one of each of these versions were paired with a description of the victim's injuries as mild (like broken ribs and short-term amnesia). In contrast, the other ones of each condition were paired with the description of much more serious injuries (like paralysis from the waste down and permanent trouble speaking). These four conditions were assigned to four groups of participants, respectively. Results revealed an increasing selection of incriminating investigations increased with both increasing evidence strength and with

increasing seriousness of injury. The authors argued that the results suggested that people are inclined to gather evidence to confirm a suspect's guilt.

In their third study, participants were presented with the strong evidence/serious injury-version of the case file, used in the previous study. Participants in the first group were immediately asked if they would convict or acquit the suspect. Participants in the second group were told that they could order further investigations from the list, also used in the previous studies, and received written results from their chosen investigations (if they chose one or more of them). Further, these participants were exposed to either of four different condtions. In the first condition, the evidence resulting from the ordered investigations were positive in that exonerating investigations produced evidence against the suspect's guilt, while incriminating investigations produced evidence of his guilt. Subsequently, participants in the second condition received negative evidence in relation both types of investigations. Participants in the third condition received guilt confirmatory evidence to incriminating investigations, but no evidence to exonerating investigations. Finally, the last condition involved receiving exonerating evidence to exonerating investigations, but no evidence to incriminating investigations. Results showed that the conviction rates differed in different condtions. The main finding was that all participants in the second group who were exposed to the third condition (receiving only incriminating evidence to incriminating investigations) decided to convict the suspect, while less than half of the participants in the fourth condition (receiving only exonerating evidence to exonerating investigations) did. Selection of incriminating investigations correlated with rates of conviction, and participants who made such selections seemed to have invented additionally evidence against the suspect. Rassin et al. argued that despite having any stated opinion or belief prior to the presentation of a task, people tend to prefer confirmation over disconfirmation. They pointed out that Wason's selection task was illustrative of this.

Hernandez and Preston (2012) carried out two experiments to investigate the effect of disfluency on errors which examplify the confirmation bias in relation on already stated political views and experimental manipulations of such views. In their first experiment, participants were instructed to answer some questions about them selves (e.g., age, religious affiliations), and then to rate their ideogly in relation to political statements on a scale from 1, representing "strongly liberal", to 7, representing "strongly conservative". Further, participants were presented with an article with arguments pro the death penalty. Some of them read this text written in a common size 12 Times New Roman font, while the others read it in a light gray Haettenschweiler font, bold and italicized. All of the participants were then instructed to evaluate statements thematically related to the text. The researchers focused on three dependent variables which were considered as measures of participants' degree of agreement with the author of the text. Participants rated the statements from 1, representing "not at all", to 5, representing "extremely", with respect to reliability and intelligence of the argument, and belief in the argument. The dependent variables were the three measures of degree of agreement with the statements in the article. Based on participants' ratings of own idelogy, the experimenters scored them using values between 0, representing "most liberal", and 1, representing "most conservative". For comparison, participants presented with the fluent condition were scored as a 1, whereas participants presented with the disfluent condition were scored as a 0. The experimenters combined the two variables to calculate an interaction term, as a third variable. Finally, they used a multiple regression of these three variables to evaluate the degree of confirmation bias. Hernandez and Preston observed a significantly reduced bias effect of idelology in the disfluent condition, compared to the fluent condition. Accordingly, differences between liberal and conservative views appeared smaller following exposure to disfluent reading, than following exposure to fluent reading. Participants' initial report of ideology significantly affected their degree of agreement was

significant in the fluency group, but this relationship was not significant in the disfluency group. However, considering that the very definition of the confirmation bias has to do with the development of response patterns summarized as "beliefs", and not simply with evaluation of them, what was measured by Hernandez and Preston seems uncertain. The authors themselves added that perhaps they did not measure confirmation bias at all in this first study, by arguing that the measures were of pre-existing standpoints and not of statements randomly presented to them.

In their second experiment, Hernandez and Preston (2012) investigated the legal jugdment of participants online at Amazon.com. All of them were presented with the information of a person, suspected to have comitted a crime, and were told to decide whether the person should be convicted or not. For some of them, the information consisted of a description of the defendant's positive characteristics, like polite and a good sense of humor. However, other participants were presented with a negative description of the defendant, like that he had a history of diciplinary troubles and was cold. Furthermore, these two conditions were paired with four different presentation of the defendant's case file: fluently, disfluently, with a time constraint, and paired with a memory task. In the fluent condition , the information was presented with a size 16 Times New Roman font, while it in the disfluent condition was presented with a size 12 Times New Roman font, put through a copier three times to reduce the quality of it. The time constraint condition involved for participants to respond to the task when 3 min had passed on a timer on a timer visible to them. Finally, the memorization task involved for participants to memorize, through the experiment, nine common nouns that they were briefly presented with. Both the time constraint and the memory task conditions were presented together with the disfluent version of the article. Together, this constituted a 2 (positive and negative bias) x 4 (fluent, disfluent, disfluent + time constraint, and disfluent + memorization) factoral design. The defendant's case file

contained description of his accusation of robbery, even though the evidence was open to interpretation. At last, participants were asked to decide a verdict by choosing “guilty” or “not guilty”, a jail sentence from 0-5 months (5-point scale), and to make an assessment of their own certainty, from 1, symbolizing “extremely certain he is not guilty”, to 7, symbolizing “extremely certain he is guilty” (7-point scale). The authors found no significant difference when comparing the time constraint and the memorization conditions. Hence, they decided to merge the two conditions into one with the common term “cognitive load”, resulting in a 2 (positive and negative bias) x 3 (fluent, disfluent, disfluent + cognitive load) factoral design. The effect on the three dependent variables was measured with a 2 x 3 ANOVA (analysis of variance). The experiment led to no statistically significant effect of confirmation bias on the measure of length of jail sentence. However, effects were observed with respect to the measure of the verdict in all conditions but the disfluent. In the disfluent group, the amount of guilty verdicts varied by only 2 % between the positive and negative bias conditions, whereas the amount varied from 10 to 40 % in the other groups. In addition, biased effects appeard under the same conditions with respect to participants’ certainty of their verdict. Confirmation biased effects appeared in the fluency but not in the disfluency condition, in accordance with the authors predictions. In the disfluency condition, participants generally focused on disconfirmative evdence. Additionally, disfluency was not observed to mitigate bias during the cognitive load conditions. Hernandez and Preston argued that the finings indicated that the participants’ cognitve resources were occupied during the cognitive load conditions. Thus, participants were not able to disengage with the article. They did, however, acknowledge that they could not be sure that deeper processing was responsible for the disconfirmation effect.

Lidén et al. (2018) tested two suggested debiasing techniques in their studies of swedish police officers. First, they investigated whether apprehension of a suspect by the interrogator him- or herself affects their questioning mode during interrogation. To search for

the answer, they compared results from such situations with results from situations where an colleague or prosecutor was the one to apprehend the suspect before the interrogation. Even though their results suggested evidence for their hypothesis of this technique of altering the decision maker as a way of reducing confirmation bias, it did not reach a level of statistical significance. Second, the experimenters examined whether reduction of cognitive load mitigate the occurrence of confirmation bias. The reduction of cognitive load was operationalized as change in interrogation mode by alternating between interrogator-generated consecutive questioning and picking questions from a predetermined list. The degree of guilt presumption was measured by categorizing questions into neutral and guilt-presumptive questions. The autors examplified these categories with "Why did you push her?" as an example of a guilt-presumptive question and "What happened in the apartment?" as a neutral one. The experimenters rated the freely generated questions on a scale from 1 to 7, where 1 was equivalent to the presumtion of total innocence, 4 was equivalent to neutral, and 7 was equivalent to the presumtion of complete guilt. Results revealed that interrogation by pre-determined questions compared to interrogation by interrogator-generated questions reduced the confirmation bias. In addition, Lidén et al. conducted a follow-up study to investigate further their hypothesis that officers are more likely to question suspects with a large degree of guilt presumtion regarding suspects they chose to apprehend than suspects apprehended by others. Law and psychology students served as participants in this study, and the procedure was the same as in their first experiment, except for simple changes to reduce the time duration of the experiment (the scale for evaluating guilt presumtion of questions ranged from 1 (innocence-presumptive) to only 3 (guilt-presumptive) and that the participants generated three instead of six such questions). The results from their first experiment was replicated in this second study: apprehension in general lead to a larger degree of guilt-presumption, while the effect of alternating between different decision makers in relation to apprehension was not

significant. As a third experiment, the experimenters let half of the participants (psychology students) generate six questions freely and let the other half pick six questions from the predetermined list. They found that freely generated questioning mode led to a larger degree of guilt presumption than the other type of questioning mode. Lilienfeld and Landfield (2008) also supported the finding of Lidén et al. that confirmation bias in law enforcement workers occurred frequently. They pointed to double-blinded experimental designs as an example of a safeguard against such errors.

Numerous authors have pointed out the possibility that the if-then construction or frame when used in everyday language usually has a different definition than that of Wason's, namely that if-then is interpreted as a biconditional relation of "If and only if P then Q" (Nickerson, 1998). However, this description of how the if-then frame is often interpreted is simply a summary of the typical finding of these experiments rather than an explanation of those findings.

Wason and Shapiro (1971) suggested three aspects of their thematic tasks that may have contributed to better performance on this task than to their abstract tasks. First, they hypothesized that there was a chance that the participants manipulated the familiar terms in a more appropriately and carefully compared to the abstract ones. They suggested a way of testing this hypothesis, by presenting familiar concepts, but without the familiar relations, for example by putting names of different metals and fruits on each side of the cards, given a rule of the type "if metal p, then fruit q". Second, the authors argued that thematic tasks contained concrete relations ("travelling"), as opposed to arbitrary relations ("the other side of"), and pointed out that it might be this relations, above the terms themselves, that makes the participants more likely to solve the problem. Conversely to the previous, they proposed a task form to test this hypothesis, in which the concepts were abstract and the relations concrete, for example "Every time I go to K I travel by 3" (p. 69). Ultimately, Wason and

Shapiro suggested that the inclusion of names of week days contributed to forming a more logical narrative, which in turn made the participants focus equally on all four cards. Contradictory, results indicated that participants facing classical selection tasks tended to pay attention almost exclusively to those cards with correspondence to concepts referred to by the rule (e.g., Wason & Shapiro, 1971; Nickerson, 1998). Perhaps a possible way to test this hypothesis could involve adding an item in the classical selection task, different for each card, similar to the name of week days, but in arbitrary form, for example four different symbols, comparing results with participants presented with the classical task without the added items. According to Nickerson (1998), a common objection to Wason's selection task has been that the concept of confirmation bias may account only for abstract problems. As demonstrated by Wason and Shapiro, people seem to be slightly more likely to solve such selection-task problems concerned with familiar relations. However, if abstract hypotheses of the form "If P then Q" are the most common in science, as noted by Mynatt et al. (1977), then the objection may be of relatively low importance – at least with respect to the relevance of confirmation bias to scientific research.

