Hypothesis Falsification in the 2-4-6 Number Sequence Test: Introducing Imaginary Counterparts

Article *in* SSRN Electronic Journal · January 2015

DOI: 10.2139/ssrn.2691667

1 author:

Michelle B. Cowley-Cunningham

Royal Statistical Society

104 PUBLICATIONS **133** CITATIONS

SEE PROFILE

Cowley, M. (2015). Hypothesis falsification in the 2-4-6 numbers test: Introducing imaginary counterparts. *Philosophy of Mind eJournal*, vol.8, Issue 41: November 25, 2015. See also the *Cognition in Mathematics, Science, & Technology eJournal*, vol. 7, Issue 42: December 3, 2015.

Cowley, M. (2015). Hypothesis falsification in the 2-4-6 numbers test: Introducing imaginary counterparts. *Political Behaviour: Cognition, Psychology, & Political Behaviour eJournal*, vol.9, Issue 66: December 3, 2015. See also *Games & Political Behaviour eJournal*, vol. 9, Issue 48: December 1, 2015.

Hypothesis Falsification in the 2-4-6 Number Sequence Test:

Introducing Imaginary Counterparts

M. Cowley

Royal Statistical Society, Fellow Member 2012

Abstract

Two main cognitive theories predict that people find refuting evidence that falsifies their theorising difficult, if not impossible to consider, even though such reasoning may be pivotal to grounding their everyday thoughts in reality (i.e., Poletiek, 1996; Klayman & Ha, 1987). In the classic 2-4-6 number sequence task devised by psychologists to test such reasoning skills in a simulated environment – people fail the test more often than not. In the 2-4-6 task participants try to discover what rule the number triple 2-4-6 conforms to. The rule is 'ascending numbers', but it is tricky to discover this rule. Participants tend to generate hypotheses with the properties of the 2-4-6 triple, for example, 'even numbers ascending in twos'. They must search for evidence to test whether their hypothesis is the rule. But experimental evidence has shown that they tend to generate confirming triples that they expect to be consistent with their hypothesis rather than inconsistent falsifying triples. Counter to the two main hypothesis testing theories this paper demonstrates that falsification is possible in five 2-4-6 task experiments when participants consider an Imaginary Participant's hypothesis. Experiment 1 and 2 show that competition with an opponent hypothesis tester facilitates falsification. Experiments 3 to 5 show that the consideration of an alternative hypothesis helps this falsification of hypotheses lead to rule discovery. The implications of the results for theories of hypothesis testing and reasoning are discussed.

Key words: Hypothesis Falsification; Imaginary Participant 2-4-6 Task; Opponent Hypothesis Testing

Hypothesis falsification, evidence that shows a hypothesis to be untrue, has been considered the cornerstone of enlightened thinking (Popper, 1959). Yet the empirical evidence suggests that people find falsification difficult if not impossible in their everyday thinking (Poletiek, 1996; 2005). Hypotheses may start out as anticipations of future events, tentative solutions to problems, or even guesses about occurrences around us (e.g., Bruner, Goodnow, & Austin, 1956; Poletiek, 2001). Many aspects of cognition require people to inspect whether their hypotheses are accurate by searching for evidence.

Searching for inconsistent evidence that falsifies a hypothesis has a long tradition of being considered more rational than searching for consistent evidence that confirms a hypothesis in the psychology of reasoning (i.e., Popper, 1959; Wason, 1960). For example, scientists may need to search for falsifying evidence to ensure their theories accurately represent the laws of nature (e.g., Fugelsang, Stein, Green, & Dunbar, 2004); military strategists may search for evidence of potential negative consequences, such as an opponent army responding to a plan in a way that was not anticipated (e.g., Cowley & Byrne, 2004); and problem solving may require searching for evidence that a hypothetical solution works out by searching for the possible ways a solution may not work out (e.g., Gobet, de Voogt, & Retschitzki 2004). Medical experts generate hypotheses to understand the causes of disease in order to develop cures for illnesses (e.g., Christensen-Szalansky & Bushyhead, 1981), and they must discriminate between relevant and irrelevant symptoms to diagnose illness (e.g., Koriat, Lichtenstein & Fischhoff, 1980). Consider the following example to illustrate the importance of falsification in scientific discovery:

You are a scientist and your job is to identify the cause of a dangerous new disease. You identify a previously unrecognized virus in tissue samples of symptomatic patients and your hypothesis is that this 'new virus' is the cause of the disease. However, other scientists have identified two viruses, including your new virus in their tissue samples. They hypothesise that it is the 'other virus' and not the new virus that is the cause. Both hypotheses have confirming evidence. A case is reported where the new virus is present and the other virus is absent. What should you conclude?

A situation similar to this one faced scientists working on the cause of one of the main contemporary major international health crises—the SARS virus. They concluded that the 'new virus' hypothesis was correct. The case where the 'other virus' was absent falsified the 'other virus' hypothesis and proved that the 'new virus' hypothesis was right. The example illustrates how falsification can be vital to the discovery of truth.

Yet the empirical investigations of hypothesis falsification show that people tend to overwhelmingly seek confirming evidence to prove the truth of cherished yet untrue hypotheses rather than seek evidence to falsify these hypotheses. This tendency to seek consistent confirming evidence and avoid inconsistent falsifying evidence is called *confirmation bias*, and it has been found to be pervasive whether thinking takes place in the context of laboratory experiments (e.g., Wason, 1960; Wetherick, 1962; Tweney *et al.*, 1980; Mynatt, Doherty, & Tweney, 1977; 1978; Poletiek, 1996), social contexts (e.g., Snyder & Swann, 1978; Wagenaar, van Koppen, & Crombag, 1993), or scientific discovery (e.g., Mitroff, 1974; Gorman, 1995; Kuhn, 1996).

This paper presents convincing evidence to the contrary and shows that people can engage in hypothesis falsification more often than has been found in the reasoning literature to date. This paper will focus on the 2-4-6 task which has been the classic

test-bed reasoning task for investigations of hypothesis falsification for over forty years. First, a detailed analysis will be presented of the precise logic which is essential to disentangling what confirmation and falsification has meant to different researchers at different times in the history of reasoning. It is constructive to focus this paper's subsequent experimental analysis on the 2-4-6 task because the results are not only directly comparable to the literature, but to the accuracy of the theoretical predictions made by the two main theories of hypothesis testing in this task (i.e., Poletiek, 1996; 2005; Klayman & Ha, 1987). Second, a novel component to the 2-4-6 task is introduced. By simply asking people to consider an Imaginary Participant, people were able to think about how to falsify a hypothesis. Third, a detailed analysis of the theoretical predictions of the two main hypothesis testing theories are explicated, and a critical analysis of these predictions is tested with the experiments that follow. Let us turn next to the 2-4-6 task.

The 2-4-6 Task

In the 2-4-6 task participants are instructed to discover a rule the experimenter has in mind that the number triple 2-4-6 conforms to. The participant is analogous to the scientist, and the experimenter's rule is analogous to the law of nature to be discovered (Wason, 1960). The experimenter's rule is simply 'any ascending numbers' but participants tend to focus on the salient features of the initial 2-4-6 triple and generate hypotheses such as 'even numbers ascending in twos'. They propose triples consistent with their hypothesis such as 10-12-14 and 16-18-20, rather than triples inconsistent with their hypothesis such as 5-10-15. If they had test their hypothesis with at least one triple that is inconsistent with their hypothesis, such as 5-10-15 which contains odd numbers, their hypothesis 'even numbers ascending in twos' would be falsified. Participants would then know that odd numbers are consistent with the experimenter's rule and they can infer that their hypothesis containing the property of evenness was incorrect. Instead participants have been found to persist in testing with triples that would lead to confirmation such as 10-12-14 (e.g., Wason, 1960; Poletiek, 1996). This tendency for people to seek out information consistent with their hypotheses and avoid inconsistent information is termed *confirmation bias*. The result has been replicated many times in the 2-4-6 task (e.g., Tweney *et al.,* 1980; Gorman, Gorman, Latta, & Cunningham, 1984; Kareev & Halberstadt, 1993), and has contributed to the view that human thinking was irrational and biased (e.g., Evans, 1989; Evans, Newstead, & Byrne, 1993).

Yet the documented inability to search for falsifying evidence to overcome untrue hypotheses presents us with a paradox (Poletiek, 1996). How can people be irrational hypothesis testers given the scientific and technological advancement they are capable of achieving? For example, how can we put a man on the moon if our thinking is inherently flawed (Mitroff, 1974)? One possibility is that people are more capable of falsification than previously shown. Another possibility, which this paper explores next, is that there has been a problem with the classification of people's hypothesis testing in the 2-4-6 task (Wetherick, 1962; Klayman & Ha, 1987).

The logic of hypothesis testing: Forty years of misdiagnosis in the 2-4-6 task?

Initial classifications of hypothesis testing as confirming and falsifying tests were equated with consistent tests (tests that were consistent with the participant's hypothesis) and inconsistent tests (tests that were inconsistent with the participant's hypothesis) (Wason, 1960). But Wetherick (1962) argued against this division. Instead, a four-way classification was suggested where confirmation and falsification

were split into two different strategies based not only on whether participants expected instances to be consistent with their hypothesis *but* also whether participants *intended* instances to be consistent with the experimenter's rule. For instance, when a participant's hypothesis is 'numbers ascending in twos' and they generate the test triple 3-5-7, it is clear that 3-5-7 is consistent with the participant's hypothesis because it ascends in twos. But this test is only a confirming test if the participant *intends* the triple to confirm by also expecting it to be consistent with the experimenter's rule. If the participant expects a 'yes' from the experimenter, they are attempting to confirm their hypothesis, and they expect their hypothesis is correct. But if the participant expects a 'no' from the experimenter, they are attempting to falsify their hypothesis as they expect their hypothesis is incorrect.

The same is true for inconsistent tests. For example, when a participant's hypothesis is 'numbers ascending in twos' and they generate the test triple 5-10-15, it is clear that 5-10-15 is inconsistent with the participant's hypothesis because it is not ascending in twos. However, it is a falsifying test only if the participant intends the triple to conform to the experimenter's rule. If the participant expects a 'yes' from the experimenter, then they expect a triple that is inconsistent with their hypothesis to be consistent with the experimenter's rule, therefore they expect their hypothesis to be incorrect. But, if the participant expects a 'no' from the experimenter, then they are in fact attempting to confirm. They expect that the triple 5-10-15 is neither inconsistent with their hypothesis 'numbers ascending in twos' nor with the experimenter's rule. The inconsistent test in this instance is intended to provide confirmation. Inconsistent tests can be intended to either confirm or falsify. Later theorists also considered the above system to be the best method for classifying confirming and falsifying hypothesis tests in the 2-4-6 task (e.g., Poletiek, 1996), but the terminology used to describe this classification has changed (Klayman & Ha, 1987; 1989). A test triple that is consistent with a hypothesis test is renamed a *positive test*, because it is a positive instance of the hypothesis, so 3-5-7 is a positive test of the hypothesis 'numbers ascending in twos'. A test triple that is inconsistent with a hypothesis test is renamed a *negative test*, because it is a negative instance of the hypothesis, so 5-10-15 is a negative instance of the hypothesis 'numbers ascending in twos'. Positive and negative tests have been split into confirming and falsifying sub-classifications: positive confirming (i); positive falsifying (ii); negative falsifying (iii); negative confirming (iv).

Although this classification has now been accepted as the best way to classify confirming and falsifying hypothesis tests in recent years, earlier research on the 2-4-6 task tended to rely only on the distinction between positive and negative tests to distinguish confirming and falsifying hypothesis testing (e.g., Gorman, Gorman, Latta, & Cunningham, 1984; Tweney et al., 1980; Kareev & Halberstadt, 1993). For example, when researchers tried to improve participant's ability to falsify in the 2-4-6 task by instructing them to falsify, they based their instructions on the concept of confirmation as a positive test and falsification as a negative test (e.g., Gorman, Gorman, Latta, & Cunningham, 1984; Gorman & Gorman, 1984). Their analysis did not record participants' intention to confirm or falsify, and as a result confirmation and falsification may have been confused in many studies (see Klayman, 1995; Poletiek, 2001 for review). Critically a test is considered to be a confirming test when it is intended to confirm a hypothesis. Likewise, a test is considered to be a falsifying test when it is intended to falsity a hypothesis. To clarify an example of each test type is presented in Table 1.

Table 1

Table 1: Categorising confirming and falsifying test types in the 2-4-6 task for the hypothesis 'even numbers ascending in twos'.

Test triple	Is the test triple a positive or negative test?	Does the person intend the test to confirm or falsify?	**Confirming or falsifying test triple**
8-10-12	positive	confirmation expected	**Confirming**
24-26-28	positive	falsification expected	**Falsifying**
5-10-15	negative	falsification expected	**Falsifying**
23-25-27	negative	confirmation expected	**Confirming**

In the 2-4-6 task the terminology of hypothesis testing has not only reflected how confirmation and falsification have been measured, but how hypothesis testing has been labelled over time. Falsification has sometimes been termed *disconfirmation* in case studies of scientific discovery (e.g., Gorman, 1995a; 1995b). Confirmation has sometimes been labelled given the processes underlying the strategy. For example, confirming has not always been seen as a conscious process but the result of a preconscious bias to attend to information that is positive rather than negative, such as attributing more relevance to triples leading to 'yes' than 'no' responses from the experimenter (Evans, 1989). Although recently participants were found to be able to use triples that generated feedback of 'no'. When participants were told there were two rules to be discovered they used these negatively labelled triples just as effectively as triples that generated 'yes' feedback (Gale & Ball, 2003).

To summarise the different ways researchers have defined hypothesis testing strategies over the last forty-five years in the 2-4-6 task, a table is presented below:

Table 2

Table 2: The different ways hypothesis testing strategies have been conceptualised in hypothesis testing research over the past forty-five years.

