When Falsification is the Only Path to Truth

Michelle Cowley (cowleym@tcd.ie)
Psychology Department, University of Dublin,
Trinity College, Dublin, Ireland

Ruth M.J. Byrne (rmbyrne@tcd.ie)
Psychology Department, University of Dublin,
Trinity College, Dublin, Ireland

Abstract

Can people consistently attempt to falsify, that is, search for refuting evidence, when testing the truth of hypotheses? Experimental evidence indicates that people tend to search for confirming evidence. We report two novel experiments that show that people can consistently falsify when it is the only helpful strategy. Experiment 1 showed that participants readily falsified somebody else's hypothesis. Their task was to test a hypothesis belonging to an 'imaginary participant' and they knew it was a low quality hypothesis. Experiment 2 showed that participants were able to falsify a low quality hypothesis belonging to an imaginary participant more readily than their own low quality hypothesis. The results have important implications for theories of hypothesis testing and human rationality.

Hypothesis Testing and Falsification

People generate hypotheses to explain the workings of the world around them. They generate hypotheses in everyday social inference as well as in expert domains of scientific inquiry. The human ability to generate hypotheses has led to the accumulation of scientific knowledge. But it is the ability to test whether or not these hypotheses correspond to the evidence that leads to a true understanding. Experimental psychologists and epistemological philosophers have long sought to understand how people test their hypotheses (e.g., Popper, 1959; Wason, 1960; Kuhn, 1962). The key research questions have been: what is the best way to test hypotheses; how can people be successful hypothesis testers; and do people sometimes employ faulty hypothesis testing methods?

To address these questions cognitive psychologists borrowed the crucial concepts of *confirmation* and *falsification* from the philosophy of science. Hypothesis testing was categorized either as confirming, that is, people search for evidence that is consistent with a hypothesis, or falsifying, that is, people search for evidence that is inconsistent with a hypothesis. No matter how much evidence confirms a theory, there is never absolute certainty that it is correct (Popper, 1959). But if a major prediction of a theory is proved false, it can be inferred that the theory is incorrect, or at least incomplete. Falsification can safeguard against having numerous theories that explain the same phenomenon, each with some confirming evidence. Theories that are falsified can be abandoned because they no longer offer the best explanations. For this reason, the attempt to falsify, that is, to search for refuting evidence, has been advocated as the best way to test the truth of hypotheses. Nonetheless the merits of falsification have been questioned.

Problems for Falsification

Recent evidence indicates that people find falsification psychologically difficult and unhelpful (Poletiek, 1996; 2001). Experimental studies in the psychological literature have shown that people are rarely capable of falsification, if indeed they find it possible at all (Poletiek, 1996). Early studies found that people tended overwhelmingly to confirm their hypotheses. They searched for evidence that was consistent with their hypothesis and they avoided inconsistent evidence. The tendency was termed *confirmation bias* (e.g., Wason, 1960). Their failure to attempt to falsify was interpreted as a failure to think rationally.

Even great scientists appear to exhibit a confirmation bias, as studies of the Apollo moon mission scientists testify (e.g., Mitroff, 1974). The hypotheses of these scientists have stood the test of time, perhaps indicating that they were high quality hypotheses with little chance of falsification in the first place. People may find confirmation to be useful. In fact, attempts to confirm may be a better strategy than attempts to falsify, depending on the relationship between the hypothesis and the general principle to be discovered (Klayman & Ha, 1987). What then is the purpose of falsification? We suggest that a major purpose of falsification is to identify low quality hypotheses so that they can be abandoned. For example, chess masters have been shown to search for opponent moves that can falsify their plans, which helps them to discover that the moves they initially hypothesize to be good ones are mistaken (Cowley & Byrne, 2004).

In this paper we outline the results of two experiments that show that people can find falsification to be consistently possible and helpful. The experiments were carried out using Wason's 2-4-6 task. This task is a standard test bed for theories of hypothesis testing. The task presents a well-defined hypothesis testing situation for which the relationship between a hypothesis and the available evidence is precisely worked out (e.g., Klayman & Ha,

1987). In addition it has a well-defined logic for classifying confirming and falsifying hypothesis testing.