A Behavior-Analytic Point of View

In a brief review of Wason's selection task, Catania (1998) pointed out that receiving confirmatory evidence is more reinforcing than receiving disconfirmatory evidence, an interpretation compatible with the very definition of the confirmation bias. A behavior-analytic view might also contribute with economical explanations. For example, Lilienfeld and Landfield (2008) claimed that interrogators may become close-minded when they have preexisting beliefs regarding a suspect's guilt. Thus, the interrogators fail to consider alternative hypotheses. Similarily, Hernandez and Preston (2012) claimed that changing the style of arguments can produce attitude change. From a behavior-analytic view, this does not serve as a satisfactory explanation (e.g., Skinner 1974; Hayes & Brownstein, 1986; Holth,

2001). Both *presumtion of guilt*, *close-mindedness*, and *attitude change* are cognitive constructs or summary labels that are entirely based upon observations of behavioral patterns. Categorization of suspects as guilty at the end of an interrogation correlates with a high degree of stated presumption of guilt in advance. Such summary labels are used as "explanations" of the phenomena they summarize (Holth, 2001). These are mentalistic explanations that do not serve as scientific explanations within the field of behavior analysis. If the entire goal is to predict behavior, mentalism sure can be useful. If different types of behavior covary, observing one type can serve as a basis for predicting the others, for example different types of behavior lumped together under the heading aggression. However, if control is an equally important goal, another kind of analysis is required (Holth, 2001). Skinner (1974) argued that such references are misleading in the sense that what really exists inside the skin is no such thing as cognitive constructs, and that what he called "mental way stations" stand in the way of the effective search for the causes of behavior to be found in the environment, or more specifically, in contingencies of reinforcement and in contingencies of survival (Skinner, 1969). In accordance with this behavior-analytic view, Gigerenzer (1991) was crystal-clear in his encouragement not to rely on the idea of "biases" as explanations of behavior considered as showing biases. He stressed that the kind of explanations popularly used in the area of social psychology has been to refer to mental flaws, often explained via the mind, which seeks after instances confirming the chosen hypothesis, and ignores discontirmatory instances. Further, social psychologists have suggested that this selective information seeking reinforces confidence. Moreover, Gigerenzer pointed out that the gap between confidence and relative frequency of correct responses has been labeled "overconfidence", and in turn, confirmation bias has been proposed to account for this overconfidence. Although both "overconfidence" and "confirmation bias" may be useful as

descreptive terms, they do not describe environmental events that constitute manipulable independent variables in behavior analysis.

Concluding remarks

The phenomenon of confirmation bias is usually defined as a tendency to search for confirming evidence for an already stated response pattern summarized as "belief" or hypothesis. McKenzie (2006) defined the phenomenon as behavior "that leads to systematic overconfidence in a focal hypothesis (i.e., the favored hypothesis or the hypothesis being tested)." (p. 577). Moreover, McKenzie interpreted this definition as the standard definition, and argued that it was supported by Nickerson (1998), among others. However, independent of how the standard definition is interpreted, the phenomenon has proven to apply even to a hypothesis presented to a given person who has no prior opinion on the case (e.g., Wason, 1960, 1968; McKenzie, 2006; Rassin et al., 2010).

McKenzie (2006) stressed that occurrence of confirmation bias is depentent on both a particular way of testing and interpreting. The phenomenon has been investigated in relation both to abstract and to thematic and concrete materials. McKenzie argued that mostly abstract materials had been used in these experiments, and pointed out that results might be limited to such abstract materials in laboratory settings. The particular combination of testing and interpreting is supposedly testing without being particularly concerned with differentially diagnosticity, and interpreting by favoring extreme questions. In his experiments, this combination was observed to a much larger degree in participants presented with abstract, than in those presented with concrete materials. Nevertheless, McKenzie pointed out that confirmation bias may have occurred across task characteristics, with concrete materials as well, suggesting that errors exemplifying confirmation bias might not be limited to abstract tasks after all.

Wason and Shapiro (1971) made some experimental suggestions on combining different task characteristics, for example by presenting tasks with abstract concepts with concrete relations or concrete concepts with abstract relations. This could be important steps towards a more general understanding of the phenomenon.

To summarize, the literature covering the phenomenon is rather widespread and important. Nevertheless, some research questions remain unanswered. Most of the research focused on variables that affect the extent of behavior that exemplifies confirmation bias have been concerned with differential outcomes of using abstract versus concrete materials. For the development of more effective strategies to overcome or avoid such traps, more comprehensive research is necessary. Additionally, further investigations are required in order to test the degree to which potential strategies continue to be effective over time. Are there ways to reduce the likelihood of making errors that exemplify the bias in addition to make sure that the potentially effective behavior generalizes between exemplars and remains in the participants' repertories after the completion of the experiment?

In order to explain the trap, numerous writers have got stuck in yet another trap of mentalistic circular explanations. I have argued that this is a rather dangerous one that researchers need to avoid in order to successfully understand, control, and manipulate such behavior, in the same way as any other socially significant types of behavior. The phenomena referred to as confirmation bias were first studied in cognitive social psychology (e.g., Wason, 1960, 1968; McKenzie, 2006; Cook & Smallman, 2008; Hernandez & Preston, 2012) and behavior-analytic treatments of the relevant phenomena have been sparse. With its focus on environmental variables that can be changed, a behavior-analytic approach is promising both with respect to a theoretically deeper understanding of such phenomena and with respect to the innovation of effective debiasing techniques.

References

Bacon, F. (n.d.). *Novum organum (New Method)*. Retrived from file:///C:/Users/Prez/Downloads/novum_organum_(new_method)_by_francis_bacon%20(1).pdf (Original work published in 1620).

Baer, D. (2016, November 9). The "Filter Bubble" explains why Trump won and you didn't see it coming. *New York Magazine*. Retrieved from http://nymag.com/scienceofus/2016/11/how-facebook-and-the-filter-bubble-pushed-trump-to-victory.html

Catania, A. C. (1998). *Learning* (4th edition). Upper Saddle River, NJ: Prentice Hall.

Cook, M. & Smallman, H. S. (2008). Human factors of confirmation bias in intelligence analysis: Decision support from graphical evidence landscapes. *Human Factors, 50(5),* 745-754. doi: 10.1518/001872008X354183

Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., & Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress? *Psychological Review, 2,* 216-229.

Gigerenzer, G. (1991). How to Make Cognitive Illusions Disappear: Beyond "Heuristics and Biases". In W. Stroebe & M. Hewstone (Eds.), *European Review of Social Psychology* (Vol. 2, pp. 83-115). Chichester, England: Wiley.

Hayes, S. C. & Brownstein, A. J. (1986). Mentalism, behavior-behavior relations, and a behavior-analytic view of the purposes of science. *The Behavior Analyst, 9(2),* 175-190.

Hernandez, I. & Preston, J. L. (2012). Disfluency disrupts the confirmation bias. *Journal of Experimental Social Psychology, 49(1),* 178-182.

Holth, P. (2001). The persistence of category mistakes in psychology. *Behavior and Philosophy, 29*, 203-219.

JJ Hoksnes. (2010, January 13). Confirmation bias: Verstingen blant tankefellene. [Web log post]. Retrieved from http://www.psykologibloggen.no/?p=2006

Kahneman, D. (2011). *Thinking, fast and slow*. London: Penguin Books.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 2*, 211-228.

Lidén, M., Gräns, M., & Juslin, P. (2018). The Presumption of Guilt in Suspect Interrogations: Apprehension as a Trigger of Confirmation Bias and Debiasing Techniques. *Law and Human Behavior, 42(4)*, 336-354.

Lilienfeld, S. O. & Landfield, K. (2008). Science and pseudoscience in law enforcement: A user-friendly primer. *Criminal Justice and Behavior, 35(10)*, 1215-1230. doi: 10.1177/0093854808321526

McKenzie, C. R. M. (2006). Increased sensitivity to differentially diagnostic answers using familiar materials: Implications for confirmation bias. *Memory and Cognition, 34(3)*, 577-588.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: an experimental study of scientific inference. *Quarterly Journal of Experimental Psychology, 29*, 85-95.

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology, 2*, 175-220.

Rassin, E., Eerland, A., & Kuijpers, I. (2010). Let's find the evidence: An analogue study of confirmation bias in criminal investigations. *Journal of Investigative Psychology and Offender Profiling, 7*, 231-246. doi: 10.1002/jip.126

Skinner, B. F. (1969). *Contingencies of reinforcement: A theoretical analysis*. Englewood Cliffs, NJ: Prentice-Hall.

Skinner, B. F. (1974). *About behaviorism*. Oxford, England: Alfred A. Knopf.

Thornton, S. (2018). Karl Popper. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2018 ed.). Retrieved from https://plato.stanford.edu/entries/popper/#ScieKnowHistPred

Wason, P. (1960). On the failure to eliminate hypothesis in a conceptual task. *Quarterly Journal of Experimental Psychology, 12,* 129-140.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20,* 273-281.

Wason, P. C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *Journal of Experimental Psychology, 23,* 63-71.