Term used	Definition and main author(s)
Severity of test	Severity of test is a philosophical term used to refer to falsification. A hypothesis tester should test their hypothesis as severely as possible. In other words, they should choose a test that can result in the strongest possible evidence against a hypothesis. This type of hypothesis testing was termed falsification (Popper, 1959).
Falsification	Falsification became the favoured scientific and psychological term used to refer to the severity of test as outlined above. Falsification has tended to be associated with the search for evidence to show a hypothesis to be untrue (e.g., Wason, 1960).
Confirmation bias	Confirmation bias is a tendency to search for evidence that is consistent with a hypothesis and avoid inconsistent evidence (e.g., Wason, 1960).
Positive and negative test strategies	A triple that is consistent with a hypothesis is a *positive test* of that hypothesis. For example, the triple 8-10-12 is generated when the hypothesis is 'even numbers ascending in twos' because it contains the target properties of evenness and ascending in twos. A triple such as 5-10-15 is inconsistent with the hypothesis because it does not contain these target properties and it is called a *negative test*. Participants may have a tendency to test cases that have the property of interest rather than those that do not have the property in the 2-4-6 task, that is, they have a tendency to follow a positive test strategy of testing positive instances, which does not necessarily constitute a bias in all reasoning contexts (Klayman & Ha, 1987).
Intentional confirmation and falsification	Confirmation and falsification depend on whether a test is consistent with a hypothesis *and* on whether it is intended to confirm or falsify. Participants' tendency to generate triples that are consistent with a currently held hypothesis may not constitute a bias, because they may not 'expect' a consistent triple to result in a confirming response from the experimenter. Participants must *expect* a confirmation for it to constitute a confirmation bias. Likewise, participants must expect a falsification for it to constitute a falsification (Wetherick, 1962).

Positivity bias	One claim is that human reasoning is biased towards attending to positive instances of a current representation at a preconscious level (Evans, 1989). This tendency corresponds to a bias for positive instances of a current hypothesis and selecting these positive instances as hypothesis tests symptomatic of confirmation bias.
Disconfirmation	There may be two levels of hypothesis testing. At the micro-level hypothesis testing corresponds to individual tests, for example one experiment may falsify a hypothesis, but at the macro-level hypothesis testing corresponds to a series of tests, for example a series of experiments which lead to disconfirmation of a theory (Gorman, 1995a).

But how do we discern ways of discriminating between confirmation bias, non-biased confirmation, and falsification? First, this paper suggests that a test can be considered an instance of *confirmation bias* in the following circumstances when a hypothesis is untrue (Cowley & Byrne, 2004; 2005): (i) when participants indicate in their responses that they intend their test to result in confirmation of their hypothesis, *even though* falsifying evidence is available (in line with Wetherick, 1962, Klayman & Ha, 1987; Poletiek, 1996); (ii) *or* when participants evaluate the result of a test as confirming their hypothesis when the test result objectively falsifies their hypothesis. Second, this paper suggests that a test can be considered an instance of *falsification* in the following circumstances when a hypothesis is untrue: (i) When falsifying evidence is available to the participant, and participants indicate in their responses that they intend their test to result in falsification of their hypothesis (in line with Wetherick 1962; Klayman & Ha, 1987; Poletiek, 1996); (ii) *or* when participants evaluate the result of a test as falsifying their hypothesis when the test result objectively leads to falsification. Third, this paper suggests that a test can be considered an instance of non-biased confirmation in the following circumstance: When the hypothesis is true or of exceptional quality such that there is very little falsifying evidence to search for, and a hypothesis test, even though it is intended to falsify, may result in confirmation (e.g., Poletiek, 1996; 2005). In other words when a person seeks to falsify their hypothesis as much as possible in order to identify falsifying cases, if any exist. But these severe tests in fact lead to confirmation of a hypothesis. In this case the hypothesis is confirmed but not in a biased way (Cowley & Byrne 2005). [1]

In the next section two main theories are outlined which have been developed to explain the findings observed in the 2-4-6 task. The main tenets of each theory are explained, and how the factors pertinent to these main tenets affect hypothesis testing. The shortcomings of each theory are considered and we describe how the experimental designs employed in this paper test the main tenets of the each theory.

[1] It is important to note the distinction between the process of a test choice and the outcome of a test choice when we refer to confirmation and falsification in the above examples. For example, when a hypothesis is generated it may actually represent the true state of affairs. To test this hypothesis a person may generate a test with the intention to falsify it, but because the hypothesis is in fact true the test outcome can only confirm the hypothesis regardless of the process the person has used (in the experimental chapters I detail this point further.

Theories of Hypothesis Testing in the 2-4-6 Task

I will now outline two main theories of hypothesis testing developed from findings in the 2-4-6 task. The first theory proposes that people find falsification difficult if not impossible in the 2-4-6 task, and that confirming and falsifying are one and the same process (Poletiek, 1996; 2001; 2005). This paper will refer to this theory as the uniformity theory of hypothesis testing in the 2-4-6 task because it proposes that confirming and falsifying testing are the same process. The second theory proposes that hypothesis testing is constrained by the mathematical structure of the hypothesis testing task at hand (Klayman & Ha, 1987). Klayman and Ha suggest that people find it difficult to falsify, not because they find falsification impossible, but because their tendency to use positive tests is not conducive to falsification due to the mathematical constraints in the standard 2-4-6 task (Klayman & Ha, 1987). This paper will refer to this theory as the mathematical relationship theory of hypothesis testing in the 2-4-6 task. Each theory is now explained in turn and the main tenets of each outlined.

The uniformity theory (Poletiek, 2001)

Are people able to perform two distinct types of hypothesis tests, that is, confirming and falsifying tests? Or do people perform just one type of test that will either lead to a confirming or falsifying outcome depending on the quality of the hypothesis rather than their own test choice (Poletiek, 1996)?

Recent evidence from the 2-4-6 task indicates that people cannot sensibly intend to confirm or falsify (Poletiek, 1996, Experiment 1). To test hypotheses, people perform a test, and the test will either confirm or falsify a hypothesis depending on the quality of the hypothesis initially generated (Poletiek, 1996, Experiment 2). In other words participants cannot deliberately intend to falsify or control test outcomes in order to falsify a hypothesis; hypothesis testing is simply experienced as performing a test and therefore confirmation and falsification are the same strategy (Poletiek, 2001; 2005). In one experiment participants were explicitly instructed to falsify their hypothesis in the standard version of the 2-4-6 task (Poletiek, 1996 , Experiment 1). The rule to be discovered was the 'any ascending numbers' rule and participants were instructed to generate their 'best guess', that is, their hypotheses about what the rule might be. Participants typically generated hypotheses pertaining to 'evenness' or 'ascending in twos'. The only type of hypothesis test a participant can use to intentionally falsify a hypothesis such as 'even numbers ascending in twos' in the standard 2-4-6 task is a *negative falsifying test* triple; such as 5-10-15 which they then *expect* to lead to falsification (See Table 1).

The ability to generate a negative-falsifying test is pivotal to the debate about whether people can falsify in a useful way. Participants were given instructions either to 'test', 'confirm', or 'falsify' (Poletiek, 1996). For the 'test' and 'confirm' conditions, the majority of tests fell into the positive confirming category (86% and 80% respectively), and few tests fell into the negative falsifying category (0% and 3% respectively). Participants in the 'falsify' condition were instructed to 'try to test in such a way as to get your hypothesis about the rule rejected' (Poletiek, 1996; p.454). The majority of tests in this condition fell into the two confirming categories, the positive confirming and negative confirming categories (32% and 54% respectively). (See Table 1). Although the participants who were instructed to falsify proposed test triples that were negative tests, they in fact intended these tests to confirm. It was concluded that people do not seem to be able to make sense of falsification because they expect their test result to confirm their hypothesis regardless of the tests they proposed. Poletiek (1996) concludes that people are unable to

intentionally perform negative falsifying tests and therefore they find falsification an impossible hypothesis testing strategy to conduct. Poletiek explains that negative tests are a first reflex to make a mismatch between the hypothesis and test item when participants are instructed to falsify, because participants appear to have little insight into their test choices. In other words confirmation and falsification are experienced as a uniform process by participants, that is, they are experienced as the process of carrying out a hypothesis test.

On the surface this claim may make intuitive sense. However, participants may not have been given adequate opportunity to show they could intentionally falsify. First, participants were requested to generate three test triples in each condition in comparison with the previous literature allowing the generation of a minimum of fifteen triples (See Klayman & Ha, 1989), or up to forty five minutes of testing (e.g., Wason, 1960). Second, the results report statistical analyses for the first test triple only, the remaining two triples were excluded, suggesting that the uniformity theory was initially developed from a small data set of ninety-four triples (ninety-four people generated one triple each) (Poletiek, 1996, p.455). But negative tests do not tend to appear until at least after the first three test triples (Klayman & Ha, 1989), and attempts at falsification may occur at a later stage in the hypothesis testing process (e.g., Mynatt, Doherty, & Tweney, 1978). The conclusion that people find it impossible to intentionally falsify may be an artifact of the limited opportunity and analysis of the data in the experiment. We summarise the main tenets of Poletiek's uniformity theory in Table 3 below. We suggest experimental tests that may falsify the theory by showing hypothesis testers can experience falsification as possible and as distinct from confirmation.

Table 3

Table 3: Tenets of the uniformity theory (Poletiek, 1996; 2001)

Tenet 1:	Falsification is impossible because it presupposes that people know where to find information to intentionally falsify a hypothesis.
Criticism 1:	Falsification may be possible. For example when testing somebody else's as opposed to one's own hypothesis people may have information that will help to generate a falsifying test with the aim of falsifying that test. Or people can intentionally generate tests inconsistent with a current hypothesis (i.e. negative tests) and expect them to result in falsification (i.e. negative tests). Given the opportunity to test more than three triples, people may begin to use these negative falsifying tests.
Tenet 2:	Falsification is indistinguishable from confirmation and they are the same process, because the strongest attempt at falsification of a hypothesis results in the most convincing type of confirming evidence should that attempt to falsify fail.
Criticism 2:	The strongest attempt at falsification may lead to the most convincing type of confirming evidence should that attempt to falsify fail. Even though hypothesis testers may choose the same test, for example a

negative test in the standard 2-4-6 task, an objective hypothesis tester may intend it to falsify, whereas a biased hypothesis tester may intend it to confirm. The process of confirmation and falsification may be distinct (See Table 1.1).

Tenet 3: A result of a hypothesis test may be as much a consequence of the quality of the hypothesis under test, as of any specific strategy employed by the hypothesis tester.

Criticism 3: The result of the hypothesis test may be a consequence of the quality of the hypothesis under test, but if hypothesis quality is responsible for the test result it implies that people do not have an active role in hypothesis testing. But it may be possible for individuals to be active. Consider an experiment in which two conditions are compared when the hypothesis being tested in each condition is equally untrue and an additional factor leads to falsification in one condition and not in the other.

We turn now to examine the second main theory of hypothesis testing in the 2-4-6 task—the mathematical relationship theory.

The mathematical relationship theory (Klayman & Ha, 1987)

The second hypothesis testing theory we describe posits that the type of mathematical relationship between the hypothesis under test and the rule to be discovered affects hypothesis testing. Consider the situation in the standard 2-4-6 task when the participant's hypothesis is 'even numbers ascending in twos' and the properties of evenness and ascending in intervals of two are embedded in the experimenter's rule 'any ascending numbers'. 'Any ascending numbers pertains to any numbers that increase by any interval.[2] Klayman & Ha (1987, 1989) suggest that this embedded relationship is the most difficult for participants, because it is the only relationship that requires them to discover that their hypothesis is incorrect by generating a *negative test* that leads to falsification, whereas positive tests which participants may find easier to generate can lead to falsification in several of the other relationship types including another type of embedded relationship.

For example, when the experimenter's rule is 'any ascending numbers' and the participant's hypothesis is 'even numbers ascending in twos' the triple 3-5-7 is a negative test because it is not an instance of 'even numbers ascending in twos' as it contains odd numbers. When the researcher replies 'yes' indicating that a triple with odd numbers is consistent with the experimenter's rule, the hypothesis pertaining to evenness is falsified.

Klayman and Ha point out that this relationship is not representative of the majority of hypothesis testing situations that can occur, and people tend to test their hypotheses using positive tests, which are more effective at producing falsification in other hypothesis testing situations. Consider a scientist researching the cause of a birth defect such as *Spina bifida*. Spina bifida is a neural tube defect affecting spinal chord development in the early stages of pregnancy. Scientists (Molloy & Scott, 2001) hypothesized that genetic factors were the cause of the defect. They noticed a high

[2] This embedded relationship is one of five possible relationships that can occur given variations of what the experimenter's rule and the participant's hypothesis could be. The paper focuses on the three relationships relevant to the standard 2-4-6 task and the experiments that follow.

incidence rate of babies being born with Spina bifida in the Celtic gene pool of Ireland and Scotland. Breakthrough research chose to examine blood samples and family history data collected in genetic studies of the Irish population. The choice of an Irish test population for examination was a positive test of the hypothesis that genetic factors were the cause. This positive test led to a theory of genetic predisposition as a major cause of neural tube defects because a significant pattern of neural tube defects was observed in relationships among families from the Celtic gene pool.