First we outline the logical properties of the task. Second we review some of the critical evidence from the task that has led researchers to question the usefulness of falsification (Klayman & Ha, 1987; Poletiek, 1996). Finally, we report a version of the task that we have designed that provides hypothesis testing situations in which participants must falsify in order to discover the truth—the 'imaginary participant' 2-4-6 task. We describe two experiments that address the question of whether or not people can falsify a hypothesis when it is rational to do so, that is, when the hypothesis is untrue.

The 2-4-6 Task

Wason (1960) first empirically investigated people's hypothesis testing using the 2-4-6 task. Participants were instructed to discover a rule the experimenter had 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.

Participants are asked to discover the rule by generating their own number triples. For each triple they are told 'yes' or 'no' by the experimenter as to whether or not it conforms to the rule or not. Each triple is taken to be a test of what the participant hypothesizes the rule to be. For example, participants tend to focus on the salient features of the 2-4-6 triple and generate hypotheses such as 'even numbers ascending in twos'. They propose triples such as 10-12-14 and 16-18-20. For each one of these triples they receive a 'yes' response from the experimenter. But the experimenter's rule is the deliberately general rule, 'any ascending numbers'. Hence, a triple such as 10-12-14 receives a 'yes' because it is consistent with the rule to be discovered ('any ascending numbers') as well as the hypothesis under test ('even numbers ascending in twos'). When each participant thinks they have discovered the rule they have to announce what they think it is. Typically participants announce an incorrect rule such as 'even numbers ascending in twos'. Only 21% of participants announced the correct rule first time round (Wason, 1960). This result has been replicated many times (e.g., Tweney *et al.,* 1980). The tendency for people to seek out information consistent with their hypotheses and avoid inconsistent information was termed *confirmation bias*. We turn now to the logic of classifying confirming and falsifying tests in the task.

The Logic of Hypothesis Testing in the 2-4-6 Task

Initial classifications of hypothesis testing as confirming and falsifying tests were equated with positive tests (tests that were consistent with the participant's hypothesis) and negative tests (tests that were inconsistent with the participant's hypothesis) respectively (Wason, 1960). For example, 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 is ascending in twos. This test is a positive test. But the classification must also take account of the *intention* of the hypothesis tester (Wetherick, 1962). The positive test described is a confirming test only if the participant *intends* the triple to conform to 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 and they expect their hypothesis is incorrect (Poletiek, 1996). Positive tests can be intended to either confirm or falsify.

The same is true for negative 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. This test is a negative test. 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, and so 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 inconsistent with their hypothesis 'numbers ascending in twos' and they also expect that it is inconsistent with the experimenter's rule. The negative test in this instance is intended to provide confirmation. Negative tests can be intended to either confirm or falsify. These four sorts of tests are considered the best classification scheme for confirming and falsifying hypothesis tests in the 2-4-6 task (e.g., Poletiek, 1996). Table 1 below shows an example of each of the four test types for the low quality hypothesis, 'even numbers ascending in twos'.

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

	Test triple	Positive or negative test	Expect to conform to rule
Confirming			
Positive	8-10-12	positive	yes
Negative	23-25-27	negative	no
Falsifying			
Positive	24-26-28	positive	no
Negative	5-10-15	negative	yes

Are people able to falsify their hypotheses, that is, do they generate negative falsifying tests? A participant who tests the low quality hypothesis 'even numbers ascending in twos' attempts to falsify only if they generate a negative test

triple such as 5-10-15, *and* they expect it to conform to the experimenter's rule. If they receive the feedback that the triple is consistent with the experimenter's rule, then they know their hypothesis 'even numbers ascending in twos' is untrue. The ability to generate negative falsifying tests is pivotal to the debate about whether people can falsify in a useful way.

Negative confirming tests may have been misidentified as falsifying tests in early studies of hypothesis testing. In one study participants were given the standard 2-4-6 task and they had to come up with their best guess about what the experimenter's rule was. They were presented with the number triple 2-4-6 and the experimenter's rule was the usual 'any ascending numbers' rule (Poletiek, 1996). Participants were given instructions either to 'test', 'confirm', or 'falsify'. 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). Although the participants proposed test triples that were negative tests, in fact they intended them to confirm. It was concluded that people do not seem to be able to make sense of falsification.