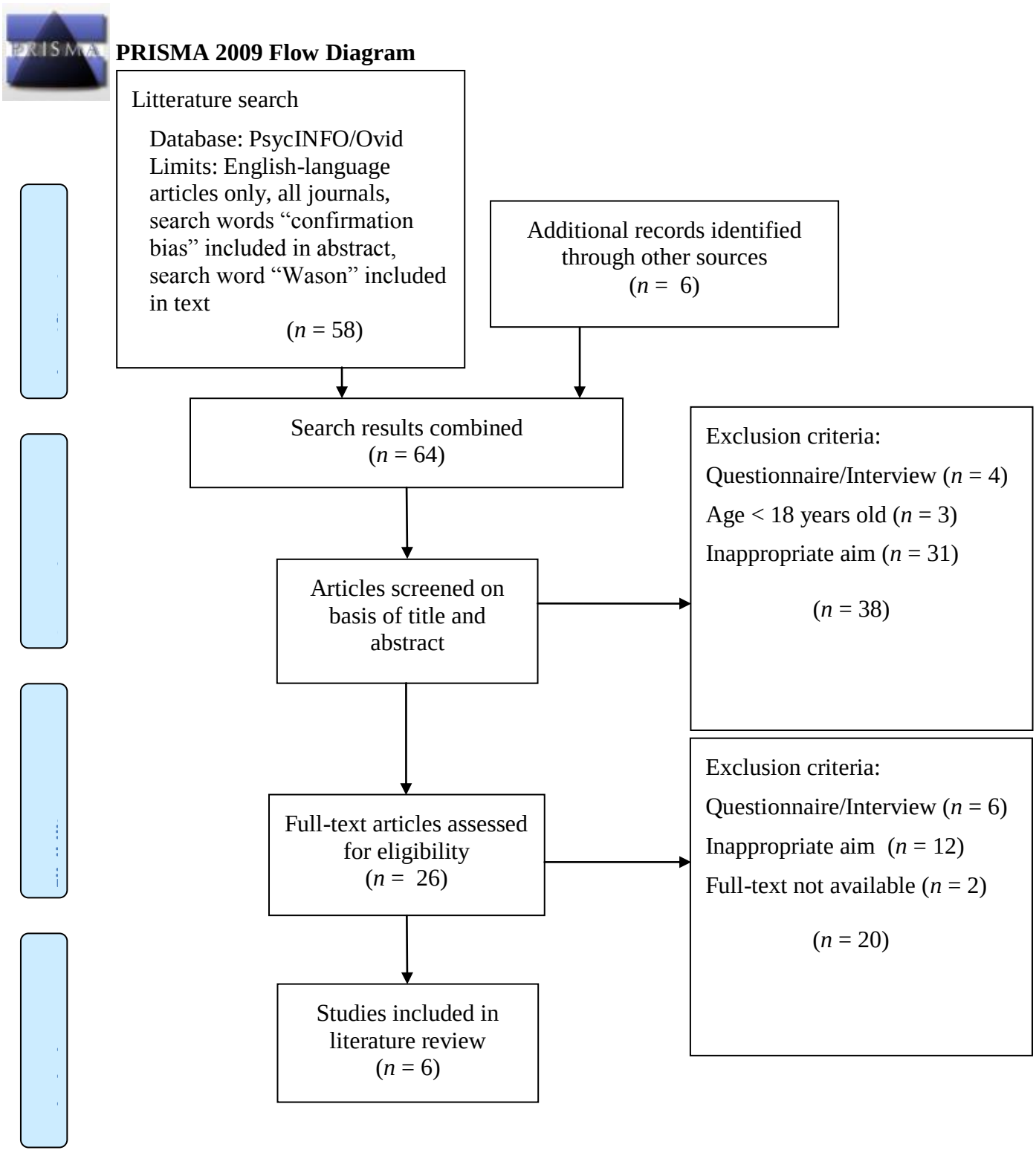


Figure 1. The literature search is presented in this PRISMA 2009 flow diagram.

(Retrieved from http://prisma.thetacollaborative.ca)

Confirmation Bias: Attempts at Debiasing Through Two Types of Training

Elise Holth

Oslo Metropolitan University

Abstract

The purpose of the present experiment was to investigate to what degree two types of training would function as debiasing techniques, reducing the probability of a response pattern exemplifying the confirmation bias. Additionally, the experimenter studied to what extent this skills would generalize to novel exemplars. Twenty-two undergraduates served as participants. The experiment was planned as a mixed N=1 and group design. The main results was that both types of training did improve the performance to a large extent in 13 out of 22 participants (two participants disqualified for further participation after demonstraded skills in a pre-experimental test), but results revealed little difference with respect to the two types of training. Moreover, the learned skills generalized to novel exemplars in 11 out of the 20 participants. The results are important with respect to the innovation of more effective debiasing techniques.

Key words: confirmation bias, explanatory feedback, correct/incorrect feedback, debiasing, behavior analysis

Confirmation Bias: Attempts at Debiasing Through Two Types of Training

The phenomenon of 'confirmation bias' is commonly defined among psychologists as the tendency to favor information compatible with an already stated hypothesis or belief (e.g., Nickerson, 1998). In November, 2018, a frigate owned by the Royal Norwegian Navy collided with an oncoming tanker. On the recent occasion of the Accident Investigation Board Norway's (AIBN) presentation of the first part of their report of the accident, this has been all over the Norwegian news. One of the conclusions in the report was that the accident was a result of errors exemplifying the confirmation bias (Stensvold & Ramsdal, 2019). Apparently, the young officers on board trusted their initial hypothesis that the oncoming ship was a stationary object, and failed to consider disconfirmatory clues before it was too late. Another contemporary example illustraive of the focus on the phenomeneon, was one in Barack Obama's (2017) presidential farewell speech. He put his audience on the alert regarding how social media may threaten democracy:

> For too many of us, it's become safer to retreat into our own bubbles, whether in our neighborhoods or college campuses or places of worship or our social media feeds, surrounded by people who look like us and share the same political outlook and never challenge our assumptions. … increasingly, we become so secure in our bubbles that we accept only information, whether true or not, that fits our opinions, instead of basing our opinions on the evidence that's out there (para. 31).

A person who believes in astrology might experience some predictions in a horoscope to come true, to a larger or smaller degree. There are wide limits to what a person can interpret in such ways that it fits with the predictions of so-called psychics, if the person really looks for it and ignores instances that do not fit the predictions. For example, the philosopher Francis Bacon (n.d.) wrote that it is human to search only for instances which confirm an already stated opinion. Further, he wrote: "And such is the way of all superstition, whether in

astrology, dreams, omens, divine jugdments … wherein men … mark the events where they are fulfilled, but where they fail, though this happen much oftener, neglect and pass them by." (pp. 9-10). Bacon also drew parallels to scientists' patterns of decision making.

Thus, the phenomenon of confirmation bias has been recognized in the literature for hundreds of years. In recent literature, the phenomenon has been discussed in different guises and in relation to numerous different areas. For example, there has been a large number of discussions of how the biggest social media pages may have affected recent political events. An advanced algorithm makes sure that information presented through an indidual's Facebook feed is news that is in accord with this individual's interests and beliefs, based on their prior hits and searches (e.g., Olson, 2016). This means that different individuals may be presented with completely different information, regardless of whether their location, nationality, or profession is the same. Schmidt et al. (2017) analyzed the news consumption of 376 million Facebook users over almost six years. They found that as a person's general Facebook activity increased, the number of news sources in focus decreased with respect to that same person. Schmidt et al. stressed that Facebook and other social media "bubbles" have proven to strenghten user polarization. Probably, a majority of humans are prone to favor confirmation over disconfirmation. This pattern is likely to be more prominent with the use of social media, which is programmed to work by presenting more information compatible with a user's history of clicks and likes and less information which challenges that particular person's views. Olson (2016) pointed out that this sort of allgorithms or 'filter bubbles' may have contributed to both the U.K. *Brexit* and Trump winning the U.S. election.

Several researchers have studied behavioral patterns that exemplify the confirmation bias in relation to law enforcement. In their book named "Mistakes Were Made (But Not By Me)", Tavris and Aronson (2007) argued that the process of profiling suspects has a significant risk of errors in that when looking for characteristics in match with a profile, the

profiler might ignore characteristics that do not. Experimenters have demonstrated the formation of verdicts before sufficient evidence is obtained (e.g., Rassin, Eerland, & Kuijpers, 2010; Lidén, Gräns, & Juslin, 2018). Lilienfeld and Landfield (2008) argued that the process of conviction carries a risk of confirmation of beliefs becoming more important than conviction of the actual guilty person, and that such a bias can have fatal consequences, litterally. For example if a person is wrongfully convicted, sentenced to death, or in a case of crises where officers make bad decisions. To fulfil the law stating that anybody should be considered innocent until proven guilty, police officers need training with respect to the avoidance of behavior exemplifying the confirmation bias. Some debiasing techniques have been proposed and studies in this area, such as alternation of interrogation mode (Lidén et al., 2018). The researchers found that interrogation by using questions from predetermined lists reduced the occurrence of bias, compared to free generation of questions. Rassin et al. (2010) also pointed out that people probably favor confirmation over disconfirmation with respect to hypotheses presented to them, even in cases where they had no prior opinion made up.

The first to refer to the phenomenon as *confirmation bias*, psychologist, Peter Cathcard Wason (1960) carried out numerous experiments in the 60's and 70's which demonstrated the phenomenon. First, he conducted an experiment to investigate whether people paid attention both to confirming and to disconfirming evidence, or simply to confirming evidence, in the process of concluding about small or elementary conceptual tasks. Participants were presented with a triplet (2-4-6), and informed that the experimenter had formed a rule to which the numbers conformed. Furthermore, partipants were instructed to attempt to reveal the rule by guessing other triplets that they thought would fit the rule. Results demonstrated the phenomenon of confirmation bias in that people tended to come up with a hypothesis early on, and only to test triplets that would confirm their hypothesis, and not to test any instances that could disconfirm the hypothesis. Later on, Wason (1966, as

referred in Wason, 1968 and in Nickerson, 1998) studied confirmation bias presenting his quite famous card *selection task*. The task typically involved four cards with one type of letter, number, or other symbol on one side, and another type of symbol on the other. Participants were asked to point out which card or cards they would need to turn around to see if a particular rule presented by the experimenter was true. For example, the participants were shown four cards, displaying *A* (P), *D* ($\overline{P}$), *4* (Q), and *7* ($\bar{Q}$), respectively, and told that every card had a letter on one side, and a number on the other. Next they were asked to tell which of the four cards would have to be turned over in order to decide whether the following rule was true: "If there is a vowel on one side, there is an even number on the other". In order to solve the problem correctly, the P- and the $\bar{Q}$-cards had to be inspected. The P-card had to have a Q (confirmation) on the other side, and the $\bar{Q}$-card had to have anything but a P (disconfirmation) on the other side. Additionally, any symbol on the other side of both the Q-card and the $\overline{P}$-card would be irrelevant with respect to confirming the rule. These experiments repeatedly resulted in an overweight of selection of the P-card. The second most common response was to turn over the P-card combined with the Q-card. Only a few participants appeared to pay attention to either the $\overline{P}$-card or the $\bar{Q}$-card. Interestingly, those cards selected by participants, almost exclusively, were those which corresponded to the symbols mentioned in the rule (e.g., Wason & Shapiro, 1971; Nickerson, 1998). The selection task is compatible with Taleb's (2007) *Black Swan*-example. In 1697, an European explorer travelled to Austrailia and observed black swans. Before this, children all over Europe were taught that all swans were white. The example illustrates that a statement like "All swans are white" can be intrepreted as a truth until a negative example, in this case a black swan, is observed.

