8-9-2008

Assessing impact of instruction treatments on positive test selection in hypothesis testing

Daniel Wade Carruth

ASSESSING IMPACT OF INSTRUCTION TREATMENTS ON POSITIVE TEST SELECTION IN HYPOTHESIS TESTING

By

Daniel Wade Carruth

A Dissertation
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in Applied Cognitive Science
in the Department of Psychology

Mississippi State, Mississippi

August 2008

ASSESSING IMPACT OF INSTRUCTION TREATMENTS ON POSITIVE TEST SELECTION IN HYPOTHESIS TESTING

By

Daniel Wade Carruth

Approved:

____
Gary L. Bradshaw
Professor of Psychology
(Director of Dissertation)

____
J. Martin Giesen
Professor of Psychology
(Committee Member)

____
Jason S. McCarley
Assistant Professor of Psychology
(Committee Member)

____
Gary McFadyen
Assistant Research Professor Center for Advanced Vehicular Systems
(Committee Member)

____
Kevin J. Armstrong
Associate Professor of Psychology
Graduate Coordinator of the Department of Psychology

____
Gary L. Myers
Dean of the College of Arts and Sciences

Name: Daniel Wade Carruth

Date of Degree: August 9, 2008

Institution: Mississippi State University

Major Field: Psychology (Applied Cognitive Science)

Major Professor: Dr. Gary L. Bradshaw

Title of Study: ASSESSING IMPACT OF INSTRUCTION TREATMENTS ON POSITIVE TEST SELECTION IN HYPOTHESIS TESTING

Pages in Study: 278

Candidate for Degree of Doctor of Philosophy

The role of factors previously implicated as leading to *confirmation bias* during hypothesis testing was explored. Confirmation bias is a phenomenon in which people select cases for testing when the expected results of the case are more likely to support their current belief than falsify it. Klayman (1995) proposed three primary determinants for confirmation bias. Klayman and his colleagues proposed that a general *positive testing strategy* leads to the phenomenon of confirmation bias. According to Klayman's account, participants in previous research were not actively working to support their hypothesis. Rather, they were applying a valid hypothesis testing strategy that works well outside of laboratory tasks. In laboratory tasks, such as Wason's 2-4-6 task (Wason, 1960), the strategy failed because the nature of the task takes advantage of particular flaws in the positive testing behavior participants learned through their experience with the real-world. Given Klayman's proposed set of determinants for the positive testing strategy phenomenon, treatments were developed that would directly violate the

assumptions supporting application of the positive testing strategy. If participants were able to identify and act on these violations of the assumptions, the number of positive tests was expected to be reduced. The test selection portion of the Mynatt, Doherty, and Tweney (1977) microworld experiment was modified with additional instruction conditions and a new scenario description to investigate the impact of the treatments to reduce confirmation bias in test selection. Despite expectations, the thematic content modifications and determinant-targeting instruction conditions had no effect on participant positive test selection.

DEDICATION

I would like to dedicate this research to my children: Rebekah, Hannah, and Jacob.

ACKNOWLEDGEMENTS

I would like to thank my wife and my children for patiently supporting me through many long nights and weekends of writing.

I would also like to thank my advisor, Dr. Gary Bradshaw, and my committee members: Dr. Martin Giesen, Dr. Jason McCarley, and Dr. Gary McFadyen. I would like to thank Dr. Kevin Armstrong for an inspirational introduction to psychology and the initial encouragement that sent me down the path that led to today. Along the way, Dr. Stephanie Doane, Dr. Jason McCarley, and Dr. Vincent Duffy have all provided invaluable support and guidance by welcoming me into their research labs and helping me to develop my own research interests.

I would also like to thank everyone at the Center for Advanced Vehicular Systems for their support over the last year. Without the encouragement of Zach Rowland, Dr. Kari Reeves, John McGinley, Chris Blackledge, Teena Garrison, Alex Morais, Mark Thomas, Bryan Robbins, and many others, I would not have been able to complete this work.

TABLE OF CONTENTS

CHAPTER

APPENDIX

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The philosopher Karl Popper (1959) proposed the falsification approach for testing scientific hypotheses: scientists should focus on the tests that are most likely to provide falsifying evidence for the current hypothesis. The falsification approach is based on an analysis of the logic of hypothesis testing. Popper pointed out that, from a logical perspective, it is never possible to conclusively prove a hypothesis to be true. There is always the possibility that some untested case exists that would falsify the hypothesis. Instead, it is only possible to conclusively falsify a hypothesis through the discovery of contradictory evidence. Therefore, Popper (1959, 1962) proposed that proper hypothesis testing requires selecting tests for experimentation that are most likely to provide falsifying evidence for the current hypothesis.

Wason (1960) reported that a remarkable 80% of participants failed to guess a specific rule on their first guess in a simple rule discovery task precisely because they did not properly seek out falsifying evidence. Following the initial guess, all but one participant provided the correct rule within five guesses. In Wason's 2-4-6 task, participants were asked to determine a specific rule that governed the acceptability of a sequence of three numbers (referred to as a triple). At the start of the task, the participant was given a single example triple: 2-4-6. Participants were then asked to generate their own triples. For each triple generated by the participant, the experimenter would respond "yes" if the triple was acceptable under the rule or "no" if it was not acceptable. After the participant was confident that he/she knew the rule, the participant reported his/her hypothesized rule to the experimenter. Despite the simplicity of the task, approximately 80% of the participants made an incorrect initial guess.

Wason (1960) argued that the participants were biased in their selection of new triples. According to Wason, the participants developed a hypothesized rule based on the initial 2-4-6 triple. If participants were using Popper's falsification approach, the participants should have generated triples that would be likely to falsify their hypothesized rule. In Wason's experiment, participants tended to generate positive triples that *would* fit their current hypothesis. In other words, participants were generating triples that, according to their current hypothesis, they expected the experimenter to respond "yes" to. Very few participants generated negative triples that did not fit their current hypothesis. Wason claimed that, by generating positive triples, the participants were seeking evidence that would confirm or support their current hypothesized rule. This tendency in participants to select positive tests that are expected to provide confirming evidence has been referred to as *verification bias* or, more commonly, as *confirmation bias* (Klayman, 1995; Wason, 1960).

Wason 2-4-6 Task

It is worthwhile to examine the Wason 2-4-6 task in more detail. In the Wason 2-4-6 task, the participant is told that the experimenter has a target rule in mind that determines the acceptability of triples of whole numbers, the participant is given an initial example triple, the participant generates triples, the experimenter responds “yes” or “no” for each triple as dictated by the target rule, eventually the participant guesses the target rule. In some versions of the task, participants that guess incorrectly are allowed to generate additional triples and make further guesses until discovering the rule or giving up.

Hypothesis Generation

For our example, assume that a participant is given the initial 2-4-6 example triple. Based on this single triple, the participant develops an initial hypothesis for a rule that matches this triple. Participants often assume that the characteristics of the initial triple are relevant to the task (Van der Henst, Rossi, & Schroyens, 2002). The 2-4-6 triple has particularly salient relationships. For example, many participants note that the numbers 2-4-6 are a set of numbers increasing by 2. Participants assume that the initial triple is especially relevant to the task and focus on these salient relationships. According to Van der Henst et al., (2002) this is why the initial hypothesized rule generated by the participant is almost always some variation of ‘increasing even numbers’ or ‘numbers increasing by 2.’ With a hypothesized rule in mind, the participant then proceeds to generate new triples designed to test the rule.

Test Selection

An infinite number of triples can be generated by participants. Every triple has a relationship to both the hypothesized rule and the target rule. The triple's relationship to the hypothesized rule is based on its acceptability according to the hypothesis generated by the participant. If the triple is acceptable according to the hypothesized rule, then the triple is a positive test of the hypothesis (+H) because the participant expects the triple to fit the rule and to receive a "yes" response. If the triple is unacceptable according to the hypothesized rule, then the triple is a negative test of the hypothesis (-H); the participant expects the triple to not fit the rule and to receive a "no" response. The relationship of the triple to the target rule is given by the experimenter's response: "yes" if the triple is acceptable according to the target rule and "no" if the triple is unacceptable according to the rule. When a participant generates a triple, their goal is to generate evidence to either confirm or falsify their current hypothesis. Table 1.1 lists the type of evidence generated for each test type and each possible experimenter's response.

Table 1.1

Evidence Generated by Triple Given Test Type and Experimenter Response

	Experimenter Response (Result)	
Test Type	"Yes"	"No"
+H	Hit	False Alarm
-H	Miss	Correct Rejection

The result of the triple will provide evidence that will either confirm or falsify the hypothesis. The result confirms the hypothesis when the experimenter's response matches the expected response dictated by the participant's hypothesized rule. There are two types of confirming evidence. The first is a *hit* which occurs when a +H test leads to a "yes" response. The second is a *correct rejection* which occurs when a –H test leads to a "no" response. If the result of a triple test does not match the expected response, then the triple falsifies the hypothesis. There are two types of falsifying evidence. The first is a "false alarm" which occurs when a +H test leads to "no" response. This is a false alarm because the hypothesized rule would predict a positive result but the result was negative. The second is a "miss" which occurs when a –H test leads to a "yes" response.

Table 1.2 lists example triples, the experimenter response (relationship to target rule), the test type (relationship to hypothesized rule), and the evidence type (relationship between test type and experimenter response) for two common hypotheses in the Wason 2-4-6 task ('increasing even numbers' and 'numbers increasing by two'). As shown in the table (indicated by boldface), the only falsifying evidence for the Wason 2-4-6 task are misses which occur when a –H triple results in an unexpected "yes" response. The other type of falsifying evidence, false alarms, occur when a +H triple results in an unexpected "no" response. In the Wason 2-4-6 task, the target rule ('all triples of increasing numbers') is very broad and participants tend to select a narrow hypothesis focused on the salient characteristics of the 2-4-6 triple. It is therefore unlikely that a participant will generate a hypothesized rule that would lead to false alarms.

Table 1.2

Relationship Between Triples and Example Hypotheses for Wason 2-4-6 Task

		Example Hypotheses			
		Increasing Even Numbers		Increasing by Two	
Example	Experimenter Response	Test Type	Evidence	Test Type	Evidence
2,4,6	Yes	+H	Hit	+H	Hit
6,8,10	Yes	+H	Hit	+H	Hit
12,14,16	Yes	+H	Hit	+H	Hit
0,10,12	Yes	+H	Hit	**-H**	**Miss**
1, 2, 3	Yes	**-H**	**Miss**	**-H**	**Miss**
6,4,2	No	-H	Correct Rejection	-H	Correct Rejection
3,2,1	No	-H	Correct Rejection	-H	Correct Rejection
5,3,6	No	-H	Correct Rejection	-H	Correct Rejection
5,5,5	No	-H	Correct Rejection	-H	Correct Rejection
1,1,5	No	-H	Correct Rejection	-H	Correct Rejection
5,1,1	No	-H	Correct Rejection	-H	Correct Rejection
1,5,5	No	-H	Correct Rejection	-H	Correct Rejection

Evidence Collection

If the participant acts as most subjects in the Wason 2-4-6 task, almost all of the triples generated and presented to the experimenter will be +H triples. For these +H triples, the participant will expect the triples to result in a “yes” response from the experimenter. For example, if the participant believes that acceptable triples consist of numbers increasing by 2, the participant might generate the triple “6, 8, 10” or the triple

"12, 14, 16." As shown in the table (column labeled "Experimenter Response" in Table 2), the experimenter would, as expected, respond "yes" for both of these triples. As the participant receives more "yes" responses, the accumulation of confirming evidence increases confidence in the hypothesized rule until the participant decides that the hypothesized rule is the target rule and reports it to the experimenter.