In our experiments we have examined whether people naturally tend to falsify their hypotheses in some situations (Cowley and Byrne, 2005). For example, a teacher who knows a student's hypothesis is incorrect may provide a counterexample to it. A scientist who believes another scientist's hypothesis is incorrect may design an experiment to falsify it. In the experiments we describe, we tested whether people can readily falsify low-quality hypotheses that have been generated by others. Existing evidence indicates that participants can understand the value of negative falsifying tests of a hypothesis when the *tests* are generated by somebody else (Kareev & Halberstadt, 1993). Our question is the opposite: can participants generate negative falsifying tests when the hypothesis has been produced by somebody else? We constructed a version of the 2-4-6 task in which participants tested someone else's hypothesis rather than their own.

The 'Imaginary Participant' 2-4-6 Task

We constructed an 'imaginary participant' version of the 2-4-6 task in which participants were told that an individual called Peter was asked to discover a rule that an experimenter had in mind, which the triple 2-4-6 conforms to (for details see Cowley and Byrne, 2005). The experimenter's rule was the standard 'any ascending numbers' rule. Participants were not required to generate the hypothesis themselves. Instead, they were asked to test Peter's hypothesis.

This task allowed us to control the precise relationship between the hypothesis under test and the truth (the experimenter's rule). The task allows us to give participants a low quality hypothesis. It must be falsified using negative falsifying tests in order to discover the truth. The low quality hypothesis 'even numbers ascending in twos' is *embedded* within the truth 'any ascending numbers'. Because the hypothesis is less general than the true rule, participants must falsify to find out that it is not the true rule.

The embedded relationship between a hypothesis and the experimenter's rule may influence a participant's potential to falsify. The embedded situation is one of five possible relationships that can exist between a participant's hypothesis and an experimenter's rule (Klayman & Ha, 1987). The relationships concern how much the participant's hypothesis and the experimenter's rule overlap with one another. Three relationships are relevant to our experiments.

The first relationship, outlined in Figure 1 (a), is the embedded relationship characteristic of the 2-4-6 task. The rule is 'any ascending numbers' and the participant's hypothesis is 'even numbers ascending in twos'. We adopted this situation for our imaginary participant 2-4-6 task. Participant's were given 'even numbers ascending in twos' as Peter's hypothesis and the experimenter's rule they had to discover was 'any ascending numbers'. The only way to intentionally falsify 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 Peter's hypothesis 'even numbers ascending in twos'). 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'. They can infer that Peter's hypothesis about 'evenness' or 'ascending in twos' cannot be true. It is not possible to falsify the hypothesis by generating a positive falsifying test.

Consider a participant who tries to falsify by generating the positive falsifying test, 24-26-28 (which is a positive test because it is consistent with Peter's hypothesis 'even numbers ascending in twos', but they expect it to be inconsistent with the true rule). When they receive a 'yes' from the experimenter they cannot infer that Peter's hypothesis pertaining to properties of 'evenness' or 'ascending in twos' is untrue. Although the positive test intended to falsify, it cannot. It is consistent with both the hypothesis and the true rule.

In the second type of embedded relationship, illustrated in Figure 1 (b), the participant's hypothesis is *more* general than the true rule. Consider a case where Peter's hypothesis is 'numbers ascending in twos' and the true rule is 'even numbers ascending in twos'. The true rule is embedded within Peter's hypothesis. It is possible to falsify by generating a positive falsifying test. A participant may generate the triple 3-5-7 (which is a positive test as it is consistent with the hypothesis 'ascending in twos'), but they expect that it is not consistent with the true rule. This time

they receive a 'no' from the experimenter, because 3-5-7 contains odd numbers. They can infer that the hypothesis is not the rule because it may pertain to numbers with the property of 'evenness'. Now consider a participant who tries to falsify by generating a negative falsifying test such as 5-10-15. This time when they receive a 'no' from the experimenter they may not infer that Peter's hypothesis 'numbers ascending in twos' is untrue. The triple is inconsistent with their hypothesis and the true rule and so it cannot discriminate between them (Klayman & Ha, 1987).