Consider if the scientists had carried out a negative test of their hypothesis by focussing on the African gene pool that was not composed of the hypothesized risk factors. They would have found close to zero percent cases of Spina bifida and their search would not yield new information because it would be like searching for a needle in a haystack (Klayman & Ha, 1987). This example shows that it may be often more useful to examine positive instances from the group composed of the hypothesized risk factors in scientific research. For this reason Klayman and Ha suggest that people have a tendency to engage in a general *positive test strategy* because they are familiar with the usefulness of engaging in hypothesis testing in real world examples. Yet the traditional 2-4-6 task does not allow a positive test to lead to falsification. A negative test leads to falsification in one relationship (the typical 2-4-6 task situation), and a positive test leads to falsification in the other (the situation akin to the Spina bifida example). Both of these relationships are called embedded situations. Critically, a negative falsifying test is best in the first situation, and a positive falsifying test is best in the second. The first embedded relationship is the one characteristic of the 2-4-6 task where the experimenter's rule applies to 'any ascending numbers' and it overlaps any triples that are even and/or ascend in twos such as when the participant's hypothesis is 'even numbers ascending in twos'. This relationship is illustrated in Figure 1 (a).

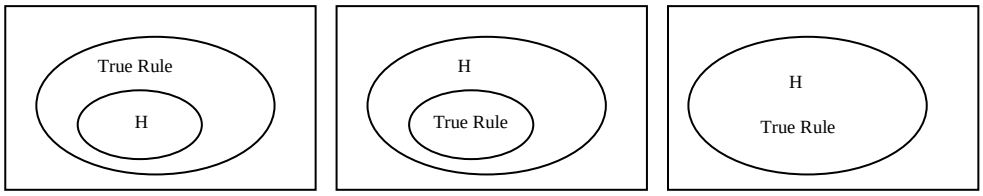

Figure 1: Embedded relationships between a participant's hypothesis (H) and the experimenter's rule (True Rule).

The only way to intentionally achieve falsification in this relationship is to use a negative falsifying test. For example, consider a participant who generates the triple 5-10-15 (which is a negative test as ascending in five or odd numbers is inconsistent with the hypothesis 'even numbers ascending in twos'), and they expect it to be consistent with the true rule. They will receive a 'yes' from the experimenter, because 5-10-15 is consistent with 'any ascending numbers', and so they can infer that the hypothesis about 'evenness' and/or 'ascending in twos' cannot be true.

Consider on the other hand a participant who tries to intentionally falsify by generating a positive falsifying test such as 24-26-28 (which is a positive test because it is consistent with the hypothesis 'even numbers ascending in twos', *but* they expect it to be inconsistent with the true rule). Perhaps the rule only corresponds to triples ending in the digits 2, 4, 6, such as 2-4-6, 22-24-26 etc . This time when a 'yes' is received from the researcher they may not infer that the hypothesis pertaining to properties of 'evenness' and/or 'ascending in twos' is untrue. Although the positive

test was intended to falsify, it cannot. It is consistent with the hypothesis and the true rule.

In the second type of embedded relationship it is possible to falsify with a positive falsifying test when the participant's hypothesis overlaps the experimenter's rule, for example, when the hypothesis is 'numbers ascending in twos' and the experimenter's rule is this time 'even numbers ascending in twos'. (See figure 1 b). (The relationship is akin to the Spina bifida example). This time the true rule 'even numbers ascending in twos' is embedded within the hypothesis 'numbers ascending in twos'. Consider when a participant generates the triple 3-5-7 (which is a positive test as it is consistent with the hypothesis 'ascending in twos'), and they intend it to falsify because they expect it not to be consistent with the experimenter's rule. This time they receive a 'no' from the researcher, because 3-5-7 contains odd numbers. They can infer that their hypothesis is falsified and it does not correspond to the experimenter's rule because it may pertain to numbers with the property of 'evenness'.

Now consider a participant who tries to falsify by generating a negative test such as 5-10-15. They may intend this test to falsify by expecting it to be consistent with the experimenter's rule. This time when they receive a 'no' from the researcher they may not infer that the hypothesis 'numbers ascending in twos' is untrue. The triple is both inconsistent with their hypothesis and the true rule and so cannot discriminate between them (Klayman & Ha, 1987).

The third situation is when the hypothesis is the same as the experimenter's rule. where the hypothesis 'any ascending numbers' completely overlaps the true rule 'any ascending numbers'. (See figure 1c). When a participant generates a positive test triple such as 24-26-28, even if it is intended to falsify it receives a 'yes' response. This leads to ambiguous confirmation because 'any ascending numbers' contains an infinite number of triples that can be confirmed. And negative test triples such as 6-4-2 receive a 'no' response (because descending numbers are not consistent with the true rule 'any ascending numbers'). When a descending triple receives a 'no' it does not help the participant infer that their hypothesis 'any ascending numbers' is certainly the true rule. It is not possible to be certain that the hypothesis is the truth, but it is still possible to attempt to falsify it. Each failed attempt to falsify indicates that a hypothesis is at least close to the truth (Popper, 1959).

Klayman and Ha distinguish between positive or negative test strategies, and they suggest that participants generally 'choose' to generate positive test strategies rather than negative test strategies (pace Poletiek, 2001). They also indicate that people may find falsification possible using a negative test strategy when they are aware of what the relationship between the hypothesis and the rule to be discovered is (Klayman & Ha, 1989). However, their account does suggest that the hypothesis tester largely plays a passive role in hypothesis testing because they suggest the mathematical relationship between the hypothesis and the truth (in this case the experimenter's rule) is the major factor controlling how effective a participant's choice in hypothesis testing is. This suggestion is akin to the view of the uniformity theory which proposes that hypothesis quality creates a passive role for the hypothesis tester (e.g., Poletiek, 1996). Both the mathematical relationship theory and the uniformity theory suggest the hypothesis tester is largely at the mercy of the properties of their hypothesis (i.e., the quality of the hypothesis and the relationship between the hypothesis and the truth), post-hypothesis generation. If this assertion is true it implies that the discussion of hypothesis testing as being biased or rational may be redundant because people may not be in a position to actively pursue a confirming or falsifying strategy, and research should start to focus more on a previous stage in

the process such as the hypothesis generation stage. An important objective for this paper is to investigate whether people choose their tests in a way that reflects an active role for a hypothesis tester.

In Table 4 below the main tenets of Klayman and Ha's mathematical relationship theory are explained. Accordingly, the experimental tests that may falsify the theory are detailed.

Table 4

Table 4: Tenets of the mathematical relationship theory (Klayman & Ha, 1987; 1989)

Tenet 1:	There is a tendency to test instances that are consistent with a hypothesis. This tendency is called a *positive test strategy* and it is usually a helpful strategy in hypothesis testing, such as in the Spina bifida example. In the 2-4-6 task a positive test strategy is not the same as confirmation bias. Even if the participant intends their positive test to lead to falsification, it can only lead to confirmation of their incorrect hypothesis. The relationship between the hypothesis and the true rule constrains the effectiveness of the positive test strategy. Only negative tests can falsify in the mathematical relationship standard of 2-4-6 task; but participants find it difficult to disengage from positive testing and that is why they are not successful.
Criticism 1:	People sometimes successfully discover the rule in the 2-4-6 task. Perhaps there are conditions under which people readily follow a negative test strategy, for example, when they are competing with an opponent. The finding that participants disengage from positive to negative tests in the embedded relationship typical of the 2-4-6 task, with or without knowledge of task constraints, would indicate that the mathematical relationship alone cannot predict hypothesis testing—people can play an active role in hypothesis testing. For example, when people consider an alternative hypothesis in addition to the initial hypothesis they may generate negative tests.
Tenet 2:	People often do not know when a positive test strategy is wise and when it is not.
Criticism 2:	If people can show that they know when it is wise not to use a positive test strategy, and rely on a negative test strategy instead, then they do know when a positive test strategy will not work. For example, when people test a hypothesis belonging to somebody else that they know is untrue, they may rely on negative tests.
Tenet 3:	The mathematical relationship between the hypothesis and the experimenter's rule affects how useful positive and negative test strategies are.

Criticism 3: The mathematical relationship only affects whether or not a positive or negative test will lead to a confirming or falsifying outcome. The relationship does not affect the part of the process where people intend to falsify or confirm. For example, when people compete with an opponent hypothesis tester, they may attempt to falsify their hypotheses by generating negative tests, but they may actually intend these negative tests to confirm.

A critical question to ask is whether the ability to falsify in an embedded relationship using a negative test is important to hypothesis testing in the real world? Consider a how a prejudiced belief may be embedded by the truth because it consists of a smaller set of positive instances confirming a belief within the total population of the group against which the prejudice is targeted. A negative instance which would falsify this untrue belief exists outside of that set. For example, if someone held the prejudiced belief that Jews were lesser beings they may cite cases which are consistent (positive instances) with this belief, such as a person with a criminal record who was also Jewish, and avoid any inconsistent instances (negative instances) which exist outside of their collection of confirming evidence. But a falsifying case would be available in the story of Anne Frank, and one falsifying case can prove that this prejudiced belief is false. The standard version of the 2-4-6 task, when the participant's hypothesis is embedded within the true rule, is analogically equivalent to this prejudiced belief (Wason, 1960). Thus if people cannot falsify in this situation then they are being irrational.

But there have been some accounts of successful rule discovery in the 2-4-6 task. While the mechanisms by which participants have been successful are presently ill-specified the presentation of an alternative hypothesis appears to help. Next the paper briefly outlines some accounts of improved rule discovery in the 2-4-6 task and show how such accounts provided hints to the development of an ecologically valid standard 2-4-6 task in the experiments that follow.

Alternative Hypotheses: the key to successful hypothesis testing

Presently there is a collection of novel experimental findings relating to the role alternative hypotheses have in helping the discovery of the truth in the 2-4-6 task. The first experimental result to highlight the facilitating role of considering alternative hypotheses in rule discovery was the DAX-MED experiment carried out in the 2-4-6 task (Tweney et al., 1980). Participants were told that there were two rules to be discovered; a DAX rule and a MED rule. The DAX rule was the standard 'any ascending numbers' rule and the MED rule was 'any other number sequence which does not ascend'. Participants were instructed to generate number triples and the researcher responded with the feedback 'DAX' or 'MED' rather than 'yes' and 'no' respectively. Participants discovered the rule 'any ascending numbers' significantly more frequently than the usual 21% rule discovery rate; 60-80% of participants tend to discover the rule in DAX-MED manipulations, even though they have generated the same number of test triples (e.g., Valle-Tourangeau, Austin & Rankin, 1995; Wharton, Cheng & Wickens, 1993; Gale & Ball, 2003; 2005).

Little is known about how considering alternative hypotheses facilitates rule discovery.

A second explanation was that people's bias to process information with a positive label 'DAX' or 'MED'a s opposed to a negative label, for example 'yes' versus 'no' allowed the processing of more triples (e.g., Evans, 1989). But

participants who were asked to either discover one rule or two rules performed well regardless of linguistic labelling (Gale & Ball, 2003). A third explanation is that the DAX-MED manipulation induces a mental representation which requires less effort to switch between two alternative hypotheses and test them both simultaneously. That is, the two hypotheses are complementary to one another; one is 'ascending' and one is 'not ascending' (Wharton, Cheng, & Wickens, 1993). But similar rule discovery rates were found in a condition with feedback inducing non-complementary representation by labeling triples 'DAX'-'MED'-'DAX or MED' (Vallee-Tourangeau et al., 1995).

Oaksford & Chater (1994) suggest that participants may find the consideration of an alternative hypothesis useful because it helps them to decide which information to include or exclude from their hypothesis. For example, if the hypothesis under test is 'even numbers ascending in twos' and the experimenter's rule is 'any ascending numbers' participants may generate the alternative hypothesis 'numbers ascending in twos' which excludes the property of evenness (see also Farris & Revlin, 1989). Participants may then generate the triple 5-7-9, which ascends in twos, and ascends but is not even. If participants receive a 'yes' response from the experimenter for the triple 5-7-9, then they can revise their hypothesis to 'numbers ascending in twos' by excluding the property of evenness (see also Gale & Ball, 2005). Oaksford and Chater's account assumes that participants need to generate a specific alternative at the outset in order to discover what should be excluded from a hypothesis. That is, participants may implicitly represent the concept that another alternative exists and that it may offer a better explanation than their hypothesis.

In sum, despite the wealth of findings from research on hypothesis testing in the 2-4-6 we do not know if people can falsify, nor do we know if they can actively choose their hypothesis testing strategy, even though an active role is vital to proving that reasoning can be biased (e.g., Kahneman, Slovic, & Tversky, 1982) or rational (Johnson-Laird, 1983). In the experiments that follow an active factor that has been traditionally overlooked in hypothesis testing—competition is examined.

The Imaginary Participant 2-4-6 Task

This paper suggests that people might have a tendency to falsify in competitive situations because hypothesis testing in realistic settings may proceed by testing other people's hypothesis or interacting with an opponent. An Imaginary Participant called 'Peter'[3] is introduced to the standard 2-4-6 task and asked people to sometimes consider Peter's hypothesis and to sometimes consider an opponent called Peter. This Imaginary Participant 2-4-6 task was intended to be akin to real world situations. For example, scientists may often proceed by attempting to confirm their own hypotheses and falsify other scientists' hypotheses (e.g., Mitroff, 1974; Gorman, 1995a; Kuhn, 1996; Fugelsang et al., 2004),legal experts need to not only compete with opposition barristers, but to ensure that the grounds on which they base their legal arguments are irrefutable (e.g., Roberts & Zuckerman, 2005), and military strategists must engage with opposition forces, and ensure that they consider each possible alternative at the disposal of the opponent to their hypothesized plans of action (e.g., Mallie, 2001).

Perhaps with competition participants may either have their own alternative hypotheses from which to generate falsifying tests or understand that there are alternative hypotheses that an opponent hypothesis tester may be considering. Or when alternatives belonging to an opponent are non-explicit, the competition may

[3] In honor of the late Peter Wason who invented the 2-4-6 task.

prompt participants to flesh out these properties to consider what the opponent's alternatives might be.

The uniformity theory predicts that competitive factors should not help people to intentionally confirm or falsify their hypotheses using negative tests because they are one and the same process (Poletiek, 1996; 2001; 2005). The mathematical relationship theory predicts that competitive factors should have no effect on hypothesis testing. The mathematical properties of the task affect how successful hypothesis testing is in the standard 2-4-6 task (Klayman & Ha, 1987; 1989). Let us examine if these predictions hold in Experiment 1.