Wason and Shapiro (1971) carried out another version of the four card-experiment, in which participants were randomly allocated into either a thematic or an abstract group.

Participants in the abstract group were presented with four cards at a time, with, for example, *D*, *K*, *3*, and *7* on them, respectively, in addition to the rule "If there is a D on one side of a card, it has 3 on the other". They were then instructed to determine which card(s) they had to turn over in order to confirm or falsify the rule. Furtermore, participants in the thematic group were presented with four cards, with, for example, *Manchester*, *Leeds*, *Car*, and *Train* on them, respectively. Additionally, each card was marked with the name of a different day of the week. The rule presented with this task was *Every time I go to Manchester, I travel by car*, and participants were instructed to determine which card(s) that had to be turned over in order to confirm or falsify the rule. Results revealed that participants in the thematic group solved the tasks in 62.5 % of the cases. On the other hand, participants in the abstract group solved the tasks in only 12.5 % of the cases. Better performance with respect to thematic tasks compared to abstract tasks has also been demonstrated by other researchers (e.g., McKenzie, 2006). Wason and Shapiro (1971) provided three possible hypotheses which might "account for" this better performance. One hypothesis was that naming each card by the name of a different day of the week made it less likely for the participants to focus on those cards with correspondence to the terms mentioned in the rule. A second hypothesis was that the thematic relationship between the concepts, rather than the thematic concepts themselves, made the tasks easier to solve. Finally, most relevant for the present experiment, Wason and Shapiro argued that thematic concepts might have led participants to more symbolic manipulations of the concepts. They suggested that this hypothesis could be tested by presenting a version of the selection task containing familiar concepts, but with no familiar connection, and exemplified the suggestion by presenting the name of a metal on one side of a card and the name of a fruit on the other, combined with a rule that said "*Every card which has iron on one side has apple on the other side*" (Wason & Shapiro, 1971, p. 69).

Catania (1998) discussed the phenomenon using the term ‘confirmatiory bias’. He briefly reviewed Wason’s selection task and pointed out that turning over a card would be an example of an *observing response*. His definition of such responses was behavior that leads to the presence of a discriminative stimulus. From a behavior-analytic point of view, he argued that confirmation is more reinforcing than disconfirmation, which is tantammount to the statement that people are more likely to produce information that is compatible with their hypotheses, than they are to seek disconformatory information.

Dinsmoor, Browne, and Lawrence (1972) investigated this in his laboratory experiments with pigeons. First, they arranged for the pigeons to learn to peck on a key to produce food. Further, he arranged the conditions so that pigeons’ peck on a second key led to either green or red light on that key. In the presence of green light, pecks on the first key led to the delivery of food, while pecks in the presence of red light did not. In other words, the pigeons pecked for confirmation and disconfirmation of the effectiveness of pecking on the first key. At this point, it was not clear whether it was information in addition to something else or simply information by itself that maintained pecking on the second key. Dinsmoor et al. then altered the conditions so that pecks on the second key either led to the presence of green light (i.e., pecks on the first key would produce food) or had no consequences (i.e., pecks on the first key would have no consequences). This green light is an example of a discriminative stimulus in the sense that pecks on the first key are effective only in its presence. They observed no changes in the pigeons’ behavior under this condition in comparison to the first. Finally, the experimenter arranged for pecks on the second key to lead to either the presence of red light, indicating that pecks on the first key would have no consequences, or no light at all, indicating that pecks on the first key would produce food. Under such conditions, the pigeons’ rate of pecks on the second key fell drastically. Dinsmoor et al. concluded that when pecks on the second key led to the presence of either red light or no

light, pecks producing the red light were punished (i.e., red light had an aversive effect in that the production of food was not possible), or, alternatively, that pecks on the second key were extinguished (i.e., the presence of green light was a reinforcer [a consequence leading to increased probability of repetition of behavior provoking it] that stopped occuring and lead to the extinction of pecks on the second key). Catania (1998) summarized these findings, pointing out that organisms do not work for information as such, but primarily for information correlated with reinforcement. He pointed out that this is simply another fact about behavior stemming from the science of behavior analysis. Catania added that this fact is sometimes resisted and that "[w]hen that happens, it's a fact that illustrates itself." (Catania, 1998, p. 183). In their experiments, Dinsmoor et al. demonstrated just that: the pigeons pecked a key for information confirmative of food accessibility and ignored information disconfirmative of it.

Researchers have demonstrated that people's performance improves on tasks with thematic terms and relations, compared with their performance on abstract tasks. The avoidance of errors exemplifying the confirmation bias would probably have a favorable effect in many arenas. Hence, further research aimed at discovering how to teach people to avoid making such errors is necessary, also with respect to abstract tasks. The present experiment was based on Wason's selection task. Multiple occurences of varying stimuli, and recordings of whether the relevant response occurs or not are necessary for the assessment of a controlling relation (Sidman, 1979). Ten out of 15 different tasks which were presented (and some of them repeatedly) to the participants were designed in accordance with Wason and Shapiro's (1971) third suggestion (familiar concepts with abstract relations), two tasks contained familiar terms in addition to familiar relations, whereas the last three consisted of a novel version of the task with eight cards, with familiar concepts with unfamiliar relations. The experiment was conducted in order to investigate to what extent the likelihood of making

errors that exemplify the confirmation bias can be reduced through two types of training, namely training with explanatory feedback and training with correct/incorrect feedback. Also, the researcher aimed to investigate to what extent the effect of training might generalize to novel eight-card tasks. Moreover, the experimenter sought to investigate whether participants who already mastered the four-card task would also solve the eight-card task and thus, demonstrate that their skills were not just automatized responses controlled by characteristics of the four-card task.

Method

Participants

Four child welfare undergraduates, five psychology undergraduates, and 13 learning disability nurse undergraduates participated in the experiment, 15 female and 7 male. All of them were students at Oslo Metropolitan University. Their reported age varied from 19-20 to 51-55 years old. The experimenter showed up in the students' classrooms and asked if they would like to participate in an experiment by answering some tasks that would take approxipately 30 min. They were informed of times and room number for participation. During recruitment, no detailed information was presented regarding the content or purpose of the experiment.

Apparatus

The experiment was carried out using three Acer Aspire Switch computers with 1920 x 1080 pixel, 10" screens, and one Lenovo Yoga 500-14IBD computer with 1920 x 1080 pixel, 14" screen. All of the computers had a touch screen. The custom made program that controlled the experiment and the data retrieval were written using Microsoft Visual Basic 1.0 (rev. 141, 2010 Express).

Setting

The experiment took place in groop rooms at Oslo Metropolitan University, campus Sandvika and campus Kjeller. Participants, up to four at a time, were sitting around a table, unable to look at each others screens.

Design

A mixed N=1 and group design was planned in the experiment. If the participants had completed five or less correct responses after the second baseline test, they were exposed to two types of training. However, if they had more than five correct responses at that point, they were sent directly to the generalization test. The first group received explanatory feedback during training immediately after the baseline test and correct/incorrect feedback during the second training phase. Participants in the second group were exposed to identical tasks, but they received correct/incorrect feedback in the first training phase and (potentially) explanatory feedback in the second. The dependent variable was the selected participants' patterns of responding on Wason-type card selection tasks. The main independent variable of interest was the training, consisting of either specific explanatory feedback or general feedback ("correct"/"incorrect"), during the experiment. The generalization measure was the extent to which the selected participants produced correct response combinations on a repetition of the baseline test following training and in the final generalization test, consisting of eight response alternatives for each task.

Procedure

Participants were given six conceptual tasks consisting of some information, including a rule, and four (or eight) different cards with a word, a symbol, an age, a letter, a picture, or a number on each side of them. Only one side of the cards was visible to the participants. All tasks involved for participants to select which card(s) had to be turned over in order for them to be certain if the rule was true or false. The response alternatives (i.e., the cards) appeared in

randomized orders. The participants selected cards by clicking on them, before clicking "Submit answer and continue to the next task" on the screen. Then the next task was presented immediately. Instructions were presented prior to all phases, except the *Generalization test*. A time limitation of 3 min per task was pre-programmed for all tasks, except for the first one in the *Generalization test*, where the limit was set to 5 min. In all cases, the combined selection of the P-card and the $\bar{Q}$-card(s) were considered the correct responses (see Table 1 for an overview), and those cards had to be selected, unless time ran out, in order to continue to the next task during the training phases. Latencies were recorded during all instructions and trials throughout the experiment. The tasks were presented in Norwegian.

Baseline. The first six tasks functioned both as a baseline and a pre-experimental selection of participants, which were conducted to separate the students who had a learning history to solve the relevant kinds of tasks correctly from those who had not. Students who responded correctly to four or more of the tasks in a row disqualified for further participation, whereas the students who failed on three or more tasks qualified. The first task presented in this phase, following the instruction, contained the following information: "All four cards have a sign on one side and a color on the other side", and the following rule: "If there is a circle on one side of the card, then the other side says *red*", together with four cards showing a circle (P), a triangle ($\bar{P}$) and the words *red* (Q) and *green* ($\bar{Q}$), respectively. In the second task, the information given was: "All four cards have a car on one side and a drinking tool on the other side", and rule was: "If there is a Volvo on one side of the card, then there is a glass on the other side". The cards presented respectively had a Volvo (P), a Toyota ($\bar{P}$), a cup ($\bar{Q}$), and a glass (Q) on them. The third task was one of the tasks made and used by Wason (1966, as referred in Wason, 1968), with the information: "All four cards have a letter on one side and a number on the other side", and the rule: "If there is a vowel on one side of the card, then

there will be an even number on the other side", together with four cards with, respectively, *A* (P), *D* ($\overline{P}$), *4* (Q), and *7* ($\bar{Q}$) on them. The fourth task contained the description: "All four cards have a name of a mountain on one side and a gender sign on the other side", followed by the rule: "If one side of a card shows *Galdhøpiggen*, then the other side will show the sign for men (♂)", together with four cards with *Galdhøpiggen* (P), *Glittertind* ($\overline{P}$), ♂ (Q), and ♀ ($\bar{Q}$) on them, respectively. In the fifth task, participants were presented for the information: "All four cards have the name of a travel document on one side and an age on the other side". They were given the rule: "If a card has *Tram ticket* on one side, then it has *15 years old* on the other side". The cards presented had *Tram ticket* (P), *Driver's license* ($\overline{P}$), *15 years old* (Q), and *20 years old* ($\bar{Q}$) on them. Finally, in the sixth task, the information presented was: "All four cards have the name of an ape species on one side and the name of a bird species on the other side", while the rule was: "If a card has *Gorilla* on one side, it has *Parakeet* on the other side. The cards had *Gorilla* (P), *Chimpanzee* ($\overline{P}$), *Parakeet* (Q), and *Macaw* ($\bar{Q}$) on them. The P's, Q's, $\overline{P}$'s, and $\bar{Q}$'s appeared in different sequences in the different tasks. No feedback was given during this first phase. The baseline was a systematic replication of Wason's (1966, as referred in Wason, 1968) experiment (apparatus and terms was switched out and a time limit added).