Unfortunately for the participant, the hypothesized rule 'numbers increasing by 2' does not match the target rule devised by the experimenter despite the many "yes" responses received in response to the test triples. In the 2-4-6 task, the target rule is 'any set of increasing numbers.' All +H triples generated under the common initial hypotheses ('increasing even numbers' or 'numbers increasing by 2') will lead to "yes" responses. In defense of the participant, even if the participant proposed one of the numerous negative triples, such as "6,4,2" or "1,1,5" (see rows 6-12 of Table 2), the experimenter's "no" response would still confirm her current hypothesis (columns labeled "Evidence" in Table 2). Participants tend to select +H triples that will only confirm their hypothesis. Participants are unlikely to select any –H tests and are even more unlikely to generate a -H test that would falsify their current hypothesis. Each "yes" response strengthens the participant's confidence in the hypothesized rule until she offers her current hypothesis as the target rule. In the majority of cases, the use of +H triples leads to increasing confidence in an incorrect hypothesis and the participant's initial guess for the target rule is incorrect.

Problems with the Wason 2-4-6 Task

Wason (1960) argued that participants in the 2-4-6 task are focused on seeking out confirming evidence for their hypothesis and that this focus leads to a bias towards positive hypothesis tests. Rather than trying to falsify their hypothesis, participants are selecting tests that they expect to confirm their hypothesis.

However, some researchers have taken issue with the design of the Wason 2-4-6 task (Van der Henst, Rossi, & Schroyens, 2002; Vallée-Tourangeau & Penney, 2005). Vallée-Tourangeau, Penney, & Payton (2005) argued that the task was purposely designed to encourage generation of a constrained initial hypothesis. They noted three features of the task that they considered problematic: First, the actual rule (all triples of increasing numbers) is extremely broad. Second, the numbers making up the example triple are related through two salient relationships: the triple consists of even numbers and numbers increasing by 2. Third, there is no context beyond a mathematical context to provide a cue to the participant regarding the relevancy of the initial 2-4-6 triple to the target rule. Vallée-Tourangeau & Penney (2005) demonstrated that participant performance on the Wason 2-4-6 task could be significantly improved by simply adding an external representation of the task. For example, in one experiment, dice were used to display the initial triple and used by participants to generate new triples. Participants using dice to represent the task generated more triples, more triples that received a “no” response, and more types of triples. A larger proportion of their participants (66%) were able to successfully identify the target rule on their initial guess.

The first task participants have in the 2-4-6 task is to generate an initial hypothesis. With only the initial 2-4-6 example to guide them and a presumption of the importance of the initial triple, participants may be lead to make an invalid assumption that the example triple is meant to provide a clue to the nature of the experimenter's rule. Participants appear to rely on what little information they are given (the initial triple and their own knowledge of experimental settings) in the development of their initial hypothesis. Unfortunately, this leads participants to adopt an initial hypothesis that is too narrow. As mentioned earlier, a narrow hypothesis can only be resolved using –H tests, tests that participants have a tendency to avoid.

In a test of the influence of the initial triple, Van der Henst, Rossi, and Schroyens (2002) reduced the presumption of relevance in the Wason 2-4-6 task by giving a group of participants the impression that the triple was randomly generated. Their results were similar to Vallée-Tourangeau & Penney (2005). Participants proposed more triples, more triples that received a "no" response and more types of triples when they believed the initial triple was randomly generated. These participants performed better (55% success rate) at the task than a control group (24% success rate) that was given the initial triple in the same manner as in the original Wason 2-4-6 task. How the participant frames the initial evidence and what features of the evidence that the participant believes to be relevant has a significant impact on the generation of hypotheses.

In order to properly frame evidence, participants must have some knowledge of the problem to guide them in determining what is and what is not relevant. In the Wason

2-4-6 task, the participants are given no context that might assist in the development of an initial hypothesis. Another manipulation performed by Van der Henst, Rossi, and Schroyens (2002) provided context for the Wason 2-4-6 task. In the contextual 2-4-6 task, the task was couched in terms of sales performance for a salesman. The authors believed that the general “increasing numbers” rule might be more salient due to the familiarity of the context. Participants’ common knowledge of sales should allow participants to better frame the initial hypothesis and to better consider the appropriate role of the factors presented in the problem. Participants in the sales condition searched a broader space in triple selection leading to a larger number of participants (29% vs. 4%) correctly identifying the rule.

Thus, the purposeful selection of the 2-4-6 initial triple encourages participants to select a narrow hypothesis that specifically requires participants to select –H tests in order to falsify the hypothesis. If the initial triple is modified to discourage participants from believing that the particular features of the initial triple are important (i.e., apparent random selection of the initial triple; Van der Henst, Rossi, & Schroyens, 2002) or if appropriate context is provided (i.e., numbers represent sales figures over 3 months; Van der Henst, Rossi, & Schroyens, 2002), participants are more likely to broadly test the space of possible triples, receive negative responses, and successfully discover the target rule. Gale and Ball (2005, 2003) have investigated the relative impact of broader testing and receiving negative responses. Increased variety in test selection is not sufficient for success. The key to success appears to be the generation of at least one test that leads to a negative response (Gale & Ball, 2005).

Under the common hypothesized rules, the only evidence that would falsify the participant's hypothesis is a "miss" (a –H triple that leads to a "yes" response). The only triples in Table 2 that are misses under the common hypotheses are triples with increasing numbers that do not fit the additional constraints of the hypothesized rule. As an example, "1, 2, 3" is not a valid triple under the common initial hypotheses, but it is valid under the actual rule. The "yes" response to "1, 2, 3" would invalidate the participant's hypothesized rule and suggest the broader nature of the actual rule.

However, participants do not have *a priori* knowledge as to what type of error their hypothesized rule leads to. Without this knowledge, participants must rely on some previous knowledge or some strategy to select the most effective triples for testing their hypothesis. By falling back on experience or a strategy that may not apply, participants make positive test selections that can be and have been interpreted as a confirmation bias in the Wason 2-4-6 task. The studies previously reviewed (Van der Henst, Rossi, & Schroyens, 2002; Vallée-Tourangeau & Penney, 2005) have provided some evidence that changing the circumstances of the task leads to reduced positive test selection and improved performance in identifying the target rule in comparison to the original Wason 2-4-6 task design.

Mynatt, Doherty and Tweney Microworld

An additional complaint regarding the Wason 2-4-6 task is that the task is artificial and does not reflect a real world scientific discovery problem (Mynatt, Doherty, & Tweney, 1977). Mynatt, Doherty, and Tweney attempted to create a more realistic scenario to determine if a bias towards positive test selection would persist in conditions similar to real scientific discovery conditions. An interactive microworld (MDT Microworld) was created that allowed participant interaction via computer software. Unlike the Wason 2-4-6 task, the microworld presented participants with a concrete representation of a task: an on-screen environment with objects interacting with one another. Other modifications to the Wason 2-4-6 task included presentation of more than one initial scenario and a target rule that overlapped with the common hypothesized rules rather than a very broad target rule for which all common hypotheses were too narrow. Finally, participants were separated into three groups and were given instructions designed to specifically encourage them to falsify, test, or prove their hypothesis.

MDT Microworld Procedure

Display. The MDT Microworld was rendered onto a 512 × 512 pixel display. Figure 1 shows an example of the Microworld screens. The crosshair in the top left of the screen represented a source of particles that the participant could “fire” across the screen. One or more of the objects were presented on the screen. Each object was one of three shapes (triangle, disc, and square) and one of two brightness levels (50% and 100%).

Participants fired a particle by entering an angle between 0 and 359 on the keyboard and pressing enter. The particle would leave the source at the specified angle and move across the screen. Sometimes the particle would pass through the objects on the screen with no apparent affect. Other times, the particle would stop as it neared certain objects. The goal of the scenario was to discover a target rule that described which objects were surrounded by the invisible wall that would stop the particle's motion.

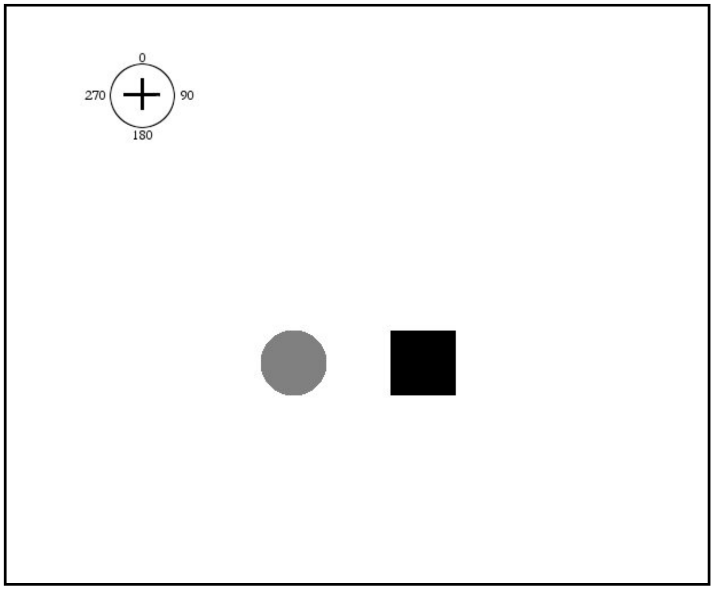


Figure 1.1. Author's recreation of a microworld screen used in Mynatt, Doherty and Tweney (1977).

Initial screens. Participants were introduced to the software and given two initial microworld screens to interact with. Participants could fire as many particles as they wished before moving on. The two initial screens were designed to encourage participants to develop an initial hypothesis that the triangle objects were surrounded by the invisible walls. In fact, objects with a 50% brightness level were surrounded by the invisible walls. After interacting with the initial screens, participants reported their initial hypothesis. Half of the participants were successfully manipulated into generating the desired incorrect triangle-based hypothesis for their initial hypothesis. The remaining participants, who identified a rule that was incompatible with the test selection materials, were dropped from the analysis of test selection.

Instructions. Participants were placed in three instruction groups. Each group received simple instructions for selecting microworld screens for further experimentation. In the *Test* instruction group (n = 7), participants were told that it is a scientist's job to test their hypothesis and that their task would be to test their own hypothesis. In the *Confirm* (n = 7) instruction group, participants were told a short story about a famous scientist working to confirm their hypothesis. Participants were then told that their task was to select tests that would confirm their hypothesis. In the *Disconfirm* (n = 6) instruction group, participants were told a short story about a famous scientist working to falsify their hypothesis. Participants were told to select tests that would falsify their current hypothesis.

Test selection. Participants were then given paper packets containing 10 pairs of microworld screens. Each page of the packet displayed two images of different possible experiments. The pairs of screens were constructed to include tests that participants would expect to generate confirming evidence and tests that would be expected to generate falsifying evidence. The participants were instructed to select whichever of the two screens that they believed would most effectively test their current hypothesis. Participants reported their selections to the experimenter.

Results. In the analysis of the participant's test selections, approximately 70% of the test selections made by the participants were tests categorized as confirming by Mynatt, Doherty, and Tweney (1977). There was no significant effect of the instruction condition on positive test selection. Despite the differences between the Wason 2-4-6 task and the MDT Microworld, participants' test selections appear to demonstrate the same bias in test selection as shown by participants in the Wason 2-4-6 task. In these tasks and in others (Wason selection task, Wason, 1966; Dual-Goal version of 2-4-6 task, Tweney et al., 1980; Gale & Ball, 2005, 2003), participants are selecting +H tests that, if their hypothesis is correct, would only result in confirming evidence. Participants are not selecting –H tests in order to falsify their hypothesis. The results suggest that a confirmation bias persists across different scientific discovery tasks (Mynatt, Doherty, & Tweney, 1977).

Alternative Interpretation of Positive Testing

This is not the only interpretation of the results. Another possibility is that participants in the Wason 2-4-6 task and the Mynatt, Doherty, and Tweney Microworld (MDT Microworld) are not seeking only to confirm their hypothesis through +H tests or simply biased towards +H tests but are exhibiting an *effective testing strategy*, in which the use of +H tests is encouraged over –H tests. In the Wason 2-4-6 task, the only tests that would generate falsifying evidence were –H triples that did not fit the participant's current hypothesis. However, in the MDT Microworld task, +H tests, that would generate confirming evidence if the participant's hypothesis is correct, *can* generate falsifying evidence. This is because, unlike the Wason 2-4-6 task, the initial hypothesis that participants in the MDT Microworld were expected to adopt ('triangles stop particles') overlaps with the target rule ('50% brightness objects stop particles'). Whereas some +H trials fit the target rule (e.g., a 50% brightness triangle), other +H trials (e.g., a 100% brightness triangle) are false alarms and do not fit the target rule. These +H trials will generate falsifying evidence. Table 1.3 gives example screens for each type of evidence that a test can provide.