This experiment examines if ownership of a hypothesis is a competitive factor that affects falsification when the mathematical relationships in each experimental condition are identical. In the first condition participants are instructed to test 'Peter's hypothesis: even numbers ascending in twos'. In the second condition participants are instructed to test 'Your hypothesis: even numbers ascending in twos'. The experimenter's rule is 'any ascending numbers'.

Participants were predicted to not only generate more negative falsifying tests of an imaginary participant's hypothesis than their own, but that they will expect these negative tests to falsify. Hypothesis ownership was predicted to be a competitive factor that affects hypothesis falsification.

Materials and Design

Participants were randomly assigned to two conditions (n = 16 in each): in one condition the low-quality embedded hypothesis was identified as belonging to the "imaginary participant" Peter (Peter's hypothesis is 'even numbers ascending in twos') and in the other the identical hypothesis was identified as belonging to the participant (Your hypothesis is 'even numbers ascending in twos'). Participants were not made aware of what the experimenter's rule was. Crucially, the relationship between the hypothesis and the true rule was identical. Several factors were controlled for. First, the hypotheses are equally incorrect, so any differences observed in testing behaviour cannot be explained by the quality of the hypothesis and the amount of available evidence for participants (Poletiek, 1996, Experiment 2; Klayman & Ha, 1987). Second, the mathematical relationship between the hypothesis under test and the experimenter's rule is the same in each condition, so any differences observed in hypothesis testing between the two conditions cannot be explained by the mathematical relationship theory either (Klayman & Ha, 1987). Third, the participants in both conditions were given the hypothesis for testing, they did not generate the hypotheses to rule out explanations related to personal investment (e.g., Kunda, 1987; 2000). See Appendix A for the complete text of instructions given to participants.

Participants and procedure

Thirty two people who were members of the general public volunteered and were paid a nominal fee of 8 Euro per hour. There were 23 women and 9 men and their age ranged from 20 years to 75 years, with a mean age of 51 years. No participants had taken courses in the philosophy of science. Participants were tested individually. The testing session lasted approximately 20 minutes. Each participant was given a three-page recording sheet which had 5 columns. Appendix B provides a copy of the recording sheet. Participants wrote down their 'number sequence', 'yes' or 'no'

answers to the questions, 'do you expect it to conform to Peter's rule?' (i.e., positive or negative test), and 'do you expect it to conform to the experimenter's rule?' (i.e., intend to confirm or falsify). Feedback was given in the form of a 'y' for a yes and an 'n' for a no as to whether or not the generated number sequence conformed to the experimenter's rule. There were 18 rows in the recording sheet. There were also spare sheets for participants to insert as many tests as they considered necessary.

Results and discussion

Number of triples

In total, 184 triples were generated, with a mean of 5.75 triples per participant. A similar number of test triples were generated for the imaginary participant's hypothesis (M = 6.0) and for participants' own hypothesis (M = 5.5). Participants did not generate more tests of the hypothesis belonging to the imaginary participant Peter significantly more than their own hypothesis.

Positive and Negative tests

Participants generated more negative tests of Peter's hypothesis (46%) than of their own hypothesis (25%), but this difference was not reliable (chi^2 = 10.492 (6), p = .052). Participants generated fewer positive tests of Peter's hypothesis (54%), than of their own hypothesis (75%, chi^2 = 18.619 (9), p = .015).

Table 5

Table5: Percentages of positive and negative tests generated in Experiment 1 .

	Peter's hypothesis	Own hypothesis
Positive tests	54	75
Negative tests	46	25

The result that participants generated negative tests readily for the imaginary participant's hypothesis does not corroborate the mathematical relationship theory or the prediction that people engage more readily in a positive test strategy than a negative test strategy in hypothesis testing (Klayman & Ha, 1989). The mathematical relationships between the hypothesis and the rule were identical in each case. Participants generated more negative tests of the imaginary participant's hypothesis than their own, even though the hypothesis quality was the same (Poletiek, 1996). The experiment shows that hypothesis ownership can affect the generation of hypothesis test types.

Falsification and confirmation

Did participants intend to use their positive and negative tests differently to test the imaginary participant's hypothesis than their own? Participants generated more negative falsifying tests, when the hypothesis belonged to Peter (32%) than when the hypothesis was their own (7%), but this difference was not reliable, $chi^2 = 2.667$ (4), $p = .307$ as Table 5 shows. Participants intended their positive tests to falsify Peter's hypothesis a similar amount to their own hypothesis (8% vs 9% respectively, $chi^2 = 3.143$ (3), $p = .185$). Overall participants tested Peter's hypothesis less with falsifying tests (40% vs 60%), although the difference was not reliable $chi^2 = 5.25$ (5), $p = .193$. Although participants generated more negative falsifying tests than is usual in the standard 2-4-6 task, hypothesis ownership does not have a significant effect on the generation of negative tests which are expected to falsify.

Participants expected their positive tests to confirm reliably more often when the hypothesis was their own than when the hypothesis belonged to Peter (67% vs 46%, $chi^2 = 17.571$, $df = 9$, $p = .02$), as Table 3.2 shows. Participants expected their negative tests to confirm as often when the hypothesis was their own as when it belonged to Peter (17% vs 14%, $chi^2 = 4.4$, $df = 5$, $p = .246$). Overall participants tested their own untrue hypotheses with confirming tests (84%) more than the imaginary participant Peter's (60%), but this difference was not significant ($chi^2 = 15.067$ (10), $p = 0.06$). The results show that confirming your own hypothesis may be easier than confirming someone else's. Hypothesis ownership affects the generation of confirming hypothesis tests even though the relationship between the hypothesis and the experimenter's rule was not made explicit to participants (Klayman & Ha, 1987), nor was there an explicit alternative presented to participants (Klayman & Ha, 1989; Oaksford & Chater, 1994; Wason & Johnson-Laird, 1972).

Using falsification to abandon low-quality hypotheses

An important question is how participants used the falsifying and confirming test triples to reach the discovery whether they thought the untrue hypothesis they were testing was the rule or not. The results showed that more participants abandoned the untrue hypothesis when it belonged to Peter (62%) than when it was their own (38%), and fewer participants decided to abandon the low-quality hypothesis when they finished testing their own hypothesis (25%) than when they finished testing Peter's (75%), and this result was reliable ($chi^2 = 4.571$ (1), $p = .016$).

The result suggests that people not only find falsification to be possible but they also find it to be useful: they can use it to abandon untrue hypotheses. The major implication of this experiment is that the role of the hypothesis tester is not totally constrained by the mathematical properties of the problem (Klayman & Ha, 1987). The effect of hypothesis ownership on the generation of positive and negative tests, and the intention to turn these tests into confirming or falsifying ones, shows that hypothesis testing cannot be completely explained by the mathematical relationship between the hypothesis and the evidence. The hypothesis tester has their own active role both in the selection of hypothesis tests and in the interpretation of the test results, and this role cannot be completely explained by the constraints placed upon the hypothesis tester by the mathematical properties of the problem.

People appear to be able to consider how other people's hypotheses might be false, but can people falsify their own in a competitive context? The next experiment addresses this question by introducing a previously unexplored factor in hypothesis testing in the 2-4-6 task—direct competition with an opponent.

Experiment 2

Let us now test if the presence of an opponent hypothesis tester prompts participants to generate negative falsifying tests more readily when they consider an opponent hypothesis tester who is also testing their hypothesis, than when they do not consider an opponent.

Materials and design

Participants were randomly assigned to two conditions (n = 16 in each): in both conditions the incorrect embedded hypothesis was identified as belonging to the participant (Your hypothesis is 'even numbers ascending in twos'). In the experimental condition participants were given additional information about an opponent hypothesis tester ('However, an opponent called Peter is also testing 'even numbers ascending in twos'). Participants were not made aware of what the experimenter's rule was. Crucially, the relationship between the hypothesis and the true rule was identical. See Appendix B for the instructions given to participants.

Participants and procedure

Thirty two people who were members of the general student population in Trinity College, University of Dublin volunteered, and were given a minor reward of one bar of chocolate. There were 26 women and 6 men whose ages ranged from 18 years to 27 years, with a mean age of 20 years. No participants had taken courses in the philosophy of science. Participants were tested individually or in a group of up to three people.

Results and discussion

Number of triples

In total 147 triples were generated, with a mean of 4.6 triples per participant. A similar number of test triples were generated when participants were told there was an opponent hypothesis tester (M = 4.75) and when they were not (M = 4.44), (Mann-Whitney U = 125.00, Z = -.118, p = .906, two-tailed). The consideration of an opponent hypothesis tester did not encourage participants to test their hypothesis with more tests than when there was no opponent.

Positive and negative tests

Participants generated more negative tests when there was an opponent (62%) than when there was no opponent (21%), and more positive tests when there was no opponent (79%) than when there was an opponent (38%); this pattern was reliable, $chi^2 = 4.5$ (1), p = .038, two-tailed as Table _ shows.

Table 6

Table 6: Percentages of positive and negative tests generated by participants for their own hypotheses when an opponent hypothesis tester was absent or present.

	No opponent	Opponent
Positive	79	38
Negative	21	62

The result that participants generated more negative tests and fewer positive tests in situations in which the mathematical between the hypothesis and the truth were identical implies that the mathematical relationship theory cannot explain how the interaction with an opponent affects hypothesis testing (i.e., Klayman & Ha, 1987). In each case the participant owns an equally untrue hypothesis but the interaction with an opponent facilitates the generation of negative tests which can lead to falsification. The major implication is that participants play an active role in the choice of their hypothesis tests (Poletiek, 1996; 2005).

Falsification and confirmation

Do participants intend to use their positive and negative tests differently to test their hypothesis when there is an opponent? Overall participants generated a similar amount of confirming tests whether or not there was an opponent hypothesis tester (91% vs 88%), but the types of confirming tests differed. Participants generated more positive confirming tests when there was no opponent (75%) than when there was an opponent (37%), although the difference was not reliable, $chi^2 = 8.067$ (7), $p = .164$. [4]

Participants who considered the opponent hypothesis tester generated negative tests reliably more than participants who did not consider an opponent, but they intended the negative tests to confirm. Participants in the opponent condition expected their negative triples to be inconsistent with the experimenter's rule, thereby confirming their hypothesis. Participants in the opponent condition generated these negative confirming tests (54%) reliably more often than participants in the no opponent condition, (13%, $chi^2 = 11.4$ (6), $p = .039$). This result is an important one because it is similar to another result in the hypothesis testing literature on problem solving in which novice chess players tend to see only how their opponent's countermoves can make their plans work even though these moves lead to negative falsifying consequences.

Participants generated too few falsifying tests to warrant a statistical analysis. However, they did generate the same amount of negative falsifying tests whether an

[4] (This raises the possible question of power because there is a 40% difference and the p value is not significant. In fact the reason for this non-significance is a result of the degrees of freedom that are sometimes elevated because participants do not generate the same number of triples in each case of the chi square matrix; Hollander & Wolfe, 1999).

opponent hypothesis tester was present or not (8% in each case). They generated a small amount of positive falsifying tests irrespective of whether an opponent hypothesis tester was present (1% vs 4%). In both conditions the participants tested their hypothesis that belonged to themselves rather than an imaginary participant, and the rates of falsification were low. For example, participants tested their own hypothesis with falsifying tests when an opponent hypothesis tester was not present only a small amount of the time (12%). This result replicates the previous experiment which showed that participants did not falsify their own hypotheses (16% were falsifying tests).

The introduction of an opponent hypothesis tester affected the types of hypothesis testing; participants changed the type of confirming tests they performed, from positive confirming tests when there was no opponent, to negative confirming tests when there was an opponent. This result does not corroborate the uniformity theory which predicts that people experience hypothesis testing as one and the same process regardless of other factors (Poletiek, 2001; 2005) Hypothesis testing cannot be explained by mathematical theories positing that the relationship between the hypothesis under test and the experimenter's rule constrains hypothesis testing; the relationships were identical in both conditions, and participants generated different types of tests (Klayman & Ha, 1987).

Using falsification to abandon low-quality hypotheses

Somewhat more participants abandoned the low-quality hypothesis when there was an opponent (56%) than when there was no opponent (38%), and somewhat fewer participants endorsed the low-quality hypothesis when there was an opponent (44%) than when there was no opponent (62%), but the difference was not reliable, chi^2 = 1.129 (1), p = .288, two-tailed. This result may indicate that participants are somewhat better able to successfully discover that their hypothesis is untrue when an opponent hypothesis tester is introduced. By generating negative tests, participants did in fact receive falsification even though they did not expect it; but they tended to ignore this falsification. This result indicates that there may be a bias not only in the search for tests of one's hypothesis, but also in the interpretation of the test result.

The imaginary participant experiments in this chapter point out that competition affects hypothesis testing. This is a novel result and it does not corroborate the tenets of the two main alternative theories of hypothesis testing, which importantly were also developed from findings in the 2-4-6 task. What facilitates the ability to generate a negative test that has the potential to falsify a hypothesis? The hypothesis testing literature has tended to show that successful hypothesis falsification depends on the consideration of explicit alternative hypotheses. (e.g., Tweney et al., 1980). Is it reasonable to suggest that competition helps people to consider alternative hypotheses, and thus help them to generate counterexamples; in this case negative tests? Experiments 3 to 5 examines how the consideration of alternative hypotheses may help hypothesis falsification. A detailed analysis of how alternative hypotheses, explicit representation of alternative possibilities, and competition relate to one another in order to create a more detailed picture of hypothesis testing.