Explanatory-feedback-first condition. Qualified participants (those who got three or more tasks wrong) were allocated into either one of the groups, randomly. To guarantee random allocation, 11 red and 11 black cards were put together and reshuffled, then drawn one by one. Red cards represented first group (explanatory feedback) and black cards represented the second group (correct/incorrect feedback). Results were noted in order next to each participant number.

Training with explanatory feedback. In the following trials, participants received explanatory feedback for every submission, stating why the selection was correct or incorrect.

For P-cards, the feedback was: “If Q does not appear on the other side of this card, the rule is false. Therefore, it is correct to turn this card over.” A specific example was: “If the other side of this card does not say *We offer weekly magazines*, then the rule is false. Therefore it is correct to turn this card over. The explanatory feedback for Q-cards was the following: “The rule does not say that P has to appear on the other side of Q. Thus, this card need not be turned over.”, for example: “The rule does not say that other stores cannot offer weekly magazines. Thus, what appears on the other side of this card is irrelevant; turning over this card is unnecessary”. For $\overline{P}$-cards, the feedback provided was: “The rule does not require anything in particular to be found on the other side of this card. Therefore, it does not have to be turned over.”, for example: “The rule does not tell us that Joker is the only store to offer weekly magazines. Therefore, nothing appearing on the other side of this card can be relevant to find out whether the rule is true or false, and it needs no turning over. Participants received the following feedback in relation to $\bar{Q}$-cards: “If P appears on the other side of this card, the rule would be falsified. Therefore, it is correct to turn this card over.”, for instance: “If the invisible side of this card says *Joker*, it means that the rule is false. Therefore, this card has to be turned over”.

Correct selection led directly to the presentation of the next task when pressing “Submit and continue to the next task”, while pressing after an incorrect selection led to the appearance of explanatory feedback. Either way, participants had to select the correct cards to continue to the next task.

In the first of these tasks, the information appearing on the screen was that “All four cards have a grocery store name on one side and information about whether or not weekly magazines are offered on the other side”, and the rule as follows: “If there is a Joker sign on one side of the card, then the other side says ‘We’re offering weekly magazines’ on the other side”. Four cards with a Joker sign (P), a REMA 1000 sign ($\overline{P}$), “We’re offering weekly

magazines" (Q), and "We're not offering weekly magazines" ($\bar{Q}$) appeared respectively. In the second task, the information provided was that "All four cards have a picture of an instrument on one side and the name of a fruit of the other side", together with the rule: "If there is a picture of a guitar on one side of the card, then the other side shows *Apple*". The four cards attached to this task had a picture of a guitar (P), a picture of a trumpet ($\bar{P}$), *Apple* (Q), and *Pear* ($\bar{Q}$) on them, respectively. The third task involved that "All four cards have a flower name on one side and a picture of an animal on the other side", and the rule: "If one side of the card shows *Poppy*, then the other side will show a picture of an elephant". Four cards were presented with, respectively, *Poppy* (P), *Tulip* ($\bar{P}$), a picture of an elephant (Q), and a picture of a rabbit ($\bar{Q}$) on them. In the fourth task, participants were enlightened that: "All four cards have the name of a profession on one side and a dish on the other side", followed by the rule: "If one side of the card says *Carpenter*, then the other side says *Fish balls*", and four cards with the text *Carpenter* (P), *Police* ($\bar{P}$), *Fish balls* (Q), and *Spaghetti* ($\bar{Q}$), respectively. Fifth, the information given was that "All cards have a famous painting on one side and a city name on the other side". The rule presented was: "If there is a copy of The Scream on one side, then it says *New York* on the other side". The four cards that appeared had a copy of The Scream (P), a copy of Madonna ($\bar{P}$), *New York* (Q), and *Moskva* ($\bar{Q}$) on them, respectively. Finally, the sixth task presented, contained the information: "All four cards have the name of a beverage on one side and an age on the other side", while the rule was: "If a card has Beer on one side, it has 22 years old on the other side". The cards had Beer (P), Coke ($\bar{P}$), 22 years old (Q), and 16 years old ($\bar{Q}$) on them.

Test. The third phase consisted of a repetition of the baseline test tasks. Participants were instructed that they would not receive feedback.

Training with correct/incorrect feedback. Participants with more than five correct responses prior to this phase skipped this phase and the third baseline test, and were presented

with the generalization test directly after the second baseline test. Only those participants who had five or fewer correct responses at this point (during the three first phases) were exposed to a second type of training. The tasks from the first training phase were repeated. However, this time the participants received simple feedback, stating only whether the selected cards were correct or incorrect. The participants had to select the correct cards to continue to the next task before time ran out.

Post-test. The six tasks from the baseline test were once again repeated.

Generalization test. As a final test, participants were presented to three tasks containing eight response alternatives. This phase did not include any instructions or feedback. The first task contained the information: "All cards have the name of a band on one side and a picture of a fruit on the other side", accompanied by the rule: "If a card has a picture of an apple on one side, then it says Queen on the other side". This task involved the following eight response alternatives: a picture of an apple (P), a picture of a pear ($\bar{P}$), a picture of a grape ($\bar{P}$), a picture of a mango ($\bar{P}$), Queen (Q), The Rolling Stones ($\bar{Q}$), The Beatles ($\bar{Q}$), and ABBA ($\bar{Q}$). In the second task, participants were informed that "All eight cards have the name of a game on one side and the name of a diagnosis on the other side" and with the rule: "If one side of a card says *Yatzi*, then it says *Bipolar disorder* on the other side." The eight cards presented had the following names on them, respectively: *Yatzi* (P), *Chess* ($\bar{P}$), *Monopoly (*$\bar{P}$), *Snakes & Ladders* ($\bar{P}$), *Bipolar disorder* (Q), *ADHD* ($\bar{Q}$), *Autism* ($\bar{Q}$), and Schizophrenia ($\bar{Q}$). At last, the third task presented the information: "All cards have a planet name on one side and the name of a country on the other", followed by the rule: "If it says *Saturn* on one side of a card, then it says *Italy* on the other side". The eight cards appearing on the screen had *Saturn* (P), *Mars* ($\bar{P}$), *Uranus* ($\bar{P}$), *Neptune* ($\bar{P}$), *Italy* (Q), *Ghana* ($\bar{Q}$), *Peru* ($\bar{Q}$), and *Mongolia* ($\bar{Q}$).

Correct/incorrect-feedback-first condition. Participants in the second group were exposed to conditions in the opposite order with respect to the provided feedback: Under the first set of training tasks, they received feedback in a nonspecific form, stating only whether the selected card was correct or incorrect, whereas they received explanatory feedback in the second round, unless they had submitted more than five correct responses during the first three phases. If so, they were presented with the generalization test immediately after the second baseline test.

Results

Two participants responded correctly on four or more of six tasks in a row during the first baseline test, hence, they disqualified for further participation. Both participants also responded correctly to all of the three generalization tasks (See Figure 1 for a representative example).

During the first baseline test, the 20 qualified participants responded to a total of 120 tasks, and selected only the P-card in response to 24,16 % of them (29 times), the PQ-combination in 35 % of them (42 times), and the correct combination of P$\bar{Q}$ in 0.05 % of them (six times).

Figure 2 demonstrates that both the explanatory feedback and the correct/incorrect feedback improved performance up to 72 % correct selection during the second baseline test, from 12 % and 0.2 %, respectively, during the first baseline test. In sum, participants exposed to the explanatory feedback responded correctly on 53 % of the generalization tasks, and participants exposed to the correct/incorrect feedback responded correctly on 50 % of the generalization tasks. However, nine participants responded incorrectly to all of these generalization tasks, while two participants responded correctly on 66.67 % of them and 11participants (including the two disqualified participants) responded correctly on 100 % of them.

Eleven participants, six presented with explanatory feedback and five presented with correct/incorrect feedback, responded correctly to at least five out of six tasks during the second baseline test, and responded correctly to at least two out of three tasks during the generalization test. A representative example of these participants' response patterns is shown in Figure 3. The six participants who were exposed to explanatory feedback responded incorrectly to 4.8 tasks (including repeated errors on the same tasks), on average, during training. The five participants who were exposed to correct/incorrect feedback responded incorrectly to 13.4 tasks, on average, during training.

Two participants, one from each feedback group, had five or less correct responses when the second baseline test started and were therefore exposed to the second type of training (see Figure 4 for a representative example). None of these participants responded correctly on any of the eight-card generalization tasks.

Seven participants had 100 % incorrect responses during the generalization test, but two of these had five or more correct responses during the second baseline test (See Figure 5 for a representative example). Those latter two participants were presented with the correct/incorrect feedback. One of them selected the P-card combined with one of the $\bar{Q}$-cards in all of the three generalization tasks, whereas the other selected seven cards, excluding one $\overline{P}$-card, in each task. The other five participants, three exposed to explanatory feedback and two exposed to correct/incorrect feedback, with no correct responses during the generalization test had few correct responses (from six to 11) through the whole experiment (See Figure 6 for a representative examples).