Table 1.3

Example of Types of Evidence by Test Type and Test Result.

	Test Result	
Test Type	Particle Stopped	Particle Passes Through
+H	Hit (50% Triangle)	False Alarm (100% Triangle)
-H	Miss (50% Disc)	Correct Rejection (100% Disc)

As an example, if the participant believes that triangle-shaped objects stop particle movement, then testing a screen with a 100% brightness triangle-shaped object is a +H test (and selection of the test could be construed as evidence of a confirmation bias). However, because the actual rule is that objects of 50% brightness stop particles, the particle will pass through the triangle-shaped object, resulting in a false alarm (see Table

1.3). In this case, a positive test, considered evidence for confirmation bias, will generate falsifying evidence leading the participant to reevaluate the current hypothesis. Instead of a bias towards confirming evidence, the participant may be purposely selecting +H tests because +H tests are an effective method of falsifying a given hypothesis.

Positive Testing Strategy

It may be possible that participants are seeking falsifying evidence in the Wason 2-4-6 task and the MDT Microworld task when they select positive tests. Klayman (1995) suggests that the positive test selection, previously interpreted as confirmation bias, is actually the result of the application of a learned strategy or heuristic: the *positive test strategy*. Klayman's argument is that the positive test strategy works on real-world problems so long as certain assumptions are met. When the assumptions are not met and participants are unaware of this violation of assumptions, participant behavior gives the appearance of a confirmation bias.

Klayman and his colleagues (Klayman, 1995; Klayman & Brown, 1993; Klayman & Ha, 1989, 1987; Slowiaczek, Klayman, Sherman, & Skov, 1992) have examined how people select tests and interpret test results for hypothesis testing. Klayman proposes three primary determinants for the apparent confirmation bias in participant behavior during hypothesis testing: 1) a positive testing strategy; 2) a preference for extremity; and 3) a preference for tests with a higher apparent diagnosticity. According to Klayman and colleagues, these three determinants lead to reasonable performance in real tasks outside of the laboratory, but the Wason 2-4-6 task exploits particular flaws created by the three determinants leading to very poor performance in the laboratory setting.

Klayman and Ha (1989, 1987) propose the existence of a positive testing strategy used by participants in hypothesis test selection and defend the general reasonableness of this testing strategy by examining the logic and probabilities involved in hypothesis test selection. Before describing the positive test strategy in detail, I will discuss the hypothesis test selection task.

In hypothesis test selection, a researcher may select tests for one of three reasons: First, the researcher may be seeking additional evidence without a clear hypothesis (Klahr & Dunbar, 1988). Second, the researcher may be seeking evidence that will support a current hypothesis. Third, the researcher may be seeking evidence that will falsify a current hypothesis. In the second and third case, according to Popper's falsification approach, the researcher should focus on experiments that may generate evidence to falsify the hypothesis. However, in order to identify the type of tests that will falsify the hypothesis, the researcher must know what type of error the current hypothesis is most likely to produce.

The relationship between the type of error and the type of tests can be represented by considering the different "spaces" of the scientific discovery problem. There exists a universe (U) of possible experimental tests related to the scientific problem. Within the universe of tests, there exist sets of tests defined by the target rule (target set, +T set) and one or more hypotheses (hypothesis set, +H set). Figures 1.2 – 1.9 represent the sets of tests in diagram form.

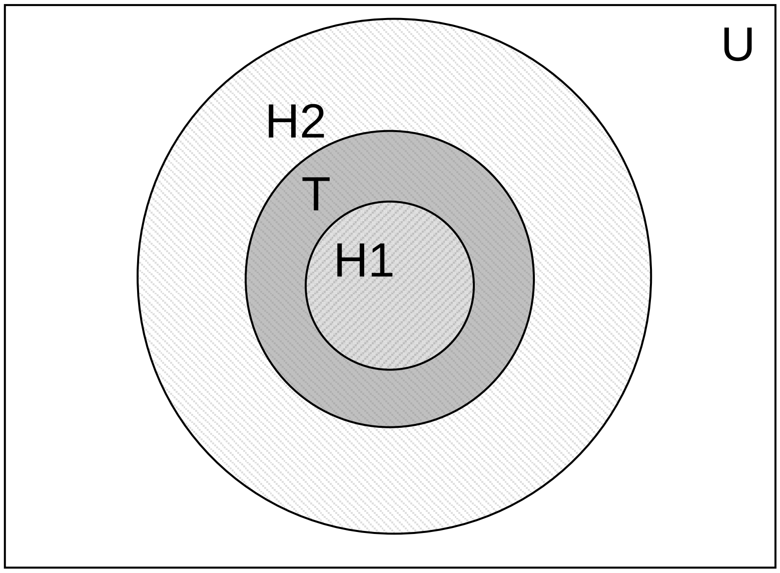


Figure 1.2. Broad and narrow hypothesis spaces. The box U represents all possible tests of the hypothesis. The circle T represents the target rule space. The inner circle, H1, represents a narrow hypothesis space. The outer circle, H2, represents a broad hypothesis space.

If the researcher is aware of the relationship between the hypothesis set and the target set, the choice of +H or –H tests is simple. If the set of +H tests is narrower (H1 in Figure 1.2) than the set of target instances (tests that fit the target rule, T in Figure 1.2), the researcher should focus on -H tests. Alternatively, if the set of +H tests is too broad (H2 in Figure 1.2), the researcher should focus on +H tests. However, the researcher often does not know the actual relationship between the current hypothesis space and the target space. Klayman's (1995, 1987) claim is that researchers mitigate this lack of knowledge by applying the positive testing strategy.

Table 1.4

Definitions and Examples of the Four Types of Tests for Wason 2-4-6 and MDT Microworld

			Examples	
Test Type	Definition	Goal	Wason 'increasing even numbers'	MDT 'triangles stop particles'
+H	Test that should result in the phenomenon of interest according to the hypothesis	Show event happens	6,8,10	100% triangle
-H	Test that should not result in the phenomenon of interest according to the hypothesis	Show event does not happen	5,3,6	50% disc
+T	Examination of a test known to include phenomenon of interest	Compare test characteristics to hypothesis	2,4,6	50% triangle
-T	Examination of a test known to not include phenomenon of interest	Compare test characteristics to hypothesis	NA	100% square and 100% disc

Types of Tests

When presented with a research problem, researchers can employ four types of tests: +H tests, -H tests, positive target (+T) tests and negative target (-T) tests. Target tests are examination of known evidence (e.g., the 2-4-6 triple or the two initial screens from the MDT Microworld) and are described in detail in the following sections. Table 1.4 lists each of the test types with a definition and examples for the Wason 2-4-6 task and the MDT Microworld task.

Positive hypothesis tests. A +H test (positive hypothesis test) occurs when the researcher examines an instance of the problem that the researcher expects to include the phenomenon of interest based on the current hypothesis. The researcher performs a +H test to determine whether the phenomenon of interest occurs as expected. For example, in the Wason 2-4-6 task, the triple “6, 8, 10” is a +H test of the hypothesis ‘increasing even numbers.’ In the MDT Microworld task, testing a screen with a 100% brightness triangle is a +H test of the ‘triangles stop particles’ hypothesis. As an additional example, assume a researcher hypothesizes that tornadoes occur during when atmospheric conditions meet certain criteria: stormy, winds between 15 and 50 mph, temperature between 70° and 90° F, and pressure between 1000 and 1500 mb. A +H test would require that the researcher seek out a storm that fits the criteria and observe whether a tornado is generated.

Negative hypothesis tests. A negative hypothesis test (-H test) occurs when the researcher examines an instance of the problem in which the hypothesized conditions are *not* true in order to determine whether or not the phenomenon of interest occurs. In the Wason 2-4-6 task, the triple “5, 3, 6” is a –H test for the hypothesis ‘increasing even numbers.’ In the MDT Microworld task, testing a screen with a 50% disc is a –H test for the ‘triangles stop particles’ hypothesis. For the tornado researcher, a –H test would occur when the researcher performs field research on a sunny, calm day and observes that a tornado does not occur or when the researcher observes a storm when the temperature is 50° F (below the 70° to 90° range believed to be required).

Positive target tests. Outside of the laboratory, researchers often have access to another type of test. Target tests are examinations of instances that are available via previous observations. A positive target test (+T test) occurs when the researcher examines the conditions associated with a known instance of the phenomenon of interest.

In the Wason 2-4-6 task, the triple “2,4,6” is an example of a +T test. The participant knows that the initial “2,4,6” triple fits the target rule because the experimenter provided it as such. In the MDT Microworld task, the participant was provided two introductory screens and provided an opportunity to observe particle-object interactions. The first screen included a 50% brightness triangle, a 100% square, and a 100% disc. The 50% brightness triangle stopped particles and the 100% square and 100% disc allowed particles to pass through. This single screen provided participants with multiple examples of particle-object interaction before participants were asked to select additional test screens. In the first screen, participants could determine that the 50% triangle object was clearly stopping particles. When considering these initial tests during test selection, the 50% triangle would act as a +T test of the hypothesis. In the tornado example, the researcher could examine all records of tornadoes to identify attributes shared amongst the occurrences and determine whether the known temperature, pressure, and other attributes were within the hypothesized ranges.

Negative target tests. A negative target test (-T test) occurs when the researcher examines the conditions when the phenomenon of interest has never occurred. In the Wason 2-4-6 task, participants do not have a –T test available at the beginning of their testing and may never generate a –T test during their testing. In the MDT Microworld task, the first introductory screen included a 100% disc and a 100% square that allowed the particles to pass through them. During test selection, the memory of the observed particle-object interactions are –T tests available to participants. The tornado researcher might look at records of storms that did not generate one or more tornadoes to attempt to determine what attribute(s) are missing from those instances.

Test types in rule discovery tasks. In the Wason 2-4-6 and the MDT Microworld tasks, people tend to use +H tests rather than –H tests. In the Wason 2-4-6 task, participants have only one known instance (+T test) of the target and can perform no other target tests (other than re-examining the results of the previous hypothesis tests). The design of the 2-4-6 task is such that +H tests will cause participants to have an undue level of confidence in their initial hypothesis. In the MDT Microworld, participants tended to select +H tests even when explicitly told to disconfirm their hypothesis. Participants were shown two screens with multiple objects that provided participants with both +T tests (i.e., 50% triangle in first screen) and –T tests (i.e., 100% square in first screen). During test selection, participants in the MDT Microworld were not allowed to perform any selected tests until all tests were selected. The only +T and –T tests available in MDT Microworld were provided by the introductory screens.

Role of +H and –H Tests in Hypothesis Testing

In order to support their claim that the positive testing strategy works well in the real world, Klayman and Ha (1987) undertook an analysis of the possible benefits of a positive testing strategy. First, Klayman and Ha determine what information can be determined from the execution of +H tests and –H tests given the different possible relationships between the hypothesis and reality. The relationship between the set of hypothesis-matching instances and the set of target instances is represented in the following figures (Figures 1.3 – 1.9) as Venn diagrams.

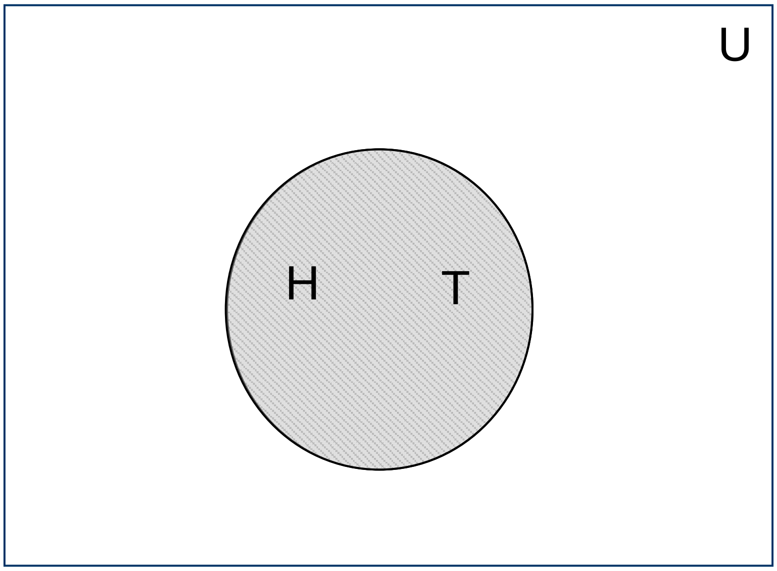


Figure 1.3. Correct hypothesis. The box U represents all possible tests related to the problem. The set of positive tests according to the hypothesis (H) matches the target set (T). All tests generate confirming evidence.