Experiment 3

This experiment aims to investigate if people can use negative tests to intentionally falsify a hypothesis. One way to show that people can intentionally falsify hypotheses is to investigate if they exhibit insight into the implications of particular test choices,

by understanding how they will be interpreted by the Imaginary Participant. This experiment tests whether people can intentionally falsify a low-quality hypothesis belonging to someone else (Peter) when they consider a higher quality alternative hypothesis that indicates that the hypothesis under test is false. The prediction that it may be possible to generate falsifying tests to communicate to another, because participants are presented with an alternative hypothesis which they know is the solution, was made. This alternative hypothesis may present participant's with the knowledge to intentionally falsify Peter's hypothesis in much in the same way a teacher falsifies a student's inaccurate hypothesis. The experiment also tests whether the quality of the hypothesis under test may determine the availability of confirming or falsifying evidence and hence the extent people can confirm or falsify. To this end it is constructive to compare participants testing a low-quality hypothesis belonging to Peter and compare their hypothesis testing to participants testing a high-quality hypothesis belonging to Peter (a high-quality hypothesis refers to a hypothesis that is representative of the truth).

Materials and design

One group of participants were told that Peter hypothesised that the rule was 'even numbers ascending in twos' (low-quality hypothesis), and another group were told that Peter hypothesized that the rule was 'any ascending numbers' (high-quality hypothesis , which is in fact the researcher's rule). Hypothesis quality in this instance refers to how closely the hypothesised rule fits the experimenter's rule. The experimenter's rule was the standard 'any ascending numbers'. Half of the participants in each case of hypothesis quality were given knowledge about hypothesis quality by being told the solution to the 2-4-6 problem. They were given the following additional sentence: "The experimenter's rule was in fact 'any ascending numbers'". Participants were assigned at random to one of four groups (known low-quality, unknown low-quality, known high-quality and unknown high-quality, n = 16 in each).

Hypothesis quality was defined in terms of how closely the hypothesis corresponded to the correctness of the researcher's rule. That is, the hypothesis quality was based on the number of numerical properties that corresponded to the correctness of the researcher's rule. For example, when the hypothesis was 'any ascending numbers' it was 100% correct, because it corresponded perfectly to the researcher's rule 'any ascending numbers'. When the hypothesis was 'numbers ascending in twos' it was half correct (50%), because 'ascending numbers' was one of two numerical properties that corresponded to the researcher's rule . When the hypothesis was 'even numbers ascending in twos' it was one third correct (33%), because 'ascending numbers' was one of three numerical properties that corresponded to the researcher's rule. This measure is a crude measure, but it was the logical criterion available given the constraints that: (i) the same embedded relationship between the hypothesis under test and the alternative hypothesis must be used; and (ii) approximately equivalent interval decreases in hypothesis quality should occur (see Klayman & Ha, 1989).
See Appendix C for the instructions given to participants.

Participants and procedure

The participants were 64 members of the the general public recruited through national newspaper advertisements, who were paid a nominal fee (8 euro) and undergraduate students who participated for course credits. The 50 women and 14 men were aged from 15 years to 73 years, with a mean age of 35 years. No participants had taken

courses in the philosophy of science. Participants were tested individually and in small groups of up to four individuals. The experimenter read the instructions aloud to participants (and the participant could re-read the instructions by themselves if they wished). The participants were told that they could take as long as they needed to complete the task. Most participants took approximately fifteen minutes to complete the task.

Results and discussion

Number of triples

Participants generated 343 number triples, and an average of 5.36 number triples per participant. Reliably more triples were generated for high-quality hypotheses than low-quality hypotheses (6.28 versus 4.44, Mann-Whitney $U_{32,32} = 363$, $Z = -2.025$, $p = .043$, two-tailed). There was no difference in the triples generated for hypotheses for which the quality was known or unknown (5.40 versus 5.31, Mann-Whitney $U_{32,32} = 463$, $Z = -.666$, $p = .505$, two-tailed).

There was no difference between the number of triples generated for the low-quality hypothesis for which the quality was known or unknown (4.25 versus 4.63, Mann-Whitney $U_{16,16} = 91.5$, $Z = -1.401$, $p = .171$, two-tailed). And there was no difference between the number of triples generated for the high-quality hypothesis for which the quality was known or unknown (6.56 versus 6.00, Mann-Whitney $U_{16,16} = 117.5$, $Z = -.400$, $p = .696$, two-tailed).

The results show that the fewest number of triples were not generated when participants tested a low-quality hypothesis, and knew they tested a low-quality hypothesis. This result suggests that when participants know a hypothesis is untrue, they test as much as when they do not know whether a hypothesis is true or not. The results also show that participants did not generate more test triples for high than low-quality hypotheses, and that the highest number of triples were not generated by participants who tested a high-quality hypothesis and knew it was a high-quality hypothesis. The result suggests that when a hypothesis is high-quality, people do not necessarily assume the best way forward is to confirm the hypothesis as much as possible.

Correct announcements

Participants' announcements of Peter's rule as being either correct or incorrect were calculated for the conditions in which the rule to be discovered was unknown. The percentages of correct announcements were 100% for the high-quality unknown condition and 56% for the low-quality unknown condition, and this difference was reliable, $chi^2 = 10.667$ (1), $p = .001$. As in real life where one scientist may test a significantly higher quality hypothesis than another scientist, participants considering the high quality alternative who do not know that it is the rule. They will tend to make the correct announcement because they have accumulated much confirmation and no falsification, even if they intend to falsify. When the hypothesis is low-quality more than half of the participants tested the hypothesis in such a way as to conclude that Peter's hypothesis was not the experimenter's rule. Falsifying evidence is available when a hypothesis is low-quality, and they can correctly announce that Peter's hypothesis is not the experimenter's rule.

Hypothesis quality and knowledge of hypothesis quality and hypothesis testing

Overall, more confirming triples were generated for testing high-quality (90%) than low-quality hypotheses (40%), and this difference was reliable, $chi^2 = 86.087$ (1), $p <$

.0001. Somewhat more falsifying triples were generated for low-quality (60%) than high-quality hypotheses (10%), although this difference was not reliable, chi^2 = 10.442 (1), p = .165 as Table 7 shows.

Table 7

Table 7: The percentage of confirming and falsifying triples generated for high and low quality hypothesis when quality type was known or unknown.

	Known		Unknown		Total	
	Confirm	**Falsify**	Confirm	**Falsify**	Confirm	**Falsify**
High-quality	100	**0**	80	**20**	90	**10**
Low-quality	10	**90**	70	**30**	40	**60**
Total	55	**45**	75	**25**	65	**35**

Note: The percentage of falsifying triples is presented in bold.

The percentage of falsifying triples is the mirror image of the percentage of confirming triples high-quality hypotheses more confirming triples were generated by participants who knew they were testing a high-quality hypothesis (100%) than those who did not (80%), and this difference was reliable, chi^2 = 4.308 (1), p = .038. More falsifying triples were generated by participants who did not know they were testing a high-quality hypothesis (20%) than those who did know (0%), and this difference was reliable chi^2 = 21.895 (1), p < .01, (although this p value may be elevated because zero cases were present in all cells for the known condition). The result suggests that even when a hypothesis corresponds to a true state of affairs, participants cannot be certain that it does and so they will still attempt to falsify the high-quality hypothesis in the unknown condition. In this way the knowledge that the hypothesis under test is a good one (by telling participants what the experimenter's rule is) affects confirming and falsifying in addition to the effect of hypothesis quality.

For low-quality hypotheses more confirming triples were generated by participants who did not know they were testing a low-quality hypothesis (70%) than participants who did know (10%), and this difference was reliable chi^2 = 34.322 (1), p < .0001. Critically, participants who knew they were testing a low-quality hypothesis falsified more often (90%) than those who did not know (30%), and this difference was reliable, chi^2 = 18.325 (1), p < .0001. This result does not corroborate the theory that people cannot make sense of falsification, especially when they consider better alternative hypotheses (Poletiek, 1996; 2001). Participants found it possible to intentionally generate falsifying tests in order to show that Peter's low-quality hypothesis was untrue.

Four types of hypothesis tests

When participants tested Peter's low-quality hypothesis 'even numbers ascending in twos' and they knew that the experimenter's rule was 'any ascending numbers', they generated the critical negative falsifying test 90% of the time as Table _ shows. All falsifying triples were negative falsifying triples and there were no positive falsifying triples. Every participant in this condition generated at least one negative falsifying test and announced that Peter should know from the evidence they gathered that his

hypothesis is incorrect. Participants found it possible to consistently falsify (Poletiek, 1996), even though the relationship between the hypothesis and the experimenter's rule required this difficult type of falsifying test (Klayman & Ha, 1987).

Table 8

Table 8: Percentages of confirming and falsifying positive and negative test types generated in Experiment 1.

		Low-quality		*High-quality*	
		Known	Unknown	Known	Unknown
Confirming					
	Positive	6	61	86	72
	Negative	4	9	14	8
Falsifying	Positive	0	8	0	14
	Negative	90	22	0	6

Negative falsifying triples were generated more often by participants who knew they were testing a low-quality hypothesis (90%) than in any other condition, that is, when they did not know they were testing a low-quality hypothesis (22%), when they tested a high-quality hypothesis and did not know it was high-quality (6%), or when they tested a high-quality hypothesis and did know it was high-quality (0%), $chi^2 = 46.938$ (21), $p = .0005$. Overall more negative falsifying tests were generated for low-quality hypotheses (56%) than for high-quality hypotheses (3%), $chi^2 = 24.737$ (7), $p = .0005$. Overall the generation of negative falsifying tests did not differ between conditions where the quality of the hypothesis was known (45%) and when it was unknown (14%), $chi^2 = 8.18$ (7), $p = .159$.

Even participants who tested Peter's low-quality hypothesis and did not know it was low-quality generated more negative falsifying tests (22%) than is usual in the 2-4-6 task. For example, 6% of tests were negative falsifying tests in Poletiek's falsifying condition (1996). As noted earlier in experiment 1 simply testing someone else's hypothesis helps people to falsify using negative tests (32%) more often than is standard in the 2-4-6 literature.

Positive confirming triples were generated less often by participants who tested a low-quality hypothesis and did know it was low-quality (6%), compared to when participants knew they were testing a high-quality hypothesis (86%), or when they did not know they were testing a high-quality hypothesis (72%), or when they tested a low-quality hypothesis and did not know it was low-quality (61%), $chi^2 = 63.161$ (33), $p = .0005$. Overall reliably more positive confirming tests were reliably generated for high-quality hypotheses (79%) than for low-quality hypotheses (34%), $chi^2 = 32.732$ (11), $p = .0005$. Overall the generation of positive confirming tests did not differ between conditions where the quality of the hypothesis was known (46%) andwhen it was unknown (67%), $chi^2 = 11.34$ (11), $p = .208$. (There were too few

negative confirming and positive falsifying tests in the data set to complete an objective chi-square analysis (Hollander & Wolfe, 1999)).

This result indicates that alternative hypotheses play a role in successful hypothesis falsification. One important implication that can be drawn from this finding is that the properties of the hypothesis which are generated, such as the quality of the hypothesis, cannot by themselves explain the hypothesis testing strategies people adopt. Other factors, such as the quality of the alternative hypothesis considered alongside the initial hypothesis may help explain the hypothesis testing strategies people adopt. This result does not corroborate the view that people do not know when a negative test is wise and when it is not (Klayman & Ha, 1987).

However, the result presents a situation in which some may argue that it is obvious that people can intentionally falsify. The experiment suggests that this situation is a good place to start a detailed analysis of the factors that may affect intentional falsification.

The sentence '…you know that the experimenter's rule is any ascending numbers' introduces several factors which may explain the resulting high levels of hypothesis falsification. First, participants are provided with an alternative hypothesis to consider alongside the initial hypothesis. This alternative hypothesis is not only the *correct rule* ('any ascending numbers') but is *higher quality* than the hypothesis under test ('even numbers ascending in twos'). Second, the sentence provides participants with the *knowledge that Peter's hypothesis is untrue*, because participants are told that the alternative ('any ascending numbers') is in fact the experimenter's rule. In the next experiment the role each one of these factors may play in hypothesis falsification is examined. The prediction that the knowledge of hypothesis quality affects hypothesis falsification was made, and whether the alternative hypothesis needs to be very high-quality (the actual experimenter's rule) in order for participants to falsify Peter's hypothesis was tested.

Experiment 4

The aim of this experiment was to determine what properties of an alternative hypothesis facilitate high levels of negative falsifying tests. In sum it was predicted that alternative hypotheses, particularly higher quality alternatives, facilitate the generation of negative falsifying tests and hence rule discovery.

Method

Materials and design

Participants were randomly assigned to four conditions (three experimental conditions and one control condition, n = 16 in each). In each condition they were given a low-quality hypothesis belonging to the imaginary participant Peter: 'even numbers ascending in twos'. Participants were then given another piece of information in the form of an alternative hypothesis. In the three experimental conditions participants were given one of three alternative hypotheses (high, medium and low-quality) to consider alongside the initial hypothesis that belonged to Peter. In the control condition they were given the alternative: 'in fact you know the experimenter's rule is 'any ascending numbers' (this is the replication of the known low-quality condition in Experiment 1). In the first experimental condition (high-quality alternative) participants were given a high-quality alternative hypothesis that was the experimenter's rule, but they did not know it: 'you know that another participant called James hypothesised that the experimenter's rule was any ascending numbers'.

In the second experimental condition (medium quality alternative) participants were given the medium quality alternative that was not the experimenter's rule but which was higher quality than Peter's hypothesis: 'you know that another participant called James hypothesized that the experimenter's rule was numbers ascending in twos'. In the third experimental condition (low-quality alternative) they were given a low-quality alternative that was lower quality than Peter's hypothesis: 'you know that another participant called James hypothesized that the experimenter's rule was even numbers ascending in twos that end in the digits 2,4,6' (adapted from Klayman & Ha, 1989). The instructions for the second experimental condition are given below to illustrate (see Appendix E for the instructions given to participants).