Instruction times, that is, the time from the instruction was presented until the participant touched the "Submit answer and continue to the next task" button, prior to the first baseline tasks varied between 16s and 235s across participants. Prior to the next block, the first teaching trials, instruction times varied from 10s to 62s. In general, latencies were higher

on, or shortly prior to the first trials with correct responses, and when new tasks were introduced. For example, participant #16 (see Figure 5) showed increased latencies on the first correct trials during feedback and during the following repeated baseline trials. Participant #20 and 22 (see Figure 3) exemplified the pattern of increased latencies on trials just prior to the first correct responses during the feedback phase. Finally, participant #13 (see Figure 4) showed increased latencies on some of the few trials with correct responses, following failures, during the first feedback phase. Initially increased latencies were sometimes followed by consistently incorrect responding, as during the generalization test for partcipant #16 (see Figure 5).

Discussion

The present experiment demonstrated a clear reduction in the number of errors that exemplify the confirmation bias, as a result of exposure both to the explanatory feedback and to the correct/incorrect feedback. The performance on the second baseline test was exactly the same in both groups, indicating that the two types of training did not differ with respect to the number of correct responses. Differences in generalized task performance were also close to non-existent, corroborating that the two debiasing techniques did not differ in sum to any considerable degree with respect to their influence on correct responding. However, based on the performances by the 11 participants who responded correctly both to more than four tasks during the second baseline test and two or more tasks during the generalization test, the number of exemplars necessary to produce correct responses during training was somewhat larger in the correct/incorrect feedback group than in the explanatory feedback group.

With respect to response combination alternatives during the eight-card generalization tasks, each card constituted two possibilities–selection and no selection–which implies 2^8=256 possibilities. Because participants were instructed to select at least one card, the select-no-card option did not exist, and 255 options remained. Thus, the possibility of

responding correctly on two of the generalization tasks in a row just by chance was one in 65,025 (255^2). Therefore, it seems safe to conclude that participants who responded correctly in this way had the appropriate skills for solving these tasks already established in their repertoire. Thus, the correct performances generalized to the eight-card task in more than half of the participants.

Evidently, the two participants who responded correctly to at least five of six tasks during the second baseline test, but did not respond correctly to any tasks during the generalization test demonstrated that their pattern of correct responding with respect to the original four-card tasks did not completely comply with the generalized correct pattern of turning over P and all $\bar{Q}$s. The correct response pattern in the four-card tasks ($P\bar{Q}$-combination, responding only to one $\bar{Q}$) seen in one of these participants during the generalization test substantiates this interpretation, whereas the other participant's response pattern (selection of all response alternatives, except one $\overline{P}$-card) may be properly labeled incomprehension. It should be noted that both of these participants were presented with the correct/incorrect feedback, suggesting the possibility that explanatory feedback might lead to a higher probability of establishing generalized skills. Moreover, Sidman (1979) pointed out that in order to be certain of what is actually controlling a change in a particular type of behavior, the observation of several instances of that particular responses and its preceding/coinciding stimuli are necessary. Participants who responded correctly to at least five of the six tasks during the second baseline test obviously behaved in accordance with the rule. Again, however, their pattern of correct responding may have been limited to the four-card task rather than being in accord with the more generalized rule of selecting P and *all* $\bar{Q}s$. The generalization task was a way of testing for this possibility.

The recorded duration of instructions were, with a few exceptions, as long as or longer than the time expected on the basis of normal reading speed–suggesting that for the most part

the instructions were read by the participants. The fact that latencies on test trials were often higher initially when new tasks were introduced as well as on, or just before the first trials with correct responses suggests that some kind of problem solving occurred. In a few cases, participants showed increased latencies followed by consistently incorrect responding, or incorrect responding which persisted with very short response latencies throughout the different training and test periods. Colloquially speaking, the response patterns displayed by these participants may indicate that they "had given up". The pattern may suggest the relevance of the phenomenon referred to as learned helplessness (e.g., Seligman, 1975). As defined by Catania (1998), learned helplessness is a phenomenon in which a learning history of inescapable and unavoidable discomfort, in this case the incorrect-feedback, produce a behavioral pattern in which the organism stops escaping or avoiding similar stimuli later on. Also, according to Seligman, additional learning is hampered.

The four-card tasks involved two response alternatives to each card–selection or no selection–which implies 2^4=16 possibilities. Again, as participants were instructed to select one or more cards, only 15 options remained. Responding correctly to four out of six four-card tasks in a row just by chance is then possible in one out of 50,625 (15^4) cases, while the possibility of responding correctly to six out of six such tasks by chance is one in 11,390,625 (15^6). Hence, clearly the two participants who disqualified for further participation in the experiment had the skills of interest established in their repertoire prior to their participation in the present experiment. Additionally, both these participants responded without errors during the generalization test, suggesting that their skills extended beyond the classical selection task.

As pointed out by Catania (1998), confirmation may typically function as reinforcement. Hence, providing people with explanatory feedback, both confirmatory and disconfirmatory, will function as differential reinforcement of responding suitable for

disconfirmation and, hence, not exemplifying the confirmation bias. This may exemplify exactly what needs to be done in order to change such response patterns that show the confirmation bias.

The present findings supported the previous findings with respect to most common response combinations. Wason (1968) reported that 46.15 % of his participants selected the P- and Q-card in a selection task, 34.61 % selected the P-card only, as opposed to 0.07 % selecting the correct P and $\bar{Q}$ combination. Similarily, in the present experiment those response combinations occurred in 35 %, 24.16 %, and 0.05 % of the tasks, respectively. Regarding previous findings on the differences in performance in relation to abstract versus thematic selection tasks, results from the present experiment did not replicate them. As demonstrated by other researchers (e.g., Wason & Shapiro, 1971), participants often perform better on tasks involving familiar concepts and relations than to abstract tasks. Results in the present experiment did not replicate these differences. No systematically better performance was shown on the two tasks with thematic relations between rule and the words printed on the cards than on the other tasks. Wason and Shapiro included some more information in their rules, such as “Every time I go to Manchester, I travel by car”, as opposed to “If a card has Beer on one side, it has 22 years old on the other side”. Perhaps the absence of such information in the rules used in the present experiment account for the lack of differences between the results from the two sorts of tasks.

Moreover, the finding by Dinsmoor et al. (1972) that the pigeons stopped responding when only the negative discriminative stimulus was contingent on observing responses, indicated that disconfirmation served as a punishing stimulus (i.e., a stimulus that reduces the rate of responses upon which it is contingent). Supposing that disconfirmation had a punishing effect also on the behavior of participants in the beginning of the present experiment, the two types of training may have counteracted disconfirmation as punishing

stimuli and the explanatory and correct-feedback were likely established as reinforcers (i.e., stimuli following a type of behavior and leads to more of that behavior) for disconfirmation seeking.

An identified weakness in the present study was that the placement of the "Submit answer and continue to the next task" button was the same for each task. This presented the opportunity to double-touch the button, and, thus, skipping a task, by "accident" during the baseline tests. Another limitation was perhaps the criterion set to determine whether to present participants with a second type of training. In order for a participant to have collected less than six correct responses at that point in the experiment, they had to have ran out of time at least once. Few participants ran out of time in any tasks at all, and only two qualified for the presentation of the second type of training. Hence, the opportunities for comparing effects of the two types of feedback within subjects were very limited. Another limitation of the present experiment was that the sample of participants consisted exclusively of undergraduates recruited from a university. The general effectiveness, as well as the relative effectiveness, of the two types of "feedback"-interventions may be different with participants recruited from other groups and environments.

The confirmation bias phenomenon has not typically been researched, or even discussed, by behavior analysts. However, whenever the goal is to debias or avoid confirmation bias, strategies concerned with manipulable independent variables are necessary. Further investigations with respect to the generalization of skills and maintenance of skills over time is still missing from the research literature. An interesting variation in a future experiment could be to investigate the degree of generalization of acquired debiased skills across different tasks, for example from the kind of tasks used in the present experiment to Wason's (1960) triplets. Maintenance over time could be studied by replicating the present study with follow-up sessions progressively longer periods of time, such as after a week, a

month, or even a year. Another aim for further research is to develop effective interventions for participants with whom the two types of training investigated in the present experiment were not effective.

Both of the two types of training presented in the present study functioned as debiasing techniques, at least to some extent on a short term. However, both types of training produced similar improvements in the number of correct responses, even though more repetitions were required before correct responses occurred in the correct/incorrect feedback group than in the explanatory feedback group. Moreover, the new skills did generalize to novel eight-card tasks in approximately 50 % of the participants, independent of what type of training they had been exposed to. The disqualified participants also showed generalization of skills to the novel eight-card tasks. So far, this research has demonstrated that it is possible to facilitate people's correct solving of selection tasks. However, the next, perhaps important, goal must be to discover and implement techniques that increase the likelihood that important decisions across guises and disciplines are characterized by confirmation bias.

References

Bacon, F. (n.d.). *Novum organum (New Method)*. Retrived from file:///C:/Users/Prez/Downloads/novum_organum_(new_method)_by_francis_bacon%20(1).pdf (Original work published in 1620).

Catania, A. C. (1998). *Learning* (4th edition). Upper Saddle River, NJ: Prentice Hall.

Dinsmoor, J. A., Browne, M. P., & Lawrence, C. E. (1972). A test of the negative discriminative stimulus as a reinforcer for observing. *Journal of the Experimental Analysis of Behavior*, 18, 79-85.

Lidén, M., Gräns, M., & Juslin, P. (2018). The Presumption of Guilt in Suspect Interrogations: Apprehension as a Trigger of Confirmation Bias and Debiasing Techniques. *Law and Human Behavior, 42(4)*, 336-354.

Lilienfeld, S. O. & Landfield, K. (2008). Science and pseudoscience in law enforcement: A user-friendly primer. *Criminal Justice and Behavior, 35(10)*, 1215-1230. doi: 10.1177/0093854808321526

McKenzie, C. R. M. (2006). Increased sensitivity to differentially diagnostic answers using familiar materials: Implications for confirmation bias. *Memory and Cognition, 34(3)*, 577-588.

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology, 2*, 175-220.