Correct hypothesis. If the hypothesis is correct, then the set of hypothesized instances is the same as the set of target instances (see Figure 1.3). In other words, every test that is hypothesized to result in the phenomenon of interest does result in the phenomenon of interest. In this scenario, neither +H tests nor –H tests will generate new evidence that would falsify the hypothesis. For example, in the Wason 2-4-6 task, if the hypothesis is ‘increasing numbers,’ every +H triple generated will fit the target rule ‘increasing numbers’ and every –H triple will not fit the target rule. There will be no misses and no false alarms generated by the results of the tests. In the tornado example, the hypothesized conditions would perfectly predict when a tornado will occur and when a tornado will not occur.

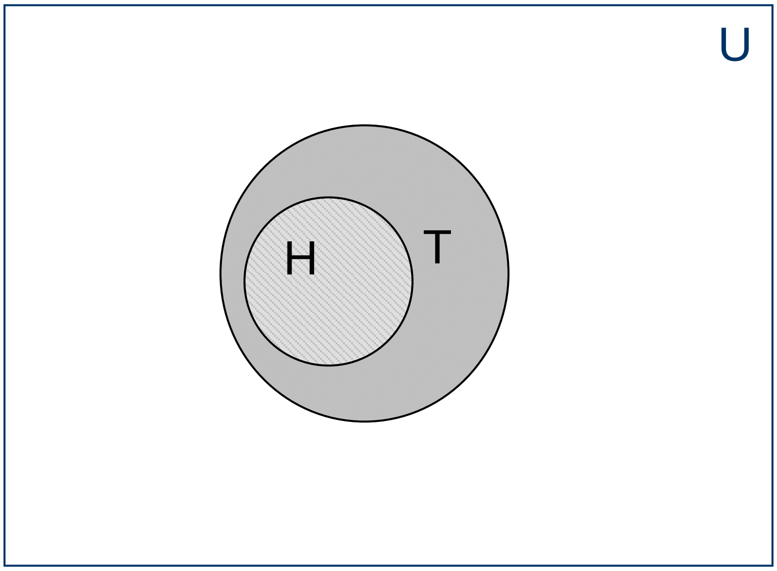


Figure 1.4. Narrow hypothesis. The box U represents all possible tests related to the problem. The set of positive tests according to the hypothesis (H) is contained within the target set (T). Only the –H tests that lie outside of the hypothesis set (H) but inside the target set (T) will provide falsifying evidence.

Narrow hypothesis. In Figure 1.4, the hypothesis is too narrow and the set of hypothesized instances (H) is completely contained by the set of target instances (T). As in the Wason 2-4-6 task, every +H test will generate evidence supporting the current hypothesis. Only certain –H tests will reveal that the hypothesis fails to account for all of the target instances. A narrow hypothesized set completely contained within the target set can only provide falsifying evidence via misses (-H test with “yes” response). False alarms cannot be generated because there are no +H tests that are not in the target set. In the Wason 2-4-5 task, the common initial hypotheses generated by an analysis of the single target instance (2-4-6) are too narrow and the +H tests performed by the participants do not reveal the error in the hypothesis. If the participant's hypothesis is 'increasing even numbers,' all +H tests (see Table 1.2) will result in a 'yes' response and provide evidence that confirms the hypothesis. Most –H tests will result in a 'no' response and provide evidence that confirms the hypothesis. The participant must select a –H test of the hypothesis that is within the set of target instances that match the target rule 'increasing numbers.' For example, the participant could select "1,2,3" which is a –H test of 'increasing even numbers' but matches the target rule. Thus, the participant would receive an unexpected 'yes' response that would falsify the current hypothesis.

In the tornado example, if the target conditions for generation of a tornado are: stormy, winds between 20 and 50 mph, temperature between 65° and 85° F, and pressure between 1000 and 1500 mb and the researcher's hypothesis is: stormy, winds between 30 and 45 mph, temperature between 65° and 85° F, and pressure between 1000 and 1500 mb, then the researcher's hypothesized wind speed condition is too narrow and is missing instances when tornadoes will be generated. Focusing only on the wind speed condition, a +H storm (e.g., 40 mph winds) will only confirm the hypothesis because any instance matching the hypothesized conditions also falls within the target rule's conditions. Some –H tests (e.g., 10 mph winds) will confirm the researcher's hypothesis because many –H tests also fall outside of the target rule's conditions. In order to falsify the hypothesis, the researcher must observe a –H storm (e.g., 25 mph winds) with a wind speed that falls outside of the hypothesized conditions but is within the target rule's conditions.

Broad hypothesis. Alternatively, if the hypothesis is too broad (see Figure 1.5), the situation is reversed. All –H tests will generate negative results as expected and provide only confirmatory evidence. Some +H tests that happen to fall within the target set will generate positive results and provide confirmatory evidence. Only the +H tests that are within the hypothesized conditions but outside of the target rule's conditions will reveal that the hypothesis is false. In this case, only false alarms (+H test with "no" response) will generate falsifying evidence and the hypothesis must be narrowed in order to match the target set.

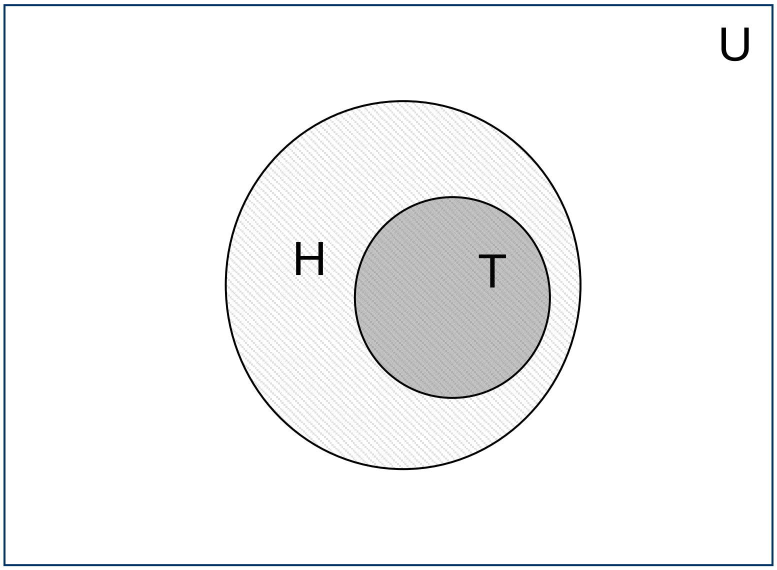


Figure 1.5. Broad hypothesis. The box U represents all possible tests related to the problem. The set of positive tests according to the hypothesis (H) contains the target set (T) and tests outside of the target set (false alarms). Only the +H tests that lie inside the hypothesis set (H) but outside the target set (T) will provide falsifying evidence.

In the tornado example, assume the same target conditions previously defined for generation of a tornado: stormy, winds between 20 and 50 mph, temperature between 65° and 85° F, and pressure between 1000 and 1500 mb. This time assume that the researcher's hypothesis is: stormy, winds between 10 and 55 mph, temperature between 65° and 85° F, and pressure between 1000 and 1500 mb. Again, for simplicity, the only difference between the target conditions and the hypothesized conditions are wind speed. A –H storm (e.g., 60 mph winds) provides confirmatory evidence because all of the –H storms under the researcher's current hypothesis are outside of the constraints of the target rule. Some +H storms (e.g., 45 mph winds) are within the conditions dictated by both the researcher's current hypothesis and the target rule. These +H storms will generate a tornado as expected and provide confirmatory evidence for the researcher's hypothesis. Only a +H storm that is within the conditions dictated by the researcher's current hypothesis (e.g., 55 mph winds) but is outside the conditions dictated by the target rule will falsify the researcher's hypothesis. These +H storms are false alarms and will fail to generate the predicted tornadoes.

Intersecting hypothesis and target sets. When the set of instances matching the hypothesis intersects with the set of target instances (Figure 1.6), both +H tests and –H tests may provide falsifying evidence. The hypothesis generates both false alarms and misses. The researcher must pursue both lines of questioning in order to collect the evidence required to correct the hypothesis. All four types of evidence (see Table 1.3) are possible in this scenario. In the MDT Microworld, the desired initial hypothesis (‘triangles stop particles’) overlaps with the actual rule (‘50% brightness objects stop particles’). The hypothesized set (H) overlaps with the set of target instances (T) and can result in hits (e.g., 50% brightness triangle). Some instances that are in H are not in T and will result in false alarms (e.g., 100% brightness triangle). Other instances are not in H but are in T and will result in misses (e.g., 50% brightness disc). Finally, some instances are outside the dictates of both the hypothesis set and the target set and will result in a correct rejection (e.g., 100% disc). In order to effectively test that the hypothesis matches the target rule, the researcher must perform both +H tests and –H tests

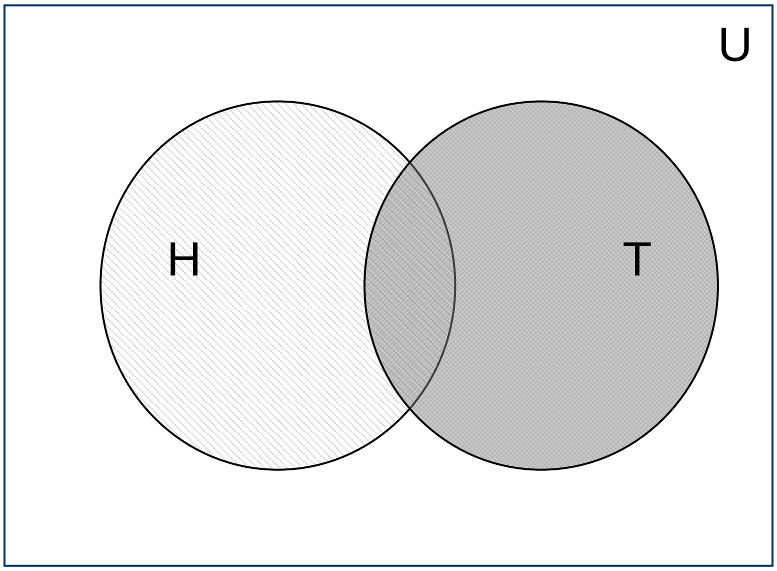


Figure 1.6. Intersecting hypothesis and target sets. The box U represents all possible tests related to the problem. The set of positive tests according to the hypothesis (H) intersects the target set (T). Since a subset of +H tests and a subset of –H tests lie inside the target set, +H and –H tests may be used to generate both confirming and falsifying evidence.

In the tornado example, let us focus only on the wind speed conditions and assume the same conditions for the target rule: winds between 20 and 50 mph. Assume the researcher’s new hypothesized conditions specify winds between 40 and 60 mph. Storms with wind speeds between 40 and 50 mph meet the conditions dictated by both the hypothesis (+H test) and the target rule. Observations of storms meeting these conditions will generate the expected tornadoes and will provide confirmatory evidence for the researcher’s hypothesis. Storms with wind speeds between 20 and 40 mph are outside of the hypothesized conditions (-H tests) but are within the dictates of the target rule. Observations of these storms will unexpectedly generate tornadoes providing falsifying evidence in the form of a missed target. Storms with wind speeds below 20 mph and above 60 mph are –H tests according to the hypothesis and outside of the conditions dictated by the target rule. These storms are correct rejections and will, as expected, not generate a tornado thereby providing confirmatory evidence for the researcher’s hypothesis. Finally, storms with wind speeds between 50 and 60 mph are +H tests that lie outside of the conditions specified by the target rule. These storms generate false alarms, do not produce a tornado as expected, and falsify the researcher’s hypothesis.