Participants and procedure

Forty eight participants completed the task (one was excluded because she said she was familiar with the task). Most participants were undergraduate students and some were individuals from the general population. The age of the participants ranged from 16 to 49 years. The mean age was 22 years, and there were 33 women and 14 men who took part. No participants had taken courses in the philosophy of science.

Results and discussion

Number of triples

A total of 245 triples was generated with a mean of 3.83 triples per participant. A mean of 3.38 triples was generated in the control condition when participants knew the alternative 'any ascending numbers' was the experimenter's rule. A mean of 4.06 triples was generated in the high-quality alternative condition when participants considered the alternative 'any ascending numbers'. A mean of 4.31 triples was generated in the medium quality alternative condition when participants considered the alternative 'numbers ascending in twos'. A mean of 3.56 triples was generated in the low-quality alternative condition when participants considered the alternative 'even numbers ascending in twos'. Somewhat fewer triples were generated by participants in the control condition who knew that 'any ascending numbers' was the experimenter's rule (M = 3.38) than participants in the high-quality alternative condition who did not know it was the experimenter's rule (M = 4.06), but this difference was not reliable (Mann-Whitney$_{16,16}$ = 96.5, Z = -1.204, p = .118). Somewhat fewer triples were generated in the control condition (M = 3.38) than in the medium quality alternative condition (M = 4.31), and but this difference was not significant (Mann-Whitney$_{16,16}$ U = 87.5, Z = -1.563, p = .064, two-tailed). There was little difference in the mean number of triples generated in the control condition (M = 3.38) and in the low-quality alternative condition (M = 3.56), and the difference was not reliable (Mann-Whitney$_{16,16}$ U = 114.00, Z = -.536, p = .308). There was no difference for the number of triples generated in the high-quality alternative condition (M = 4.06) and in the medium quality alternative condition (M = 4.31, Mann-Whitney$_{16,16}$ U = 111.00, Z = -.658, p = .268). There was a marginal difference in the number of triples generated for the medium quality alternative condition (M = 4.31) and in the low-quality alternative condition (M = 3.56, Mann-Whitney$_{16,16}$ U = 88.5, Z = -1.522, p = .069). These results imply that the quality of the alternative hypothesis may sometimes affect the number of tests participants generated when testing a low-quality hypothesis. There was a small indication that participants could have a tendency to test fewer triples in the control condition because they are sure that Peter's hypothesis is untrue and that their test falsifies his hypothesis.. This result replicates the same finding reported in Experiment 1. And there was a small

indication that participants could have a tendency to test fewer triples in the low-quality alternative condition than in the other experimental conditions, perhaps because the consideration of an alternative that is even lower quality than their hypothesis may constrain their ability to generate other possible test triples or alternatives. Participants test more when the alternative hypothesis is higher quality than the hypothesis under test and they do not know that the alternative is higher quality, perhaps indicating that there is a tendency to test more once a falsification is achieved.

Correct announcements and rule discovery

The experiment predicted that as the quality of the alternative hypothesis decreased the number of correct announcements would decrease, that is, the number of participants who would announce that Peter's low-quality hypothesis 'even numbers ascending in twos' was incorrect would decrease.[5] The alternative hypothesis may present the participant with an explicit set of possibilities from which to generate falsifying tests. If participants are using the alternative hypothesis to generate test triples such as 5-11-22 when they consider the alternative hypothesis 'any ascending numbers' they cannot falsify Peter's hypothesis 'even numbers ascending in twos' when they receive a 'yes' from the experimenter, but they may conclude that the alternative hypothesis is necessarily the experimenter's rule.

Participants in the high-quality alternative condition announced that Peter's hypothesis was not the rule almost as often (81%) as participants in the medium quality condition (94%), but less often when they were presented with the low-quality alternative hypothesis (69%). This difference was not significant, $chi^2 = 3.282$ (2), $p = .097$, two tailed). Participants discovered what the experimenter's rule was more often in the high-quality alternative condition (50%) and in the medium quality alternative condition (44%), than in the low-quality alternative condition (12%, $chi^2 = 5.647$ (2), $p = .03$).

Table 9

Table 9: The percentages of participants who discovered the experimenter's rule.

	High-quality	Medium quality	Low-quality
Rule discovered	50	44	12

The result implies that even when one of the hypotheses under consideration is correct participants may not always discover that it is correct (50%). Moreover, even when participants consider a medium quality alternative hypothesis they can sometimes discover the rule (44%). Participants who considered a lower quality alternative hypothesis rarely discovered the rule (12%). The implication is that it is not enough to consider two alternative hypotheses to discover the rule; discovery may depend on considering at least one good quality hypothesis (e.g., Tweney et al., 1980).

[5] The control condition is not relevant to this section because participants know what the experimenter's rule is. We compare the three experimental conditions only.

Alternative hypothesis quality and hypothesis testing

The experiment predicted that participants would falsify more when the alternative hypothesis was higher quality, and that as the alternative decreased in quality participants would confirm more. Although participants confirmed somewhat more in the low-quality condition (67%), compared to the medium quality condition (52%) and in the high-quality condition (49%), the differences were not reliable, chi^2 = 13.017 (12), p = .184. As predicted participants falsified more in the high-quality condition (51%), and in the medium quality condition (48%), than in the low-quality condition (33%), chi^2 = 20.323 (10), p = .013. The results imply that as the quality of the alternative hypothesis decreases the amount of falsification decreases. High-quality alternative hypotheses facilitate falsification of low-quality hypotheses.

Four types of hypothesis tests

Participants falsified reliably more often with negative falsifying tests in the high-quality condition (42%), and in the medium quality condition (48%), than in the low-quality condition (23%), chi^2 = 22.167 (10), p = .007, as Table 2.7 shows. Participants confirmed somewhat more often with positive confirming tests in the low-quality condition (44%), than in the medium quality condition (20%), or in the high-quality condition (23%), but this difference was not reliable, chi^2 = 7.725 (10), p = .328.

A similar amount of negative confirming was observed in the high-quality condition (26%), as in the medium quality condition (32%), and in the low-quality condition (23%), chi^2 = 6.686 (10), p = .378. There were too few positive falsifying tests in the data set to justify a statistical analysis (See Siegel and Castellen, 1994). The results imply that the quality of the alternative hypothesis does not have a strong affect on the amount of negative falsifying triples. Regardless of the quality of the alternative hypothesis, negative falsifying tests were generated.

Knowledge of alternative hypothesis quality

Participants generated falsifying tests when they knew the alternative hypothesis was the experimenter's rule (61%) and when they did not know (51%), and this difference was not reliable, chi^2 = 7.244 (6), p = .15. The result that the majority (61%) of the tests were falsifying when participants knew the alternative was the experimenter's rule replicates our finding in Experiment 1, although the effect in this experiment was not as large.

Participants confirmed somewhat less often when they knew the alternative hypothesis was the experimenter's rule (39%) than when they did not know (49%), but this was not reliable, chi^2 = 3.352 (5), p = .323. Negative falsifying tests were generated somewhat more often when participants knew the alternative was the experimenter's rule (61%) than when they did not (42%), but this difference was not reliable, chi^2 = 6.819 (6), p = .169. Positive confirming tests were generated as often when participants knew the alternative was the experimenter's rule (33%) than when they did not know (23%), chi^2 = .400, (4), p = .491. More negative confirming tests were generated when participants did not know the alternative was the experimenter's rule (26%) than when they did know (6%), and this difference was marginally reliable, chi^2 = 7.133 (4), p = .065. The knowledge that the alternative hypothesis is the experimenter's rule has a small effect on hypothesis testing, but the consideration of a higher quality alternative hypothesis may be the clearest predictor that people will falsify a low-quality hypothesis.

The results do not corroborate the view that falsification is impossible; people can falsify when they consider a higher quality alternative hypothesis (Poletiek, 1996). However, the higher quality alternative may have given participants information that made relationship between Peter's hypothesis and the truth explicit (Klayman & Ha, 1987). The next experiment examines whether the alternative needs to be explicit and shown to embed Peter's hypothesis in order for participants to falsify.

Experiment 5

The aim of this experiment was to investigate whether hypothesis falsification is facilitated by presenting participants with an alternative explicit set of possibilities from which to generate the negative falsifying tests. Our previous experiments found that falsification was facilitated by the consideration of an alternative hypothesis, but each alternative hypothesis stated explicit numerical properties, such as 'even numbers ascending in twos' from which a negative test triple such as 3-5-7 could be generated. This experiment examines whether people can generate negative falsifying tests when they consider a non-explicit hypothesis such as 'something else'. Counter to Klayman and Ha (1987; 1989) who suggest that people may be able to generate negative tests in the standard 2-4-6 task when the relationship is made explicit to them, we predict that the consideration of a non-explicit alternative hypothesis prompts people to generate their own alternatives from which to generate negative falsifying triples.

Method

Materials and design

Participants were randomly assigned to three conditions (n = 16 in each). In each condition they were given a low-quality hypothesis belonging to the imaginary participant Peter: 'even numbers ascending in twos'. In the first condition they were given an explicit alternative hypothesis: 'Another participant called James hypothesised that the experimenter's rule was any ascending numbers'. In the second condition they were given a non-explicit alternative: 'Another participant called James hypothesised that the experimenter's rule was something else'. In the third condition they were given no alternative at all. See Appendix F for the instructions given to participants.

Participants and procedure

Forty eight participants completed the task. They were undergraduate students who gained course credit for their participation. Their age ranged from 17 to 49 years and the mean age was 21 years. There were 33 women and 15 men who took part. No participants had taken courses in the philosophy of science. The recording sheet and procedure were the same as in Experiment 1.

Results and discussion

Number of triples

A total of 222 triples were generated. A mean number of 4.63 triples were generated per participant. There was no difference in the number of triples generated when the alternative was explicit (M = 4.75) and non-explicit (M = 4.81, Mann-Whitney$_{16,16}$ U = 122.5, Z = -.210, p = .417). There was no difference in the number of triples generated when the alternative was non-explicit (M = 4.81) andwhen there was no

alternative (M = 4.31, Mann-Whitney$_{16,16}$ U = 109, Z = -.7224, p = .469, two-tailed). And there was no difference in the number of triples generated when the alternative was explicit (M = 4.75) andwhen there was no alternative (M = 4.31, Mann-Whitney$_{16,16}$ U = 106, Z = -.840, p = .401, two-tailed). The result implies that neither the consideration of nor the explicitness of an alternative hypothesis, affects how much people test their hypothesis.

Correct announcements and rule discovery

Participants announced correctly that Peter's low-quality hypothesis 'even numbers ascending in twos' was not the experimenter's rule somewhat less often when they were presented with the explicit alternative (69%), than the non-explicit alternative (81%), or no alternative (81%), but this difference was not reliable (chi^2 = 0.943 (2), p = 0.312). The rate of correctly announcing that Peter's hypothesis is not the experimenter's rule appears to be elevated in this experiment compared to the previous experiment. Nonetheless the first condition (50% discovered the rule) replicates the result of the same condition in Experiment 2 (50% also discovered the rule), suggesting there were no new extraneous variables.

Participants in this experiment were asked what they thought the experimenter's rule was once they announced Peter's low-quality hypothesis 'even numbers in twos' was incorrect. The rate of rule discovery was highest when participants considered the explicit alternative 'any ascending numbers' (50%), than the non-explicit alternative 'something else' (31%), or when there was no alternative (19%), and this difference was reliable (chi^2 = 5.101 (2), p = .039).

The results suggest that the discovery of the rule appears to depend on the consideration of an explicit high-quality alternative hypothesis. Falsification and the consideration of an alternative that is both explicit and high-quality may go hand in hand to facilitate rational hypothesis testing (Wason & Johnson-Laird, 1972; Kuhn, 1996).

Confirming and falsifying

More confirming triples were generated when the alternative was explicit (57%), than when it was the non-explicit (47%), or when there was no alternative (46%), but this difference was not significant (chi^2 = 28.374 (16), p = .058).

There was no difference in the amount of falsifying triples generated when the alternative was explicit (43%), than non-explicit (53%), and when there was no alternative (54%), (chi^2 = 10.044 (16), p = .216). It is not clear from this result if the consideration of explicit and non-explicit alternatives help people to falsify. Falsifying was found in each condition even when there was no alternative. As noted earlier, it is possible that simply considering someone else's hypothesis helps a participant to falsify.

Four types of hypothesis tests

More positive confirming tests were generated when the alternative was explicit (37%), than when it was non-explicit (22%), or when there was no alternative at all (27%), but this was not reliable, chi^2 = 11.379 (12), p = .249. There was no difference in the amount of negative confirming tests generated when the alternative was explicit (20%), than when it was non-explicit (25%), than when there was no alternative (19%), chi^2 = 9.128 (8), p = .166).

Table 10

Table 10: The percentages of positive and negative confirming and falsifying triples generated when the alternative hypothesis was explicit, non-explicit, and when there was no alternative.

		Explicit	Non-explicit	No alternative
Confirming				
	Positive	37	22	27
	Negative	20	25	19
Falsifying	Positive	8	8	12
	Negative	35	45	42

There was no difference in the amount of positive falsifying tests generated when the alternative was explicit (8%), than when it was non-explicit (8%), or when there was no alternative (12%, the number of cases was not large enough to carry out a reliable chi-square test). There was no difference in the amount of negative falsifying tests generated when the alternative was explicit (35%), than when the alternative was not explicit (45%), or when there was no alternative (42%, $chi^2 = 12.875$ (14), $p = .268$). The results are important not only because they replicate the results of our earlier experiments to show that negative falsifying is possible more often than the literature has ever shown, but they imply that the consideration of an alternative need not necessarily be explicit in order to falsify using a negative falsifying test.