Obama, B. H. (2017, January 11). Text of Obama's farewell speech as prepared for delivery. *AP News*. Retrieved from https://apnews.com/5f2a5b8bf38e4bd58852cfaee5864430

Olson, P. (2016, November 9). How Facebook helped Donald Trump become president. *Forbes*. Retrieved from https://www.forbes.com/sites/parmyolson/2016/11/09/how-facebook-helped-donald-trump-become-president/#3b6050ac59c5

Rassin, E., Eerland, A., & Kuijpers, I. (2010). Let's find the evidence: An analogue

study of confirmation bias in criminal investigations. *Journal of Investigative Psychology and Offender Profiling, 7*, 231-246. doi: 10.1002/jip.126

Schmidt, A. L., Zollo, F., Vicario, M. D., Bessi, A., Scala, A., Caldarelli, G., . . . Quattrociocchi, W. (2017). Anatomy of news consumption on Facebook. *Proceedings of the National Academy of Sciences of the USA*. https://doi.org/10.1073/pnas.1617052114

Seligman, M. E. P. (1975). *Helplessness: On Depression, Development, and Death*. San Francisco: W. H. Freeman.

Sidman, M. (1979). Remarks. *Behaviorism, 7*, 123-126.

Stensvold, T. & Ramsdal, R. (2019, November 8). Her er konklusjonene fra Havarikommisjonens rapport om KNM Helge Ingstad. *NTB*. Retrieved from https://www.tu.no/artikler/her-er-konklusjonene-fra-havarikommisjonens-rapport-om-knm-helge-ingstad/478525

Taleb, N. N. (2007). *The Black Swan: The Impact of the Highly Improbable*. New York: Random House. Abstract retrieved from https://www.economist.com/media/globalexecutive/black_swan_taleb_e.pdf

Tavris, C., & Aronson, E. (2007). *Mistakes Were Made (But Not By Me): Why We Justify Foolish Beliefs, Bad Decisions, and Hurtful Acts* (1st ed.). Orlando, Fla.: Harcourt.

Wason, P. (1960). On the failure to eliminate hypothesis in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*, 129-140.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*, 273-281.

Wason, P. C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *Journal of Experimental Psychology, 23*, 63-71.

Table 1

Response Alternatives

P	**Q**	**$\bar{P}$**	**$\bar{Q}$**
Circle	Red	Triangle	Green
Volvo	Glass	Toyota	Cup
A	4	D	7
Galdhøpiggen	♂	Glittertind	♀
Tram ticket	15 years old	Driver's lisense	20 years old
Gorilla	Parakeet	Chimpanzee	Macaw

P	**Q**	**$\bar{P}$**	**$\bar{Q}$**
Joker	Offering magazines	REMA 1000	Not offering magazines
Guitar	Apple	Trumpet	Pear
Poppy	Elephant	Tulip	Rabbit
Carpenter	Fish balls	Police	Spaghetti
The Scream	New York	Madonna	Moskva
Beer	22 years old	Coke	16 years old

P	**Q**	**$\bar{P}$**	**$\bar{P}$**	**$\bar{P}$**	**$\bar{Q}$**	**$\bar{Q}$**	**$\bar{Q}$**
Apple	Queen	Pear	Grape	Mango	The Rolling Stones	The Beatles	ABBA
Yatzi	Bipolar disorder	Chess	Monopoly	Snakes & Ladders	ADHD	Autism	Schizophrenia
Saturn	Italy	Mars	Uranus	Neptune	Ghana	Peru	Mongolia

Note. The upper panel overviews the different response alternatives during the baseline tests, the middle panel overviews the alternatives during the training phases, and the lower panel overviews the alternatives during the generalization test. Correct responses were all of the P- and $\bar{Q}$-cards.

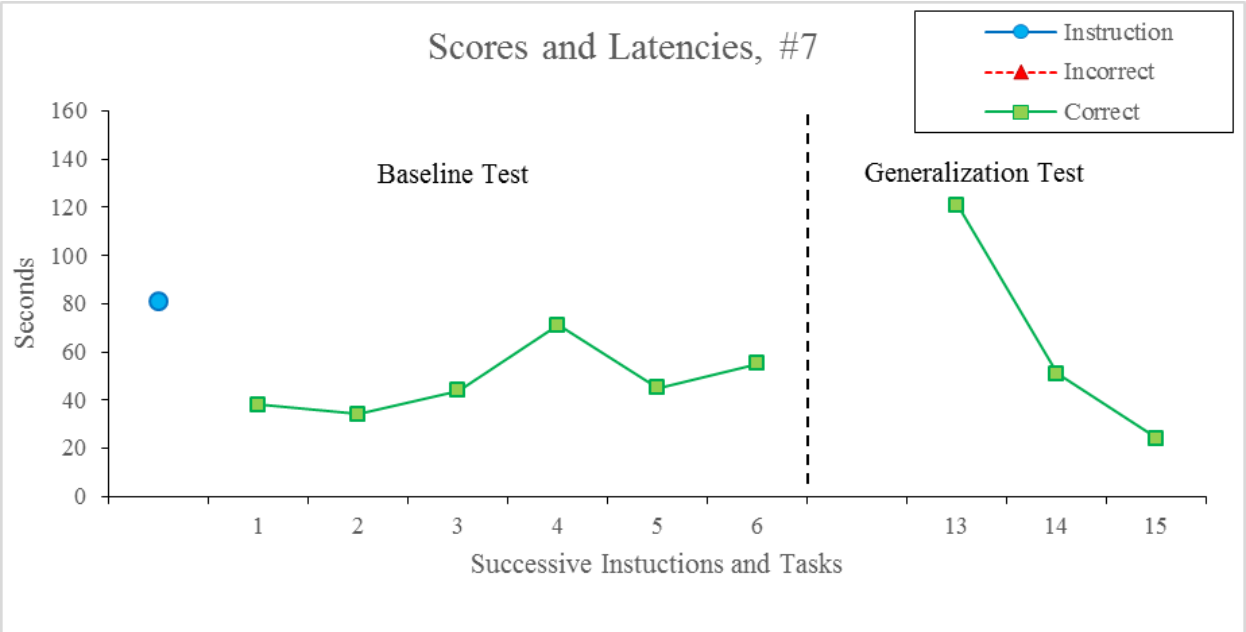


Figure 1. Shows the responses pattern produced by participant #seven, one of the two participants who disqualified for further participation. The X-axis represents successive instructions and tasks, while the Y-axis represents time used in s. The blue circle display reading of instruction. Green squares display correct responses.

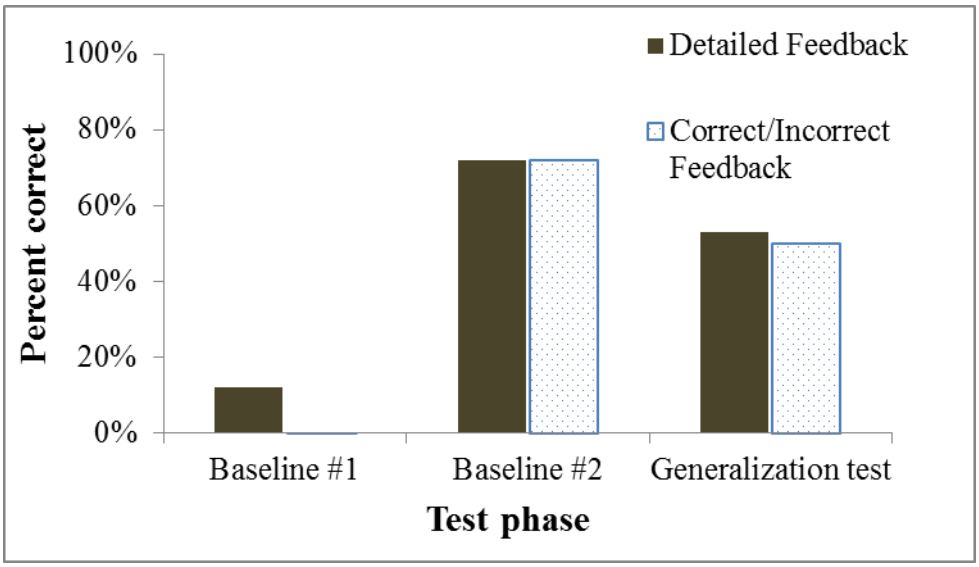


Figure 2. The dark scribbled column and the light scribbled column show the distribution of correct responses during the first and the second baseline and the generalization test for participants exposed to the explanatory feedback and the correct/incorrect feedback, respectively.