Incorrect hypothesis. If the hypothesis is completely incorrect (see Figure 1.7), all +H tests and some –H tests will falsify the hypothesis. In the tornado example, focus on the wind speed and assume the same target rule conditions: wind speeds between 20 and 50 mph. Assume the researcher’s current hypothesis is that wind speeds between 10 and 15 mph generate tornadoes. If the researcher observes any storm with a wind speed that falls within the hypothesized conditions (e.g., 12 mph winds), the storm will fail to generate a tornado since all storms within the hypothesized conditions have wind speeds too slow to generate a tornado. In this scenario, all +H storms produce false alarms and provide falsifying evidence. If the researchers observes a –H storm with a wind speed that falls within the conditions required by the target rule (e.g., 25 mph winds), the researcher will observe an unexpected tornado also providing falsifying evidence for the researcher's current hypothesis. However, not all –H storms produce misses. If the researcher observes a storm with wind speeds outside of the hypothesized conditions and outside the conditions dictated by the target rule (e.g., 18 mph winds), there will be no tornado and the researcher will receive confirmatory evidence for the current hypothesis.

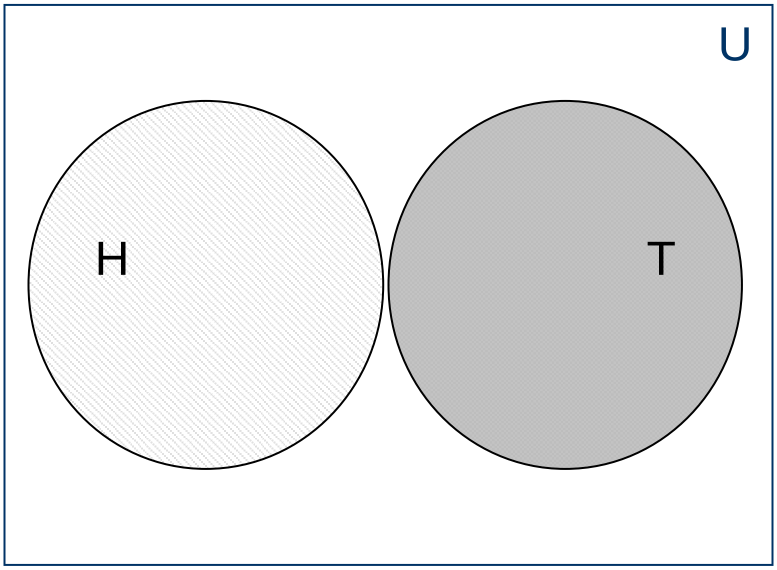


Figure 1.7. Incorrect hypothesis. The box U represents all possible tests related to the problem. The hypothesis set (H) is completely disjoint from the target set (T). All +H tests and any –H tests that lie within the target set will generate falsifying evidence.

As can be seen from the analysis of the types of tests and the role of the +H and –H tests in hypothesis testing, the appropriateness of each test is dependent on the relationship between the hypothesis and the target rule. In hypothesis testing, the critical element is to select tests that are most likely to falsify the hypothesis; not just to select –H tests (Klayman & Ha, 1987). In the Wason 2-4-6 task, -H tests are the only tests that will falsify the hypothesis but, in many other scenarios, +H tests or a combination of +H and –H tests may be the tests that are most likely to falsify the hypothesis.

Klayman's Determinants of Confirmation Bias

As previously stated, the researcher is unlikely to be aware of the relationship between their hypothesis and the target rule that determines whether the phenomenon of interest occurs. However, a researcher can make a reasonable assessment of the situation (Klayman & Ha, 1987). The simplest guide for selecting tests depends on accurately assessing what type of error is important (Friedrich, 1993; Klayman & Ha, 1987). If the researcher determines that it is important to avoid false positives but misses are unimportant, then the researcher can focus on +H tests to narrow the hypothesis. If the researcher can afford false positives but must not miss a target, then the researcher must employ –H tests to ensure that the hypothesis is broad enough to include all of the target instances (Friedrich, 1993). For example, in the tornado experiment, the researcher is likely to determine that minimizing misses is most important to ensure that alarms are sounded any time there is a chance that a storm might generate a tornado. In this case, the researcher will focus on –H tests to ensure that the hypothesized conditions cover all possible conditions that might generate a tornado.

Likewise, in the MDT Microworld task, if a participant determines it is most important to know when the particle will be stopped, the participant should use –H tests (e.g., 50% disc) to ensure that the participant's hypothesis includes all of the instances when the particle will be stopped. If a participant determines it is most important to know when the particle will not be stopped, the participant should use +H tests (e.g., 100% triangles) to narrow the current hypothesis and reduce false alarms.

If both types of error are equally important to the researcher and the researcher is uncertain about the relationship between the current hypothesis and the target rule, the researcher can apply a testing strategy such as Klayman's (1995; Klayman & Ha,1987) positive testing strategy. However, the suitability of the positive testing strategy depends on two assumptions.

Positive Testing Strategy

Uncommon event assumption. First, the appropriateness of the positive testing strategy in the real world assumes that most real-world hypothesis testing is focused on investigating an uncommon event (Klayman & Ha, 1987). If the target instances are in the minority, then it is likely that the number of –H tests is very large. If the number of –H tests is very large, then the effectiveness of –H testing is reduced.

For example, if the tornado researcher's hypothesis predicts that only 5% of storms generate a tornado, that leaves 95% of storms as –H tests that would need to be examined to ensure that there are no missed tornadoes. Without some boundary conditions that significantly pare down the total set of instances, -H testing will be very inefficient.

Accurate hypothesis assumption. Second, the positive testing strategy is appropriate when the researcher has some confidence that the hypothesis base rate is close to the target set base rate (Klayman & Ha, 1987). This is not an unlikely assumption: if a researcher is presenting a hypothesis to account for some phenomenon of interest, it seems probable that the researcher has a reasonable level of confidence in the hypothesis. If the hypothesis set is known to be smaller than the target set, then, as previously discussed, the researcher should already be aware that –H testing is necessary to broaden the hypothesis. If the hypothesis set is the same size or significantly larger than the target set, +H testing will catch the false alarms generated by the hypothesis.

Given that the +H set is about the same size as the target set (T) and that the target set contains fewer than 50% of the instances, then the greater part of the universe of instances are –H tests. In Figure 1.8, you can see that the set of –H tests (instances outside of H) is much larger than the set of + H tests (instances inside of H). If the target event is uncommon and the hypothesis is reasonably accurate, then the search for falsifying evidence should be limited to the smaller set of +H tests and not the larger set of –H tests. This is analogous to searching for a needle in a haystack. If you have a large haystack (universe of instances, U in Figure 1.8), a small number of needles (target set, T in Figure 1.8), and you have some idea where the needles are (hypothesis set, H in Figure 1.8), then it makes more sense to search where you believe the needles to be (+H tests, H in Figure 1.8) rather than search the rest of the haystack (-H tests, instances outside of H in Figure 1.8).

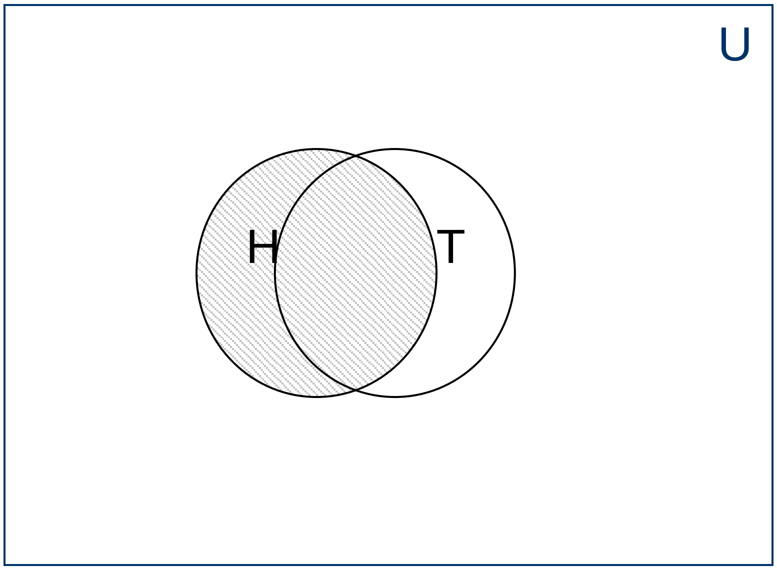


Figure 1.8. The set of +H tests (H) is significantly smaller than the size of the set of –H tests. The box U represents all possible tests related to the problem. The hypothesized set (H) is much smaller than the set of –H tests (white area). When positive testing strategy assumptions are valid, -H testing is not likely to efficiently falsify the hypothesis.

Violation of the assumptions. However, if the assumptions are violated, then the positive testing strategy is no longer effective (Klayman & Ha, 1987). In the Wason 2-4-6 task, the target set (all triples of increasing numbers) is 1/6 of the universe of instances (Klayman & Ha, 1987). The first assumption of the positive testing strategy is that the set of target instances is less than 50% of the universe of instances. Therefore, the first assumption for applying the positive test strategy holds. The second assumption of the positive testing strategy is that the hypothesis set is similar in size to the target set. Participants accept the assumption that their hypothesis is similar in size to the target set in the absence of evidence to the contrary. However, in the Wason 2-4-6 task, the participants are incorrect. The hypothesis "increasing even numbers" is significantly smaller than the target set. If the participant was aware of this violation of assumptions, the participant should employ –H tests to search for those targets that lay outside the hypothesis set. Participants are not aware that the assumptions underlying the positive testing strategy have been violated. Therefore, Klayman and Ha (1987) claim, that the participants in the Wason 2-4-6 task continue to apply the positive testing strategy and select +H tests based on the positive testing strategy and not based on a bias towards confirming evidence.

If the same analysis is applied to the MDT Microworld task, it is clear that the assumptions underlying the positive testing strategy are again violated but in a slightly different way. Participants in the MDT Microworld task could be presented with 27 possible screens given the constraints that all objects have an equal likelihood of appearing and allowing a single object, two of the same object, and combinations of two objects. Given the target rule (objects of 50% brightness stop particle movement), there are 18 screens in the target set (3 single object screens, 3 screens with two of the same 50% object, and 12 combinations including at least one 50% object). 67% (18/27) of the universe of screens are in the target set. Therefore, the first assumption is violated in the

MDT Microworld task because the target set is greater than 50% of the universe of instances. However, only 13 screens (2 single object screens, 9 combinations with at least one triangle-shaped object, and 2 screens with two of the same triangle-shaped objects) are in the hypothesis set (triangle-shaped objects). This mismatch between the base rates for the target set (67%) and the hypothesis set (48%) may lead participants to believe that the first assumption has not been violated. The relative sizes of the target set, hypothesis set, and the universe of instances can be seen in Figure 1.9. The set of target instances (67%) is represented by the solid grey circle. The set of hypothesis instances (48%) is represented by the diagonal filled circle. The crosshatch shows the overlap between the target set and the hypothesis set (screens containing 50% brightness triangles). The white space is the set of MDT Microworld screens that are neither target nor hypothesis instances (e.g., screen containing 100% brightness disc).

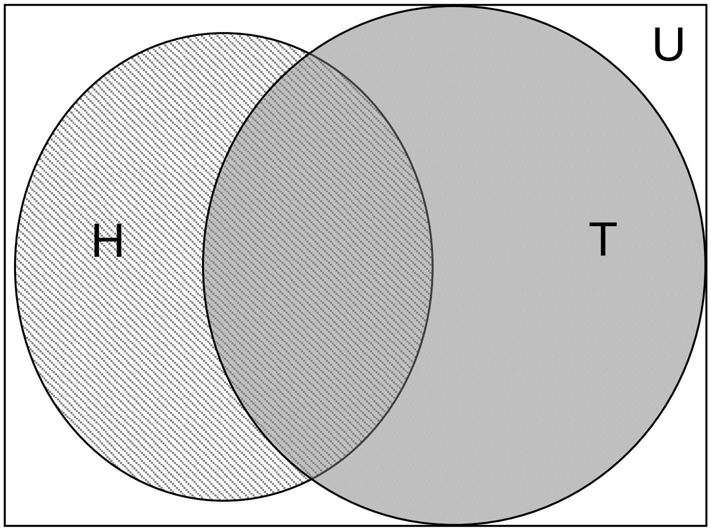


Figure 1.9. Rendering of the relative sizes and relationships between the MDT Microworld target (T), hypothesis (H), and universe sets.