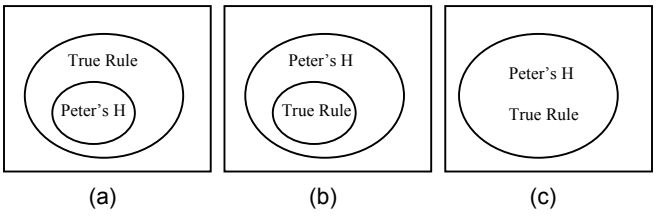

Figure 1: Three relationships between Peter's hypothesis and the true rule in the 2-4-6 Task

The third situation, illustrated in Figure 1 (c), is when the hypothesis is the same as the true rule, for example, Peter's hypothesis 'any ascending numbers' completely overlaps the true rule 'any ascending numbers'. 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. The test leads to confirmation. 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 falsify a true hypothesis.

Experiment 1

The aim of the experiment was to test whether people can falsify hypotheses consistently (see Cowley and Byrne, 2005). We gave participants a version of the 2-4-6 task with the crucial difference that their task was to test someone else's hypothesis. We presented some of them with an alternative hypothesis in addition to the hypothesis to be tested, in that we told some of them not only about Peter's hypothesis but also about the experimenter's rule. In so doing, we placed participants in a superior knowledge state much like a teacher who must present a counterexample to a student in order to show the student that a hypothesis is inaccurate. Participants must falsify and announce that Peter will know from this evidence that his hypothesis is inaccurate. Participants falsifying under these conditions can show they understand the implication of the test result if they demonstrate they *intend* to falsify.

The participants were sixty four members of the general public who volunteered and were paid a nominal fee, or who were undergraduate students and participated for course credits. There were 50 women and 14 men whose ages ranged from 15 years to 73 years, with a mean age of 35 years. No participants had taken courses in the philosophy of science.

They were assigned at random to one of four groups (n = 16 in each). They were told that an individual called Peter was asked to discover a rule that a researcher had in mind, to which the number sequence 2-4-6 conforms. Two of the groups were told that Peter hypothesises the rule to be 'even numbers ascending in twos' (a low quality hypothesis). One of these groups was told what the experimenter's rule was ('any ascending numbers'), and so they knew the hypothesis was a low-quality one, and the other group was not told what the experimenter's rule was. The other two groups were told that Peter's hypothesis is 'any ascending numbers' (a high quality hypothesis). Again, one group was told the experimenter's rule ('any ascending numbers'), and the other group was not. Participants in all four groups, low quality known, low quality unknown, high quality known and high quality unknown, were instructed to generate number triples of their own. They had to test Peter's hypothesis in such a way as to help him discover if his rule was the experimenter's rule (for further details see Cowley and Byrne, 2005).

Participants were tested individually and they recorded their number triples on a recording sheet which had five columns in which they wrote the triple, their reason for choosing the triple, whether they expected the triple to conform to Peter's hypothesis (i.e., is it a positive or negative test), whether they expected the triple to conform to the experimenter's rule (i.e., did they intend to confirm or falsify). In the last column the experimenter (the first author) wrote 'yes' or 'no' as to whether or not the triple did in fact conform to the rule. They were asked to announce when they were highly confident if their triples would show Peter if his rule was the experimenter's rule.

Results

The results show that people can consistently falsify hypotheses. In the low quality known condition**,** Peter's hypothesis is a low quality embedded one, 'even numbers ascending in twos' and the participants know that the rule to be discovered is in fact 'any ascending numbers'. Participants found it possible to falsify in this condition. 90% of their tests were falsifying ones. One example of a negative falsifying test triple generated by one of the participants in this condition was: 15-17-19. They said they expected 15-17-19 not to be an instance of the hypothesis they were testing (even numbers ascending in twos) yet they expected it to conform to the experimenter's rule (any ascending numbers). Each participant produced at least one negative falsifying test and more negative falsifying tests were generated in this condition than in any of the other conditions, low quality unknown (22%), high quality known, (0%), and high quality unknown (6%, $chi^2 = 18.325$ (1), $p < .0001$). Table 2 presents the percentages of each test type for all conditions.

Table 2: Percentages of hypothesis test types generated in each condition of 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

The result that 90% of the triples that participants generated were negative falsifying hypothesis tests shows that people can indeed falsify in some circumstances. What aspect of the low quality known condition facilitates falsification? One possibility is that people have a tendency to falsify someone else's hypothesis rather than one's own. Often, real life scientific hypothesis testing proceeds by attempting to falsify another scientist's hypothesis. Our second experiment tested whether people tend to falsify other people's hypotheses more than their own.