General Discussion

The experiments revealed that people find it possible to falsify an incorrect hypothesis that is typical of the standard 2-4-6 task. The introduction of an imaginary participant to the 2-4-6 task led to several novel and important findings for hypothesis testing. Experiment 1 reported the novel result that participants find it possible to generate the negative tests that have the potential to lead to falsification of a hypothesis. Participants could consistently generate negative tests of a hypothesis that belonged to *someone else* rather than their own equally untrue hypothesis in the 2-4-6 task. While the negative tests of the imaginary participant's hypothesis in Experiment 1 were not significantly intended to falsify there were many more intentionally falsifying instances than has previously been shown in the literature (e.g., Poletiek, 1996). Experiment 2 showed the novel result that participants could this time consistently generate negative tests of *their own* hypothesis when they imagined an opponent hypothesis tester who was also trying to discover the rule. But participants tended only to see how these negative tests had the potential to confirm their hypotheses. They could not anticipate how these negative tests could show them to be wrong.

One explanation is that considering an Imaginary Participant and an Imaginary Opponent prompt the consideration of instances outside of what is presently being considered such as alternative hypotheses. Experiment 3 showed that participants intentionally falsified Peter's untrue hypothesis with negative tests when they considered an alternative hypothesis that made the researcher's rule explicit.

Participants consistently overcame their tendency to test a hypothesis with positive tests when it was more accurate to test with negative tests and they predicted that these negative tests would lead to falsification.

A number of different factors were separated out to examine what facilitate the falsification observed. Experiment 4 found that participants did not necessarily need to know that the alternative was the experimenter's rule in order to falsify. They intentionally falsified Peter's hypothesis as often when they considered the experimenter's rule, and did not know it was the experimenter's rule, as when they did know. They also falsified as often when the alternative was higher quality than Peter's hypothesis, even though it was not as high in quality as the experimenter's rule. Participants discovered the rule as often when the alternative was higher quality, regardless of whether it was the experimenter's rule or not. When the alternative was lower quality than Peter's hypothesis it led to falsification, but participants were not able to use this falsification to discover the experimenter's rule. The major implication of this result is that falsification of a hypothesis can be facilitated by the consideration of higher and lower quality alternative hypotheses, but falsification in light of a higher quality alternative leads to the discovery of the rule.

Experiment 5 produced the novel result that participants intentionally falsified as often when the alternative was explicit and non-explicit, and when there was no alternative. But participants reliably discovered the rule more often when the alternative was explicit than non-explicit, and than when there was no alternative at all. The major implication of this finding is that falsification is sufficient to announce that a hypothesis is untrue, but perhaps an explicit alternative hypothesis that explains the falsifying result is necessary for truth discovery.

The results do not corroborate the mathematical relationship theory that asserts participants have a tendency to engage in a positive test strategy in the hypothesis testing situations they encounter (Klayman & Ha, 1987). In our experiments participants knew when it was accurate to test a hypothesis with a negative test; when they considered an alternative hypothesis they could often reliably generate negative tests and they reliably expected them to falsify (Klayman & Ha, 1987). The prediction that participants need to know what the mathematical relationship between the hypothesis and the truth is in order to generate negative tests was not supported by our results. Participants generated negative tests and expected these tests to falsify when they considered a non-explicit hypothesis telling them nothing about what the relationship between the hypothesis and the rule was (Klayman & Ha, 1989).

The results also do not corroborate the prediction that participants find falsification impossible; participants not only generated negative tests but they intended these negative tests to falsify. They showed that they understood the implications of their test choice by predicting that Peter would know from their negative falsifying tests that his hypothesis was incorrect (Poletiek, 1996).

The consideration of a non-explicit alternative could not have made the mathematical relationship between the hypothesis under test and the truth directly explicit to participants. Yet participants generated negative falsifying tests and intended them to falsify when they considered a non-explicit alternative hypothesis more often than has been usual in the hypothesis testing literature (for a review see Poletiek, 2001). Counter to alternative hypotheses accounts based on the premise of considering two alternative explicit hypotheses, the non- explicit alternative did not hamper the generation of falsifying tests, or negative falsifying triples (Oaksford & Chater, 1994). Counter to the mathematical relationship theory participants may be able to make the relationship explicit for themselves simple by thinking that the truth

is something other than what they presently consider (Klayman & Ha, 1987; 1989). Although the generation of negative falsifying tests did not depend on the consideration of an explicit alternative, it is possible that participants subsequently fleshed out the non-explicit alternative to generate their own explicit alternative (e.g., Byrne, 2005). Furthermore the theoretical view that participants find it difficult to intentionally falsify a low-quality hypothesis because there is no new information available to them, cannot offer a complete explanation either (Poletiek, 1996; 2001). A non-explicit hypothesis does not give participants new information, but it may encourage them to search for new evidence by either generating their own negative tests or alternative hypothesis. The results imply that falsification *and* the consideration of alternative hypotheses that are higher quality than the hypothesis under test, may go hand in hand in discovering the truth in hypothesis testing (Wason & Johnson-Laird, 1972). The falsifying test is only any good if it leads to the endorsement of an explicit alternative that is higher quality than the quality of the hypothesis under test. For example, in scientific reasoning a theory is sometimes falsified, but unless there is an explicit alternative theory to explain the falsifying result, the falsification remains an anomaly until such a time as a new theory is generated (see Kuhn, 1993).

Implications for theories of reasoning

Our results have important implications for theories of reasoning. The results on falsification are important to the consideration of negative information in reasoning, and the results on the consideration of alternative hypotheses are important for understanding the consideration of alternatives in reasoning in general.

Hypothesis falsification implies that people can think about negative instances (e.g., Klayman & Ha, 1987; Kareev & Halberstadt, 1993; Vallee-Tourangeau et al., 1995). Consider when someone is asked to think about the statement 'If Sharon is in Spain, then Justina is in Holland', and they encounter a piece of information that is inconsistent with this statement such as 'Justina is not in Holland', they can deduce that 'Sharon is not in Spain'. This type of inference is called *Modus Tollens* and it has been investigated extensively in the literature on deductive reasoning (e.g., Byrne, 1989; Byrne & Tasso, 1994), and is logically equivalent to hypothesis falsification (e.g., Popper, 1959; Klayman & Ha, 1987).

To consider the inference people may construct a *counterexample* that is a possibility which is inconsistent with the possibility currently under consideration. A counterexample may be similar to a refutation of a theory in science (Kuhn, 1993). Little is known about how people search for counterexamples when they reason (e.g., Byrne, Espino, & Santamaria, 1999), and the results of our experiments in which people imagine another participant or opponent show that it may be important to further investigate the ecologically valid circumstances which prompt people to explicitly mentally represent what is being reasoned about as fully as possible in order to discover counterexamples. Otherwise deductive errors could be made, such as concluding 'nothing follows' when you are told 'Justina is not in Holland' (e.g., Johnson-Laird & Byrne; Johnson-Laird & Byrne, 2002).

The experimental analysis in the imaginary participant 2-4-6 task showed that falsification by itself could not be predicted by how explicit the alternative hypothesis was. Yet the representation of an alternative hypothesis as explicit was critical in the use of falsification to abandon an untrue hypothesis. This condition may parallel scientific reasoning that tends not to abandon a falsified theory unless a viable alternative theory presents a better explanation (e.g., Kuhn, 1993), or labels

falsification as an anomaly until a better alternative theory is generated (e.g., Koslowski, 1996).

Implications for hypothesis testing

The results have several implications for current theories of hypothesis testing. First, the effect of competition in hypothesis testing corroborates the separation of falsification into falsifying one's own hypothesis, and falsifying someone else's hypotheses (Poletiek, 2005).

It is possible that when the testing of a hypothesis leads to an encounter with evidence to prove that the hypothesis is false, it may lead to the generation of new knowledge (Popper, 1963). In scientific terms a falsification of theory is termed a refutation (Kuhn, 1993). When theories are refuted either an alternative theory which explains the result is accepted as superior, or an alternative theory is developed which can explain the falsifying result (e.g., Wason & Johnson-Laird, 1972; Kuhn, 1993). A theory is revised to incorporate the new result rather than abandoned altogether (Howson & Urbach, 1993; Klayman & Ha, 1989; Kowslowki, 1996), or occasionally the refutation is labelled as an anomaly until a viable alternative theory is generated (Kuhn, 1993; Koslowski, 1996).

Refutations are generated by rival theorists (e.g., Mitroff, 1974; Kuhn, 1993), and to safeguard against many refutations being labelled anomalies by scientists who disagree with one another it is important to test hypotheses with specific alternatives in mind (e.g, Platt, 1964). For example, successful hypothesis testers who use falsification to overcome hypotheses which are untrue, consider at least one alternative hypothesis in rule discovery tasks (e.g., Klayman & Ha, 1989). Identifying falsifying evidence indicates what is wrong with a hypothesis or theory (e.g, Fugelsang et al., 2004). Falsification drives hypothesis revision because it hints at what should be incorporated into the hypothesis. When we encounter inconsistent evidence relevant to a current state of knowledge we may update our knowledge by revising it to include the new piece of information (e.g., Gardenfors, 1988; Harman, 1986). Falsification is possible in competitive contexts which promote the consideration of alternative hypotheses. Thus falsification ensures that theories which have outlived their usefulness are either improved or abandoned in favour of theories which offer better explanations (Popper, 1963).

The research leads to an important future question. How can the competition and the consideration of alternative hypotheses make it possible for people to seek out negative evidence to challenge their own hypotheses which they *believe* to be true (Popper, 1959; Wason, 1960)? Trying to confirm again what we already believe can lead to the maintenance of incorrect ideas, such as those concerned with prejudiced stereotyping (e.g., Snyder & Swan; 1978). Consider prejudiced hypotheses in which the prejudice tends to be embedded within the truth identical to the standard 2-4-6 context. For example, a prejudice about an ethnic minority, in which there is a collection of a small set of instances with a negative connotation. If we consider a contemporary example such as someone in Northern Ireland held the prejudiced belief that all Catholics were involved in paramilitary activities they may cite cases which are consistent (positive instances) with this belief, such as a person with a criminal record for paramilitary activity who was also Catholic, and avoid any inconsistent instances (negative instances) which exist outside of their collection of confirming evidence (Mallie, 2001). But one falsifying case can prove that this prejudiced belief is false. The standard version of the 2-4-6 task, when the participant's hypothesis is

embedded within the true rule, is analogically equivalent to this prejudiced belief (Wason, 1960).

When people compete with an opponent hypothesis tester the competition may help them create other salient alternative possibilities, or the competition may facilitate the need to use negation to falsify an opponent's hypothesis.

People tend to think of few possibilities in their reasoning because their working memory is limited (Johnson-Laird & Byrne, 2002). When people test hypotheses they often represent only one hypothesis at a time in working memory (Mynatt, Doherty, & Dragan, 1993), but when they compete with an opponent hypothesis tester they may represent two possibilities; their own hypothesis and the opponent's hypothesis. These possibilities may not necessarily correspond to false possibilities, but two possibilities that may be true (e.g., Tweney et al., 1980; Johnson-Laird & Byrne, 1991). Competitive hypothesis testing may provide a forum in which people can consider two possibilities, their own hypothesis and their opponent's hypothesis, and the difficulty of representing two possibilities by oneself and falsifying one's own hypothesis may be slightly less. Third, competition may help participants to be better at making possible alternative hypotheses explicit for themselves in their mental representations of hypotheses, and the alternative set of possibilities may help them to generate negative falsifying test triples. Perhaps with competition participants may understand that there are alternative hypotheses that an opponent hypothesis tester may be considering. Even though the alternatives belonging to an opponent are non-explicit, the competition may prompt participants to flesh out these properties to consider what the opponent's alternatives might be.

In conclusion, the Imaginary Participant experiments in this paper show how reasoning with falsification facilitates the comparison of internal thoughts with external facts allowing us to interact with the world in a way that reflects reality. Thus falsification may present us with one of the cornerstones of enlightened thinking associated not only with scientific progress, but to the deep insights associated with educational excellence, and a free thinking society that asks questions and challenges prejudices.

References

Allport, G. W. (1979). The nature of prejudice (2nd Ed.). London: Addison-Wesley.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking.* New York: John Wiley and Sons.

Byrne, R. M. J. (2005). *The Rational Imagination*. Cambridge, MA: MIT Press.

Byrne, R. M. J., Espino, O., & Santamaria, C. (1999). Counterexamples and the suppression of inferences. *Journal of Memory and Language, 40,* 347-373.

Byrne, R. M. J., & Walsh, C. A. (2005). Resolving contradictions. In V. Girotto & P. N. Johnson-Laird (Eds.), *The Shape of Reason: Essays in honour of Paolo Legrenzi*, 91-105. Sussex, UK: Psychology Press.

Cherubini, P., Castelvecchio, E., & Cherubini, A. M. (2005). Generation of hypotheses in Wason's 2-4-6 Task: An information theory approach. *The Quarterly Journal of Experimental Psychology,* forthcoming.

Christensen-Szalansky, J. J. J., & Bushyhead, J B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 928-935.

Cowley, M., & Byrne, R. M. J. (2004). Chess Masters' Hypothesis Testing. In K. D. Forbus, D. Gentner, & T. Rogers (Eds.). *Proceedings of the Twenty- Sixth*

Annual Conference of the Cognitive Science Society. pp. 250- 255. Mahwah, NJ: Erlbaum.

Cowley, M. & Byrne, R. M. J. (2005). When Falsification is the Only Path to Truth. In B. G. Bara, L. Barsalou, & M. Bucciarelli (eds.). *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society*. pp. 512-517. Mahwah, NJ: Erlbaum

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review, 85*, 386-416.

Evans, J. St. B. T. (1989). *Bias in Reasoning*. Hove, UK: Erlbaum.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human Reasoning: The Psychology of Deduction*. Hove, UK: Lawrence Erlbaum.

Farris, H., & Revlin, R. (1989). The discovery process: A counterfactual strategy.

Social Studies of Science, 19, 497-513.