The MDT Microworld task also violates the second assumption that the hypothesis set is similar in size to the target set. The hypothesis set size (48%) is smaller than the target set (67%) and is too narrow. As discussed earlier, a narrow hypothesis requires –H testing to resolve. In addition to the size discrepancy, the MDT Microworld task is different from the Wason 2-4-6 task in that the microworld target and hypothesis sets overlap (similar to Figure 1.6). The triangle hypothesis only matches a relatively small (narrow) portion of the target set; it also matches a relatively large (broad) portion of the universe of instances outside of the target set. The triangle hypothesis generates both false alarms and misses and requires both –H and +H tests to fully identify the flaws in the hypothesis.

In Figure 1.9, there are four spaces indicated by combinations of coloring and pattern. The white space indicates the universe of tests that are outside the hypothesized set and the target set. A –H test in the white region will result in a correct rejection. The solid grey space is the set of target instances outside of the hypothesized set. A –H test in the solid grey space will result in a positive event providing falsifying evidence in the form of a miss. The white space with the diagonal pattern is the set of hypothesized instances that are not target instances. A +H tests will result in a negative event providing falsifying evidence in the form of a false alarm. The overlap between the diagonal pattern and the solid grey space indicates the hypothesized set of instances that match the target set. A +H test in this space is a hit and provides confirming evidence for the hypothesis.

The participants have no knowledge that the assumptions underlying the positive test strategy are violated by the facts of the task. Because of their limited exposure to targets, the participants have no knowledge of the size of the target set. Participants also have no reason to believe that there is a significant discrepancy between the set of hypothesized instances and the target set. Without evidence that the assumptions are violated, participants will continue to apply the positive testing strategy to the problem. Based on Klayman and Ha (1987), it seems reasonable to conclude that, participants are incorrectly applying the positive testing strategy in the Wason 2-4-6 task and the MDT Microworld task. This inappropriate use of the strategy leads participants to select +H tests when these tests will reveal little or no falsifying evidence. At least in the MDT Microworld task, the nature of the relationship between the hypotheses is such that +H tests can reveal that the hypothesis is incorrect.

Target tests. An important additional consideration in defense of the positive testing strategy is the possibility of utilizing +T tests. In most real-world scenarios, it is likely that the researcher has access to multiple cases in which the target event is known to occur. When available, +T tests can be used in place of –H tests to falsify the hypothesis. Klayman and Ha (1987) show that positive target tests have a higher probability for falsifying hypotheses. In Figure 1.8, the set of +T tests is much smaller than the set of –H tests.

According to Klayman and Ha (1987), avoiding –H tests and using both +H tests and +T tests provides testers with the highest probability for falsifying their hypothesis when the target event is uncommon, which is often the case in real-world hypothesis testing. In the Wason 2-4-6 task, participants are presented with a particular case in which +H tests will fail to falsify the hypothesis and relatively few +T tests are available for consideration. In the MDT Microworld task, participants are presented with a case in which +H tests may falsify the hypothesis but –H tests would be more effective for testing the hypothesis. By inappropriately applying a strategy that works in the real-world, participants demonstrate significant positive test selection on the Wason 2-4-6 and MDT Microworld tasks. Klayman & Ha (1987) claim that participants inappropriately use a series of +H tests, not because they are biased toward confirming evidence, but because they are unaware that the assumptions underlying the positive testing strategy are not met in these tasks. This misapplication of the positive testing strategy leads researchers to interpret participant behavior during hypothesis testing as evidence for confirmation bias.

Preference for Extremity

In the Wason 2-4-6 task and many other laboratory rule discovery tasks, the rule is always either correct or incorrect. Slowiaczek, Klayman, Sherman, and Skov (1992) examined hypothesis testing in a probabilistic task. Participants are asked to select questions for an alien being with features that are more likely in one species and less likely in another species. When faced with probabilities, participants tend to have a bias towards testing cases that are either extremely likely or extremely unlikely to include the target event given the current hypothesis. Suppose, for example, that 90% of the Glom alien species and 50% of the Fizo alien species eat rocks and 55% of the Gloms and 75% of the Fizos have fur. According to the preference for extremity account, participants will tend to select the 90% case and ask whether the alien eats rocks (Slowiaczek, Klayman, Sherman, & Skov, 1992). This preference for extremity, combined with the positive testing strategy, leads participants to prefer +H tests that are most likely be positive. In the Glom and Fizo example, if the participant believes the alien to be a Glom, the participant will still ask whether the alien eats rocks. However, if the participant believes the alien to be a Fizo, the participant would ask whether the alien has fur.

In the MDT Microworld task, if an object in the microworld meets the requirements of the rule, the object always has an invisible wall. If participants adopt an all-or-nothing approach to their hypothesis, the preference for extremity will not impact participants' test selections in the proposed research. However, participants may probabilistically weigh the likelihood of the particle being stopped based on some function of the object's shape and brightness. In the initial screens, participant's see the particle stopped by a 50% brightness triangle and a cluster including a 100% triangle and a 50% disc. If the participant has a more sophisticated internal model of the problem than "if triangle, 100% probability; else 0% probability," then the participant may estimate the likelihood slightly differently than expected. For example, when faced with a 50% square-shaped object, the participant has not previously observed a square stopping a particle's motion so the participant estimates a 0% (or very close to 0%) likelihood that the 50% square will stop a particle. However, when faced with a 50% disc, the participant might estimate a likelihood of 25% since it shares features with a cluster that stopped a particle in one of the initial screens.

If participants do generate a probability estimate for screens based on the features of the objects, then test selection will be impacted by this estimate. When given two screens to select from, participants will estimate the likelihood of each screen and will select the screen that is most likely or most unlikely to be a target instance. Given the influence of positive testing, participants would be expected to tend to select tests that are most similar to their internal model and are expected to be target instances. This selection of tests that participants should expect to be target instances leads to the appearance of a confirmation bias.

Sensitivity to Diagnosticity

In a probabilistic environment, different questions provide different levels of information. Slowiaczek et al. (1992) investigated participants' sensitivity to the different levels of diagnosticity of different questions. Participants tend to select the test case that they believe will be most informative. However, participants appeared to not account for the different level of diagnosticity of different answers. For example, given two alien species (Gloms and Fizos) and two features that appeared probabilistically in both groups (such as having fur and eating rocks), participants were able to identify that whether the alien eats rocks (a feature present in 90% Gloms – 50% Fizos) was more diagnostic than whether the alien has fur (a feature present in 75% Fizos – 55% Gloms). They were not able to recognize that the alien not eating rocks was stronger evidence for the Fizo than the alien eating rocks was for a Glom. The different answers for the same question were treated similarly leading to poor revision of beliefs.

For the selection of tests during hypothesis testing, participants appeared to consider the difference in probability (40 and 20 in the previous example) as a simple estimate of the diagnosticity of the question. Participants will tend to select the more diagnostic question.

As previously discussed, the event in the MDT Microworld task is not probabilistic. Therefore, if participants are basing their estimates of likelihood purely on the hypothesis, participants are not expected to be influenced by their sensitivity to the diagnosticity of the question. If participants have a more sophisticated method for estimating the likelihood of the particle stopping, then the interpretation of test results could be skewed by the lack of sensitivity to the diagnosticity of the results. Because the proposed research is investigating only test selection, error in participants' sensitivity to diagnosticity should not affect participant performance.

Summary of Klayman's Determinants

The three primary determinants proposed by Klayman and his colleagues based on a logical analysis of hypothesis testing include a general positive testing strategy (Klayman, 1995; Klayman and Ha, 1987, 1989), a preference for extremity, and a partial sensitivity to diagnosticity (Klayman, 1995; Slowiaczek et al., 1992). Klayman and Ha (1987) claim in their analysis that the behavior of participants in hypothesis testing that has been interpreted as a confirmation bias is due to the combination of the general positive testing strategy and the preference for extremity. The proper application of the positive testing strategy depends on certain assumptions that are violated in the Wason 2-4-6 task and the MDT Microworld task. If the tasks can be manipulated so that participants were aware of the violation of assumptions and participants have an implicit or explicit understanding of the relationship between the assumptions and the utility of +H tests, I would expect the participants to appropriately modify their test selections and demonstrate reduced positive test selection.

Additional Determinants of Positive Test Selection

Besides the assumptions underlying the positive test strategy, other aspects of scientific discovery tasks can reduce positive test selection. The following sections will discuss some aspects related to the presentation of the task to participants and the possibility that individual differences may explain some of the common results for scientific discovery tasks.

Task Representation

Testing alternative hypotheses. In the Wason 2-4-6 task, usually only 20% of the participants successfully guess the target rule on their first attempt. Tweney et al. (1980) created a logically equivalent task where over 60% of the participants are successful on their first attempt. To achieve this dramatic improvement in performance, Tweney et al. modified the instructions of the 2-4-6 task and asked participants to discover not one rule but two complementary rules: the ‘Dax’ and the ‘Med’ rule. Participants are asked to generate triples that conform to either the ‘Dax’ rule or the ‘Med’ rule. When a triple is presented, the experimenter responds ‘Dax’ if the triple is acceptable according to the ‘Dax’ rule (increasing numbers) and ‘Med’ if the triple is acceptable according to the ‘Med’ rule (all other triples). Since the rules are mutually exclusive and exhaustive, any +H test of the ‘Dax’ rule is a –H test of the ‘Med’ rule and vice versa. This increases the likelihood of participants generating a –H triple and receiving disconfirmatory evidence for their ‘Dax’ rule.

Thematic content. Griggs and Cox (1982) dramatically demonstrated the benefits of thematic content in the drinking age variant of the Wason selection task. In the Wason selection task (1966), participants are given a rule: If *P*, then *Q*. The participants are then given four cards showing half of the information relevant to the rule: *P*, ~*P*, *Q*, or ~*Q*. Participants are asked to select which cards need to be investigated to determine if the rule is being broken. The normative response is to select *P* to see if the other half of the card shows *Q* and to select ~*Q* to make sure the other half does not show *P*. In abstract versions of the selection task, participant performance is dismal as almost all participants will select *P* and *Q* rather than *P* and ~*Q*.

Griggs and Cox (1982) added thematic content to the Wason selection task in the drinking age variant. In the drinking age variant, the participants are given the rule: If a person is drinking beer, then the person must be over 18 years of age. The participants are told that there are four people holding drinks at a party: one is drinking a beer (*P*), one is drinking a coke (~*P*), one is known to be 21 years of age (*Q*), and one is known to be 16 years of age (~*Q*). Given this scenario, many more participants correctly select the person drinking the beer to make sure the person is over 18 years of age and the person known to be 16 years of age to make sure the person is not drinking beer. The thematic content allows the participant to use knowledge of the domain to appropriately select the tasks.

Worth noting is the possibility that the dramatically improved performance on the Griggs and Cox (1982) drinking age variant of the Wason selection task over the original Wason selection task may be due to a domain-specific adaptation, possibly for social exchange or, in this case, the detection of violators of a social rule.

Griggs and Cox (1982) suggest a memory-based, familiarity effect of content that allows participants to apply experience to a problem with a familiar content. However, further research has indicated that the benefits of thematic content are domain-specific.

Gigerenzer and Hug (1992) examined two theories that propose that the benefits of thematic content are largely limited to conditional rules that involve permission and obligation (pragmatic reasoning schema; Cheng and Holyoak, 1989) or, more strictly, (social contract theory; Cosmides & Tooby, 2005). Cheng and Holyoak (1989) propose that reasoning uses structured knowledge based on experience including 'permissions,' 'obligations,' and 'causations.' Successful reasoning is limited to the domains supported by this structured knowledge. If a task does not include permissions and obligations, reasoning is significantly more difficult. According to Cheng and Holyoak (1989), if the selection task is reframed in terms of permissions and obligations (i.e., the drinking age variant), then participants can readily solve the problem by applying pragmatic reasoning schemas.