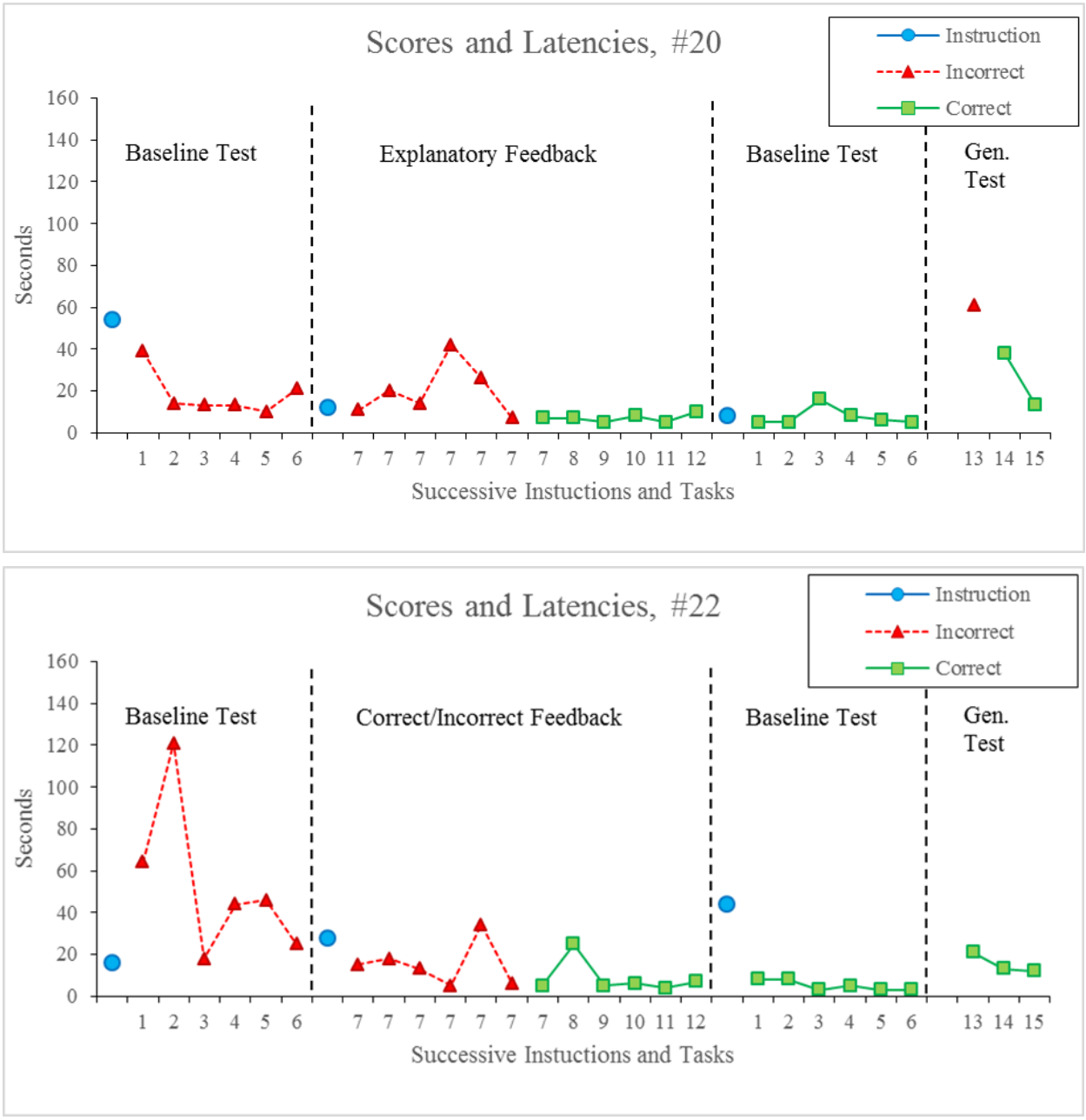


Figure 3. The upper and lower panel illustrates the response pattern produced by participant #20 and 22, respectively. These were two of the 11 participants who responded correctly to at least five of the six second baseline tasks and at least two of the three generalization tasks. The X-axis represents successive instructions and tasks, while the Y-axis represents time used in s. Blue circles display reading of the instructions. Green squares display correct responses, while red triangles display incorrect responses.

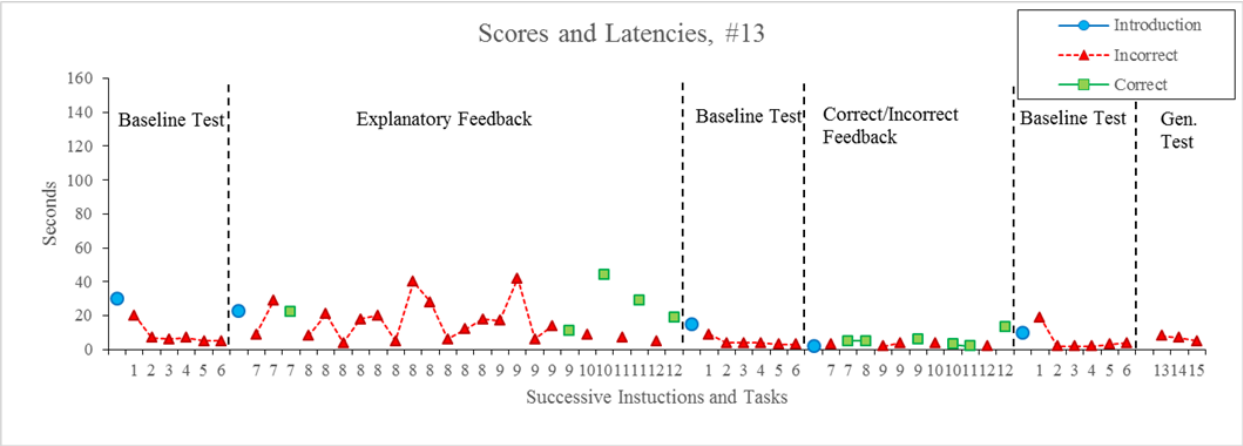


Figure 4. The response pattern produced by participant #13, one of the two participants who were exposed to both types of training. The X-axis represents successive instructions and tasks, while the Y-axis represents time used in s. Blue circles display reading of the instructions. Green squares display correct responses, while red triangles display incorrect responses.

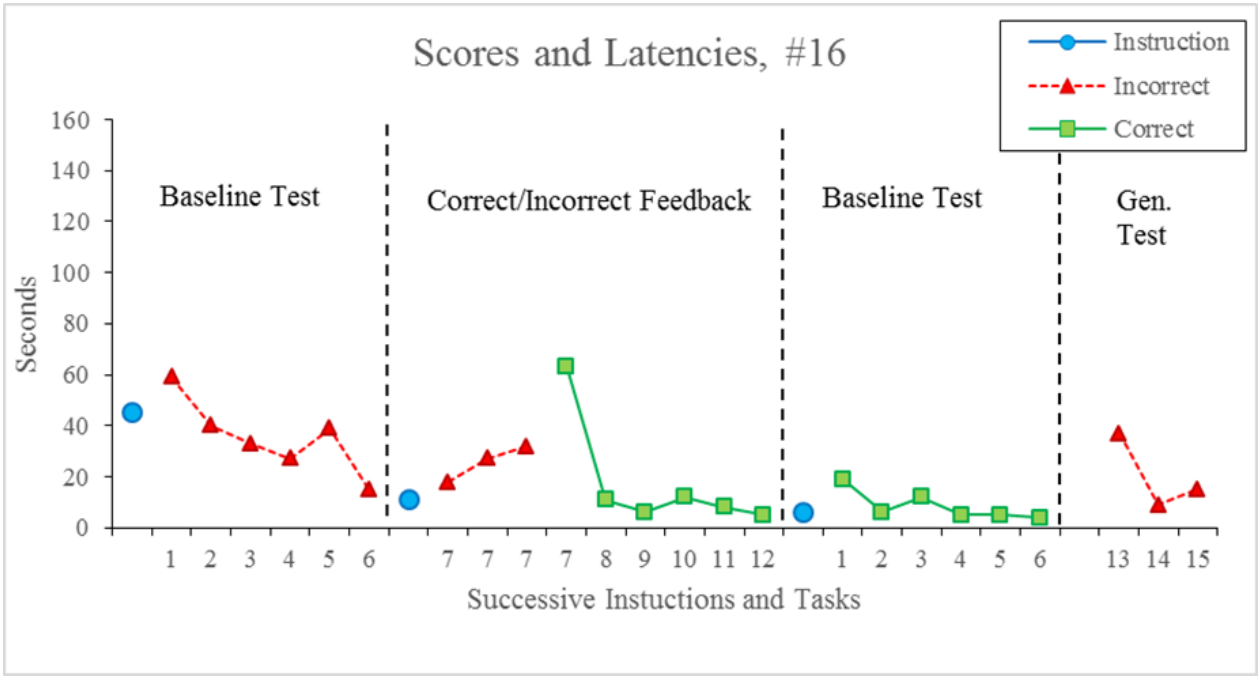


Figure 5. The response pattern produced by participant #16, one of the two participants who did not pass the generalization test, but responded correctly to at least five out of the six tasks during the second baseline test. The X-axis represents successive instructions and tasks, while the Y-axis represents time used in s. Blue circles display reading of instructions. Green squares display correct responses, while red triangles display incorrect responses.

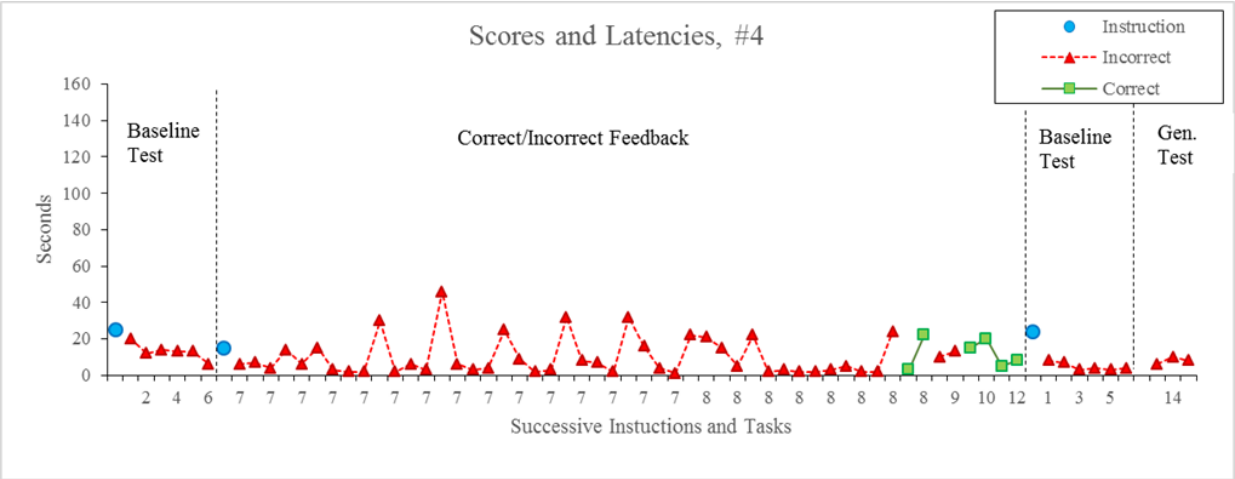


Figure 6. The response pattern produced by participant #four, one of the five participants who did not pass the generalization test and responded correctly to few tasks though the whole experiment. The X-axis represents successive instructions and tasks, while the Y-axis represents time used in s. Blue circles display reading of instructions. Green squares display correct responses, while red triangles display incorrect responses.

Appendix

Reflection note:

Regional Committees for Medical and Health Research Ethics (REK) have some guidelines for medical and health research. This involves mandatory notification in cases of personal data collection, medical and health research, research biobanks, or to get dispensation from secrecy. None of the above has been relevant for the present study.

According to Norwegian law, experimenters are obligated to inform participants in cases where personal data are collected. Additionally, voluntary consents from participants are required. In such cases, experimenters are obliged to submit notification to Norwegian Centre for Reseach Data (NSD). According to NSD's checklist, the present study did involve collection of any directly or indirectly identifiable data. No photo, video, or sound recordings were collected.

The experimentor showed up in student's class rooms and asked if some of them wanted to participate in an experiment that involved problem solving tasks on a computer. The volunteers were informed that they had the right to withdraw from the study at any point. They registered their gender and their age class (e.g., "26-30 years old"). The experimenter also recorded whether the participants were psychology or "vernepleier" student. No report or register of personal data was necessary, seeing that the participants only needed to show up once.

To summarize, the present experiment did not involve any collection of person identifiable data and was therefore not subject to notification.