Frank, A. (2001). *The diary of a young girl*. London: Penguin books.

Fugelsang, J., Stein, C., Green, A., & Dunbar, K. (2004). Theory and data interactions of the scientific mind: Evidence from the molecular and the cognitive laboratory. *Canadian Journal of Experimental Psychology, 58*, 1392-1411.

Gärdenfors, P. (1988). *Knowledge in flux*. Cambridge, MA: MIT Press.

Gale, M., & Ball, L. J. (2003). Facilitation of rule discovery in Wason's 2-4-6 task: The role of negative triples. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society*, 438-443. Boston, MA: Cognitive Science Society.

Gale, M., & Ball, L. J. (2005). Dual-goal facilitation in Wason's 2-4-6 task: What mediates successful rule discovery. *The Quarterly Journal of Experimental Psychology, 59*, 1-13.

Girotto, V., Legrenzi, P., Rizzo, A. (1991). Event controllability in counterfactual thinking. *Acta Psychologica, 78*, 111-133.

Gorman, M. E. (1995a). Confirmation, disconfirmation and invention: The case of Alexander Graham Bell and the telephone. *Thinking and Reasoning, 1*, 31-53.

Gorman, M. E. (1995b). Hypothesis Testing. In S. E. Newstead & J. St. B. T. Evans (Eds.), *Perspectives on thinking and reasoning: Essays in honour of Peter Wason.* Hove, UK: Laurence Erlbaum Associates Ltd.

Gorman, M. E. & Gorman, M. E. (1984). A comparison of disconfirmatory, confirmatory and a control strategy on Wason's 2-4-6 task. *Quarterly Journal of Experimental Psychology, 36A*, 629-648.

Gorman, M. E., Gorman, M. E., Latta, R. M., & Cunningham, G. (1984). How disconfirmatory, confirmatory, and combined strategies affect group problem solving. *British Journal of Psychology, 75*, 65-79.

Harman, G. (1986). *Change in View*. Cambridge, MA: MIT Press. .

Hedden, T., & Zhang, J. (2002). What do you think I think you think? Strategic reasoning in matrix games. *Cognition, 85*, 1-36.

Hollander, M., & Wolfe, D.A. (1999). *Nonparametric statistical methods*. New York: John Wiley & Sons.

Howson, C., & Urbach, P. (1993). *Scientific reasoning* (2nd Ed.). Chicago: Open Court.

Johnson-Laird, P. N. (1983). *Mental Models.* Cambridge: Cambridge University Press.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hove, UK: Lawrence Erlbaum Associates.

Johnson-Laird, P. N., & Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review, 109*, 646-678.

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from Inconsistency to Consistency. *Psychological Review, 111, 3*, 1-23.

Kareev, Y., & Halberstadt, N. (1993). Evaluating negative tests and refutations in a rule discovery task. *Quarterly Journal of Experimental Psychology, 46A*, 715-727.

Klayman, J., & Ha, Y-W. (1987). Confirmation, Disconfirmation, and Information in Hypothesis Testing. *Psychological Review, 94, 2*, 211-228.

Klayman, J., & Ha, Y-W. (1989). Hypothesis Testing in Rule Discovery: Strategy, Structure and Content. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15 (4)*, 596-604.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 107-118.

Koslowski, B. (1996). *Theory and Evidence: The development of scientific reasoning*. Cambridge, MA: MIT Press.

Kruglanski, A. W., & Webster, D. M. (2000). Motivated Closing of the Mind: "Seizing" and "Freezing". In E. T. Higgins & A. W. Kruglanski (Eds.), *Motivational Science: Social and personality perspectives*, 354-375. USA: Taylor & Francis.

Kuhn, T. S. (1993). *The Structure of Scientific Revolutions.* (3rd Ed.). Chicago: Chicago University Press.

Kunda, Z. (1987). Motivation and inference: Self-serving generation and evaluation of evidence. *Journal of Personality and Social Psychology, 53*, 636-647.

Kunda, Z. (2000). The Case for Motivated Reasoning. In E. T. Higgins & A. W. Kruglanski (Eds.), *Motivational Science: Social and personality perspectives*, 313-335. USA: Taylor & Francis.

Molloy, A. M., & Scott, J. M. (2001). Folates and prevention of disease. *Public Health Nutrition (Review), 4(2b),* 601-609.

Mynatt, C. R., Doherty, M. E., & Dragon, W. (1993). Information relevance, working memory, and the consideration of alternatives. *The Quarterly Journal of Experimental Psychology,* 46A, 759-778.

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.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1978). Consequences of confirmation and disconfirmation in a simulated research environment. *Quarterly Journal of Experimental Psychology, 30*, 395-406.

Oaksford, M., & Chater, N. (1994). Another look at eliminative behaviour in a conceptual task. *European Journal of Cognitive Psychology, 6*, 149-169..

Poletiek, F. H. (1996). Paradoxes of Falsification. *The Quarterly Journal of Experimental Psychology, 49*, 447-462.

Poletiek, F. H. (2001). *Hypothesis Testing Behaviour*. UK: Psychology Press.

Poletiek, F. (2005). The proof of the pudding is in the eating: Translating Popper's philosophy into a model for testing behaviour. In K. I. Manktelow (Ed.). *Historical and Theoretical Perspectives on Reasoning*, forthcoming.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.

Popper, K. R. (1963/1978). *Conjectures and Refutations* (4th ed.). London: Routledge

and Kegan Paul.
Roese, N. J., & Olson, J. M. (1995). *What might have been: The social psychology of counterfactual thinking*. Hillsdale, NJ: Lawrence Erlbaum.
Snyder, M., & Swann, W. B., Jr. (1978). Hypothesis-testing in social interaction. *Journal of Personality and Social Psychology, 36*, 1202-1212.
Tukey, D. D. (1986). A philosophical and empirical analysis of subject's modes of inquiry in Wason's 2-4-6 task. *Quarterly Journal of Experimental Psychology, 38*, 5-33.
Tweney, R. D., Doherty, M. E., Worner, W. J., Pliske, D. B., Mynatt, C. R., Gross, K. A. & Arkkelin, D. L. (1980). Strategies of rule discovery on an inference task. *Quarterly Journal of Experimental Psychology, 32*, 109-123.
Vallee-Tourangeau, F., Austin, N. G., & Rankin, S. (1995). Inducing a rule in Wason's 2-4-6 task: A test of the information-quantity and goal complementarity hypotheses. *Quarterly Journal of Experimental Psychology, 48A*, 895-914.
Van der Henst, J. B., Rossi, S., & Schroyens, W. (2002). When participants are not misled they are not so bad after all: A pragmatic analysis of a rule discovery task. *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society*, 902-907. Mahwah, NJ: Erlbaum.
Waganaar, W. A., van Koppen, P. J., & Crombag, H. F. (1993). *Anchored narratives: The psychology of criminal evidence*. London: Harvester Wheatsheaf.
Wason, P. C. (1960). On the failure to eliminate hypothesis in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*, 129-140.
Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content*. London: Batsford.
Wetherick, N. E. (1962). Eliminative and enumerative behaviour in a conceptual task. *Quarterly Journal of Experimental Psychology, 14*, 246-249.
Wharton, C. M., Cheng, P. W., & Wickens, T. D. (1993). Hypothesis-testing strategies: Why two goals are better than one. *Quarterly Journal of Experimental Psychology, 46A*, 743-758.

Appendices:

Appendix A: Instructions used in experiment 1

“In a previous study investigating human thinking a participant called Peter was asked to discover a rule a researcher had in mind that the number sequence 2,4,6 conforms to. Peter hypothesised that the experimenter’s rule was: “even numbers ascending in twos”.

Your aim is to go about testing if Peter’s rule “even numbers ascending in twos” is the experimenter’s rule. You are to do this by writing down other number sequences with sets of three numbers. You will then be informed if these number sequences conform or do not conform to the rule the researcher has in mind.

You should try to go about testing if Peter’s rule “even numbers ascending in twos” is the rule the researcher has in mind by citing as few number sequences as you can. Please note that you have three pages on which to test your number sequences if you need to. When you feel highly confident that you have discovered if Peter’s rule is the experimenter’s rule, *and not before*, you are to write down “Peter’s rule is the experimenter’s rule” or “Peter’s rule is not the experimenter’s rule”. You are to write this under your most recent number sequence. The experimenter will then write whether or not you are correct beside your announcement.

The words ‘your’ and ‘you’ replaced the words ‘Peter’s’ and ‘Peter’ respectively for the condition where the hypothesis belonged to the participant themselves.

Appendix B: Instructions used in Experiment 2

“In a previous study investigating human thinking you were a participant who was asked to discover a rule a researcher had in mind that the number sequence 2,4,6 conforms to. You hypothesised that the experimenter’s rule was: “even numbers ascending in twos”.

Your aim is to go about testing if your rule “even numbers ascending in twos” is the experimenter’s rule. However, an opponent called Peter is also testing “even numbers ascending in twos”. You must discover if “even numbers ascending in twos” is the experimenter’s rule before he does.

You are to do this by writing down other number sequences with sets of three numbers. You will then be informed if these number sequences conform or do not conform to the rule the researcher has in mind. Please remember your aim is specifically to test if your original rule “even numbers ascending in twos” is the experimenter’s rule, and not to test any new ideas of your own that you think the experimenter’s rule might be.

Please note that you have three pages on which to test your number sequences if you need to. When you feel highly confident that you have discovered if your rule is the experimenter’s rule, *and not before*, you are to write down “My rule is the experimenter’s rule” or “My rule is not the experimenter’s rule”. You are to write this under your most recent number sequence. The experimenter will then write whether or not you are correct beside your announcement.”

Appendix C: Instructions used in experiment 3

"In a previous study investigating human thinking a participant called Peter was asked to discover a rule a researcher had in mind that the number sequence 2,4,6 conforms to. Peter hypothesised that the experimenter's rule was: 'even numbers ascending in twos'. The experimenter's rule is in fact 'ascending numbers'.

Your aim is to go about testing Peter's rule 'even numbers ascending in twos' in a way you think would help him to discover if his rule is the experimenter's rule. You are to do this by writing down other number sequences with sets of three numbers. You will then be informed if they conform or do not conform to the rule the researcher has in mind.

You should try to go about testing Peter's rule 'even numbers ascending in twos' in a way that would help him discover that his rule is not the experimenter's rule by citing as few number sequences as you can. Please note that you have three pages on which to test your number sequences if you need to.

When you feel highly confident that you have helped Peter discover that his rule is not the experimenter's rule, and not before, you are to write down "Peter now knows his rule is not the experimenter's rule". You are to write this under your most recent number sequence. The researcher will then write whether or not you are correct beside your announcement."

Appendix D: Instructions used in experiment 4

"In a previous study investigating human thinking a participant called Peter was asked to discover a rule a researcher had in mind that the number sequence 2,4,6 conforms to. Peter hypothesised that the experimenter's rule was: 'even numbers ascending in twos'. You know that another participant called James hypothesised that the experimenter's rule was 'numbers ascending in twos'.

Your aim is to go about testing Peter's rule 'even numbers ascending in twos' in a way you think would help him to discover if his rule is the experimenter's rule. You are to do this by writing down other number sequences with sets of three numbers. You will then be informed if they conform or do not conform to the rule the researcher has in mind.

You should try to go about testing Peter's rule 'even numbers ascending in twos' in a way that would help him discover that his rule is or is not the experimenter's rule by citing as few number sequences as you can. Please note that you have three pages on which to test your number sequences if you need to. When you feel highly confident that you have helped Peter discover that his rule is or is not the experimenter's rule, and not before, you are to write down 'Peter now knows his rule is the experimenter's rule' or 'Peter now knows his rule is not the experimenter's rule'. You are to write this under your most recent number sequence and raise your hand. The experimenter will then write whether or not you are correct beside your announcement."

Appendix E: Instructions used in experiment 5

"In a previous study investigating human thinking a participant called Peter was asked to discover a rule a researcher had in mind that the number sequence 2,4,6 conforms to. Peter hypothesised that the experimenter's rule was: 'even numbers ascending in twos'. You know that another participant called James hypothesised that the experimenter's rule was 'something else'.

Your aim is to go about testing Peter's rule 'even numbers ascending in twos' in a way you think would help him to discover if his rule is the experimenter's rule. You are to do this by writing down other number sequences with sets of three numbers. You will then be informed if they conform or do not conform to the rule the researcher has in mind.

You should try to go about testing Peter's rule 'even numbers ascending in twos' in a way that would help him discover that his rule is or is not the experimenter's rule by citing as few number sequences as you can. Please note that you have three pages on which to test your number sequences if you need to. When you feel highly confident that you have helped Peter discover that his rule is or is not the experimenter's rule, *and not before*, you are to write down 'Peter now knows his rule is the experimenter's rule' or 'Peter now knows his rule is not the experimenter's rule'. You are to write this under your most recent number sequence and raise your hand. The experimenter will then write whether or not you are correct beside your announcement."

Appendix F: Recording booklet templates

				***Feedback from experimenter**
Number sequence	Reasons for choice	Do you expect it to conform to Peter's rule	Do you expect it to conform to the researcher's rule	Does your number sequence conform to the researcher's rule
2,4,6	…	yes	yes	**y**

Recording sheet (18 lines per participant, common to all conditions in which participants tested Peter's hypothesis).

				***Feedback from experimenter**
Number sequence	Reasons for choice	Do you expect it to conform to your rule	Do you expect it to conform to the researcher's rule	Does your number sequence conform to the researcher's rule
2,4,6	…	yes	yes	**y**

Recording sheet (18 lines per participant, common to all conditions in which participants tested their own hypothesis).

Appendix G

For the purpose of the study please circle yes or no where applicable below if you have ever done a problem like this before Yes / No,

Or

if you have taken courses dealing with the concepts of confirmation and falsification in the past Yes / No.