In their review of adaptations for social exchange, Cosmides and Tooby (2005) discuss social contract theory which constrains the domains to a subset of the Cheng and Holyoak proposal. Social contract theory relates to scenarios with perceived benefits and perceived costs associated with actions. Cosmides and Tooby (2005) suggest that specific reasoning capabilities have developed adaptively for social exchange scenarios and other adaptively useful scenarios (such as identifying dangerous situations). In a comparison of social contract theory and pragmatic reasoning schema theory, Gigerenzer and Hug (1992) reported that the key element in producing correct behavior on the Wason selection task was not only a social exchange context but also a context that puts the participant in the perspective of someone who may be cheated by the exchange. Cosmides and Tooby (2005) discuss research that has shown that selections in variants of the selection task depend on the perspective taken during selection.

Adding a simple familiar context may not be sufficient to improve hypothesis test selection. At least, a context designed with social contract theory in mind, that puts the participant in the perspective of being cheated, would have a higher likelihood of modifying test selection behavior.

Individual Differences in Positive Test Selection

In most of the scientific discovery tasks, a small number of participants do appropriately select tests and successfully complete the tasks. In the original Wason 2-4-6 task, approximately 20% of the participants guessed the target rule on their first try. Approximately 30% of the test selections reported by Mynatt, Doherty, and Tweney (1977) were not classified as confirmatory test selections. At least some participants in these tasks are selecting appropriate tests and performing normatively. Stanovich (1999; Stanovich & West, 2000) focused his analysis on these normatively performing participants.

Cognitive ability. Stanovich and West (2000) examined performance on a series of problems and found that performance on some tasks was related to measures of cognitive ability (SAT scores). Stanovich (1999) reviewed performance of participants on four different classes of tasks: syllogistic reasoning, Wason selection, statistical reasoning, and argument evaluation. For each of these tasks, only a small number of participants generate the normative response. However, the normatively responding participants' responses differed systematically from the other participants on all the tasks. The presence of a systematic difference strongly suggests some individual differences that allow some participant's to generate normative responses on these scientific discovery tasks (Stanovich, 1999). Certain measures of cognitive ability (SAT scores, Raven Matrices, vocabulary tests) were significantly correlated with four classes of reasoning tasks (SAT score correlations: r = .358 to .470). Stanovich (1999) also reported a significant negative correlation (r = -.223) between SAT scores and hypothesis testing bias from a separate experiment.

Stanovich (1999) also examined the possibility that training in math or statistics would improve performance on one of the four tasks. A math/statistics background did correlate with performance on the argument evaluation task but none of the other tasks. Stanovich also considered that the normatively responding participants might be responding in the manner they believed was desired by the experimenter. The participants were scored for socially desirable response tendencies but the scores were not correlated with normative responding.

Experience and education. If cognitive ability correlates with appropriate responses to some rule discovery tasks, perhaps experience or education also improves performance. Wason (1960) reported no significant effect of Arts and Sciences background on the 2-4-6 task. Tweney and Yachanin (1985) used a variant of the Wason selection task to directly investigate whether experienced researchers would perform better than undergraduates on a scientific discovery class. Tweney and Yachanin presented active researchers and undergraduates with two variants of the Wason selection task: the drinking age variant (Griggs & Cox, 1982) and a risk factor variant. The risk factor variant required participants to assume the role of a foreman that must determine whether workers are in violation of the following conditional rule: if a worker's risk factor is greater than 7, the worker must wear a hard hat. The participants chose from risk-7(P), risk-2 (~P), safety helmet (Q), and safety glasses (~Q). The normative response is to select the risk-7 (P) and the safety glasses (~Q) to determine if the worker is violating the rule. On both tasks, the scientists performed better than the undergraduates.

However, performance was not as improved in the risk factor variant as in the drinking age variant.

Griggs and Ransdell (1986) extended the Tweney and Yachanin (1985) results by looking at the performance of university researchers on an abstract variant of the Wason selection task and the risk factor variant of the task. Despite the additional training and experience of the researchers, the researchers performed poorly on the abstract task and only slightly better than Tweney and Yachanin's undergraduates on the risk factor variant of the task. Griggs and Ransdell suggest that the apparent effect of education in the Tweney and Yachanin study is actually an effect of thematic content (the scientists have some familiarity with workplace safety) and not an effect of scientific experience or education.

Like other content-related results, the Griggs and Ransdell (1986) results may also be explained by a domain-specific theory (i.e., social contract theory). In the Cosmides and Tooby review of neurocognitive adaptations for social exchange (2005), they discuss the adaptive importance of detecting danger due to violation of a rule. The Tweney and Yachanin (1985) risk factor variant of the Wason selection task requires participants to identify when worker's are in violation of the safety rules and participants' improved performance on the risk factor variant task may be due to a specialized ability developed for social adaptation.

Creativity and divergent thinking. Vartanian, Martindale, and Kwiatkowski (2003) investigated the possibility that normatively responding participants were more creative or employed divergent thinking to perform the tasks. Participants completed a measure for divergent thinking and then performed the Wason 2-4-6 task. Two factors contributed to a successful outcome: First, the number of –H triples generated that resulted in a “yes” response (misses). Participants that generated a large number of misses were more likely to determine the correct rule. Additionally, Vartanion, Martindale, and Kwiatkowski’s measure of divergent thinking was also correlated (*r* = .24) with successful performance of the task (Vartanion, Martindale, & Kwiatkowski, 2003) suggesting that divergent thinking directly or indirectly leads to improved performance on the 2-4-6 task.

Summary of Determinants

To summarize, a number of factors that may influence positive test selection in scientific discovery tasks such as Wason 2-4-6 and the MDT Microworld have been identified and many have been tested to some extent. The relative importance of misses and false alarms may dictate the use of +H or –H tests (Klayman, 1995; Friedrich, 1993). If a participant wants to ensure that no targets are missed, then the participant must use –H tests to detect targets that lie outside of the set of hypothesized tests. If misses and false alarms are of equal importance, Klayman and colleagues (Klayman, 1995; Klayman & Ha, 1987) have suggested that participants may apply a positive testing strategy that is appropriate for real-world scenarios where two assumptions generally hold: the phenomenon of interest is an uncommon event and the set of +H tests is similar to the target set. If participants are made aware that either assumption is violated and participants are adaptive, then they should be able to modify their test selection strategy and generate normative responses. Klayman & Ha (1987) do not claim that participants are necessarily adaptive and did not empirically test whether test selection behavior was modified when assumptions were violated.

In addition to the importance of the assumptions underlying positive test strategy, Griggs and Cox (1982) have demonstrated the successful use of thematic content for generating normative responses in a selection task. Another successful method for generating normative responses is the use of alternative hypotheses (Tweney et al., 1980; Gale & Ball, 2005, 2003). Finally, Stanovich (1999) and Vartanian, Martindale, and Kwiatkowski (2003) described the role of individual differences (cognitive ability, *r* between 358 to .470, and divergent thinking, *r* = .24) for describing the difference between the normative responding participants and other participants.

However, despite this bulk of work, several open questions remain regarding hypothesis test selection. First, Klayman and Ha (1987) is a statistical analysis of the logic and probabilities associated with hypothesis testing and not an empirical study of the proposed positive testing strategy. Klayman and colleagues (Klayman, 1995; Slowiacek et al., 1992) have examined sensitivity to diagnosticity and preference for extremity, but I am not aware of a study directly examining participant's ability to recognize violations of the assumptions required for positive testing strategy or participant's response to a violation of an assumption. In my opinion, the ability for a participant to respond to the violation of the assumption is a fundamental argument for a positive testing *strategy* rather than a positive testing *bias*. In order to investigate participant awareness and adaptability to violations of assumptions, the current research presented participants with information that was expected to lead them to determine that the uncommon event assumption had been violated.

Second, Klayman and Ha (1987) and Friedrich (1993) propose that researchers should be able to recognize when misses would have more significant ramifications (high miss cost) than false alarms. To avoid missing target events, a researcher must use –H testing. Neither Klayman and Ha (1987) nor Friedrich (1993) provide empirical data to support their claim that researchers will adjust their testing behavior based on the importance of avoiding misses over false alarms. In order to investigate participant response to a high cost of misses, the current research presented participants with information detailing the cost of accepting a hypothesis that missed target instances.

Third, in the dual-goal version of the Wason 2-4-6 task, the target rule (DAX, ‘increasing numbers’) and the alternative rule (MED, ‘everything else’) that participants must discover are mutually exclusive and exhaustive. A +H test for DAX is a –H test for MED and vice versa. There is no overlap between the rules as there is in the MDT Microworld study. If a positive test of any hypothesis is considered a +H test, every test performed in the dual-goal version of the Wason 2-4-6 task is a +H test and, rather than reducing any bias towards positive tests, the dual-goal version simply takes advantage of the bias to improve performance on the task. In order to investigate the impact of an alternative hypothesis on hypothesis testing for a more realistic and complex scenario, the current research provided some participants with an alternative hypothesis.

Fourth, thematic content has been used to modify participant behavior in rule discovery (Van der Henst, Rossi, & Schroyens, 2002) and selection tasks (Griggs & Cox, 1982). McKenzie (2006) used familiar materials to increase participant sensitivity to the differential diagnosticity of tests. The MDT Microworld task is a fairly abstract task with no stated connection between the objects in the microworld and real-world scenarios. Given a scenario with real-world meaning, participants may be able to better comprehend the problem, respond to violations of the assumptions underlying the positive test strategy, recognize the need for –H tests, and reduce the number of positive test selections compared to the original, less familiar scenario. In order to investigate the impact of thematic content on the MDT Microworld design, the current research provides participants with two scenarios: the original MDT Microworld scenario and a new scenario that reframes the MDT Microworld scenario as a story of predator and prey.

Fifth, although some research has determined that participants who respond normatively to the Wason 2-4-6 task also respond normatively to other discovery tasks, little research has been performed to investigate the nature of the individual differences that determine this behavior. The available research has focused on measures of divergent thinking (r = .24, Vartanian, Martindale, & Kwiatkowski, 2003), cognitive ability (r between .358 and .470 depending on the task, Stanovich, 1999; Stanovich & West, 2000), education and experience (no effect, Griggs & Ransdell, 1986). None of these research efforts completely explains the differences between those who successfully complete test selection tasks and those who do not. The current research explores the possible influence of personality on positive test selection using scales drawn from the International Personality Inventory Pool (IPIP). The selected scales are similar to the NEO PI-R big five scales and other scales that were judged to have a face valid relationship to scientific discovery (i.e., creativity, intellect).

CHAPTER II

PILOT STUDY

The experimental design used for the current research included a number of novel additions to the methodology and modifications or extensions to the original Mynatt, Doherty, and Tweney (1977) microworld study. First, revised version of the experiment scenario was created to add thematic content to the microworld that was expected to improve participant understanding of the task and lead to changes in test selection. Second, participants were given an initial hypothesis rather than being asked to generate one to constrain the possible hypotheses participants would be testing. Third, rather than asking participants to simply select effective tests of the hypothesis, participants were asked five different questions designed to investigate how participants assessed the figures and made their test selections. Fourth, a personality inventory was designed and included to assess individual differences. Fifth, the questionnaires were significantly longer than those used in the original study and there were concerns regarding how long participants would require for completing two questionnaires. Finally, the study was constrained to the paper-based test selection phase of the original study to examine test selection only.

Given the significant modifications from the original study, a pilot study with a single within-subjects factor (scenario) was performed in order to ensure that the new materials and measures were effective and did not impair participant understanding or performance of the task. Additional new manipulations designed to modify test selection behavior were not included in the pilot study.