Experiment 2

The experiment was designed to discover whether participants falsify a low quality hypothesis when it belonged to someone else compared to when it belonged to themselves. Thirty two people who were members of the general public volunteered and were paid a nominal fee. There were 23 women and 9 men. Their ages ranged from 20 years to 75 years, with a mean age of 51 years. No participants had taken courses in the philosophy of science.

There were two conditions. In one condition the low quality embedded hypothesis was identified as belonging to the 'imaginary participant' Peter. Participants were told that Peter's hypothesis is 'even numbers ascending in twos'. In the other condition the identical hypothesis was identified as belonging to the participant. Participants were told that 'your hypothesis is even numbers ascending in twos'. Participants were not told what the experimenter's rule was in either condition. The recording sheet and procedure were the same as in experiment 1.

Results

Participants tended to falsify a hypothesis belonging to someone else more often than the same hypothesis belonging to themselves. Participants who tested Peter's hypothesis generated more negative falsifying tests than participants who tested their own hypothesis (32% versus 7%), although the difference was not reliable ($chi^2 = 2.667$ (4), $p = .307$), as Table 3 shows.

Table 3: Percentage of hypothesis test types generated in each condition of experiment 2.

	Peter's hypothesis	Self hypothesis
Confirming		
Positive	46	67
Negative	14	17
Falsifying		
Positive	8	9
Negative	32	7

Participants could abandon the hypothesis, that is, they announced that the hypothesis was not the researcher's rule when they had finished testing triples, or they could endorse it, that is, they announced that the hypothesis was the researcher's rule when they had finished testing triples. Participants tended to abandon the low quality hypothesis rather than endorse it when it was Peter's hypothesis (62% versus 38%); in contrast, they tended to endorse the low quality hypothesis rather than abandon it when it was their own (75% versus 25%), as Table 4 shows. The difference was reliable ($chi^2 = 4.571$ (1), $p = .016$).

Table 4: Percentages of abandoned and endorsed hypotheses in each condition of experiment 2.

	Peter's hypothesis	Own hypothesis
Abandoned hypothesis	62	25
Endorsed hypothesis	38	75

The results suggest that people falsify a hypothesis belonging to someone else more often and use the falsifying evidence to abandon a low quality hypothesis belonging to someone else more often than their own equally low quality hypothesis.

Conclusion

Our experiments demonstrate that people can consistently engage in falsification, that is, they can generate negative tests that genuinely falsify a hypothesis. They can falsify in the most difficult hypothesis testing situations, that is, when the hypothesis is low quality and it is embedded within the true rule. They falsified with the intention that their chosen tests would falsify. They appear to be aware of the implications of their test choice because they announced that the imaginary participant would realize from the test

results that a hypothesis was low quality (pace Poletiek, 1996).

The experiments suggest that people can falsify other people's hypotheses more readily than their own even without knowledge that the hypothesis is low quality or knowledge about what the relationship between the hypothesis and the truth is (pace Klayman & Ha, 1989). They tended not to abandon a low quality hypothesis when it was their own, compared to when it belonged to someone else**.** Participants may be able to rely on a falsification strategy in a rational way to test somebody else's hypothesis but not their own. We are currently examining the role of alternative hypotheses in helping people to falsify (Cowley and Byrne, 2005).

Acknowledgments

We thank Gry Wester for helping to collect and score the data. Thanks to Clare Walsh, Henry Markovits, Jonathan Evans, Alan Baker, and Phil Johnson-Laird for helpful comments. This research was funded by the Irish Research Council for the Humanities and Social Sciences and the Trinity College Dublin Arts and Social Sciences Benefactions Fund.

References

Cowley, M., & Byrne, R. M. J. (2004). Chess Masters' Hypothesis Testing. *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). Falsification in Hypothesis Testing: An Imaginary Participant 2-4-6 Task. *Manuscript in preparation.*

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. (1987). Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review, 94*, 211-228.

Kuhn, T.S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Mitroff, I. (1974). *The subjective side of science*. Amsterdam: Elsevier

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

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

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

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.

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

Wetherick, N. E. (1962). Eliminative and enumerative behaviour in a conceptual task. *Quarterly Journal of Experimental Psychology, 14*, 246-249.