Experiment Design

Importance of Context

In the original MDT Microworld task (Mynatt, Doherty, & Tweney, 1977), participants were asked to interact with a microworld of particles and objects. An image representing a test case (see Figure 1.1) included a circle and crosshair in the top left of the image and one or two objects usually located near the center of the image. The objects had one of three possible shapes (disc, triangle or square) and one of two possible brightness levels (50% or 100%). Participants fired a particle across the screen and observed whether the particle moved through the objects on screen or was stopped near the objects by an invisible wall. Participants were provided no context or explanation of what the objects and particles might represent.

Providing a concrete real-world context has been shown to improve participant performance in confirmation bias problems as seen in variants of the Wason selection task (Griggs & Cox, 1982). McKenzie (2006) also used familiar materials to increase participant sensitivity to the differential diagnosticity of tests. In the current proposed research, the Mynatt, Doherty, and Tweney microworld task has been reframed in a familiar context: the struggle of predator and prey. Given a scenario with real-world meaning, participant's may be able to better comprehend the problem, recognize the need for –H tests, and reduce the number of positive test selections compared to the original, less concrete scenario.

In order to assess the impact of scenario, two scenarios were selected for the current research: a replication of the original MDT Microworld scenario and a new scenario with thematic content that would improve participant's understanding of the nature of the task by giving them a familiar context for the problem. The new scenario required the development of new materials including a main theme (scorpions and spiders), a story for presenting the theme, instructions describing the scenario, and figures to represent potential experimental tests.

MDT Microworld. Participants were given a description of the microworld: the types of objects and the interaction between the objects and the particles. Only the selection of tests of the hypothesis was replicated. Participants were asked a series of questions about possible experiments that could be used to test a given hypothesis. Participants were not given the opportunity to actually run the tests.

Scorpion-Spider scenario. In the new scenario, the objects were replaced by scorpions and the particle was replaced by a spider. Instead of different shapes and colors, the scorpions had different claws and tails. The goal in the new scenario was to determine a rule that describes which scorpions would eat the spider. The thematic content of the Scorpion-Spider scenario should have allowed participants to bring some common knowledge to bear on understanding the problem when asked to select screens to test the given hypothesis. The pilot study was used to assess the materials associated with the new scenario to ensure that participants understood the new scenario as well as the original MDT Microworld.

Providing the Initial Hypothesis

In the original Mynatt, Doherty and Tweney (1977) study, approximately 50% of the participants generated an alternative hypothesis that made it impossible to analyze the pattern of test selection. The pairs of test screens were designed with the assumption that participants would approach them with a particular hypothesis. By giving participants an initial hypothesis, the design avoided losing participants that did not propose the expected initial hypothesis but risked introducing an alternative explanation for any significant reduction in confirmation bias. Prior research has shown that other-generated hypotheses may be met with more skepticism than self-generated hypotheses. Schunn & Klahr (1993) suggest that the increased skepticism leads participants to more thoroughly investigate the hypothesis but does not appear to significantly impact the research process. Instead, participants appeared to simply spend more time selecting and running tests. Although the initial hypothesis may change participant test selection behavior slightly, the benefit of providing a specific initial hypothesis was considered more valuable.

Participants

Twenty-six undergraduate students were recruited from the Mississippi State University Psychology Department research pool. Participants received credit as partial fulfillment of classroom requirements. Data was collected in experimental sessions that included no more than 9 participants. All participants completed the experiment in less than 1 hour. All procedures were reviewed and approved by the Mississippi State University Institutional Review Board (IRB Docket #07-154; see Appendix F for copy of IRB approval letter).

Apparatus and Materials

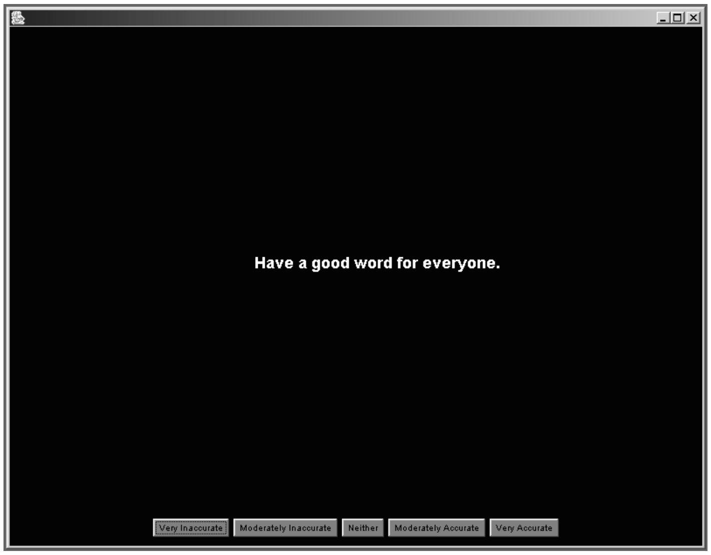
The pilot study was presented in three parts: two questionnaires and one computer-based personality inventory.

Questionnaire Materials

The questionnaires consisted of a question packet and a separate answer sheet (See Appendix B for copies of the pilot study questionnaires and answer sheets). Each questionnaire began with one page of instructions that described the scenario, explained how the images represented possible experiments, gave the initial hypothesis, and provided basic guidance for responding to the questions. Following the instructions, participants were asked a series of questions regarding possible experimental tests of the hypothesis given in the instructions. The questionnaires included five types of questions.

IPIP Personality Inventory

The IPIP personality inventory was administered via a computer application. The IPIP application was developed in Java 1.4 and was similar in presentation on the different personal computer platforms used for the experiment. The platforms included two iMac G3s, 4 PowerMac G4s, 2 PowerMac G4s, and 1 Dell PC. Figure 2.1 is a screenshot of the Java application. Responses were made by using a mouse to select one of five response buttons across the bottom of the window.



Have a good word for everyone.

Figure 2.1. Screenshot of IPIP application presenting an inventory item and five options for participant response.

Procedure

Prior to participant arrival, 36 random participant identification numbers between 1 and 500 were generated using a computer script. Before each session, all materials were setup at the computer stations and each station was assigned a participant ID. Upon arrival, participants were asked to seat themselves at one of the computer stations. The participant ID was used to associate the personality inventory data with the responses from the paper questionnaire.

After participants provided consent, participants were given basic instructions verbally. The verbal instructions described the content of the paper questionnaires, the personality inventory, and the correct order to complete the experiment (questionnaire #1, personality inventory, then questionnaire #2). The final verbal instruction directed participants to open questionnaire #1 to the instruction page and to begin.

Questionnaire #1 was the MDT Microworld scenario for all participants. After participants completed the first packet, they were instructed to complete the International Personality Item Pool inventory on the computer at their station. At the end of the personality inventory, participants were instructed to complete questionnaire #2 (the Scorpion-Spider scenario).

Questionnaire Procedure

The questionnaire packets consisted of a page of instructions and several pages of questions. The pilot study question packets included five different question types: basic event prediction questions, positivity assessments, catch trials, explicit positive selections, and hypothesis test selections. The five question types were blocked and presented in five sections. There were a total of 23 questions per packet. The only differences between the two scenarios were the experiment figures used in the questions and the specifics of the questions. The following sections describe the contents of the packets in detail (also, see Appendix B for the complete contents of the pilot study questionnaires and answer sheets).

Instructions. On the first page of the question packet, participants were presented with one page of instructions. The instructions included an introduction to the scenario, a description of the participant's goals for the packet, an example situation that described how images were used to represent possible experiments, a current hypothesis, an example question, and some general guidance for completing the packet.

Participants were introduced to the scenario by a description of a scientist investigating a particular phenomenon. In the MDT Microworld scenario, the scientist is attempting to understand which objects stop particles and which do not. In the Scorpion-Spider scenario, the scientist is attempting to understand which scorpions are attacking and consuming a particular species of spider.

The scenario description introduces the basic idea of the scenario, the scientist's goal, and the features that differentiate the different objects in the scenario. In both scenarios, the objects are differentiated by two features: one with two levels and one with three levels. In the MDT Microworld scenario, objects have shape and color features. Objects may be of three different shapes (triangle, square, or disc) and two different colors (white or black). In the Scorpion-Spider scenario, scorpions have claw and tail features. Scorpions have one of three claw types (pincher, serrated, or thick) and one of two tail types (up-turned or down-turned). (See Figure 2.2.)

The instructions explained that the participant's goal was to assist the scientist by considering certain experiments and answering questions about the possible experiments. Participants were also advised that they would be asked different questions about the experiments.

Participants were presented with an example situation that introduced the use of figures to depict possible experiments. The instructions included an example figure (see Figure 2.3) and an explanation of how the figure represents two objects with particular features in an experimental setting. In the MDT Microworld scenario, the instructions explain the role of an arrow present in the experiment figures. The arrow represents the entry point and path of a particle in the experiment. Additionally, participants were presented with a limitation of the experiments: when two objects (or scorpions) are in one experimental setting and an event occurs, the scientist cannot discern which object (or scorpion) was responsible for the event.

Following the example experiment figure, participants were given a hypothesis and were asked to adopt the hypothesis as they answered the questions contained in the questionnaire packet.

In the pilot study, participants were given no specific instruction. Participants were simply asked to give the best advice possible to the scientist by carefully considering the questions before answering. Finally, participants were instructed to direct any questions to the experimenter or to begin to answer the questions.

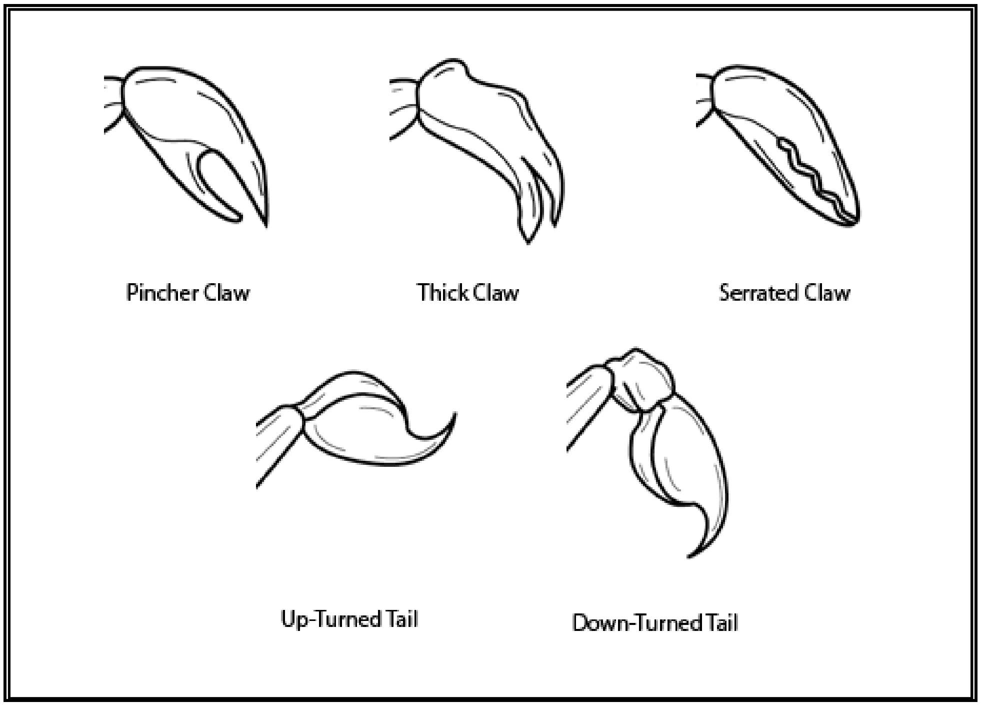


Figure 2.2. The claw features (top row) and tail features (bottom row) that differentiated the six different scorpion types in the pilot study.

Questions. Each pilot study questionnaire consisted of 23 questions. Each question asked participants to make judgments with regard to one or two experiments related to the scenario presented in the instructions. The questionnaires included five types of questions presented in sections: basic event prediction questions, positivity assessments, catch trials, explicit positive selection questions and hypothesis test selection questions.

The basic event prediction questions required participants to consider a single possible experimental test and determine if, based on the current hypothesis, the phenomenon of interest (particle stopped or spider eaten) would occur. Participants were instructed to circle “yes” on the answer sheet if they believed it would occur and “no” if they believed it would not occur.

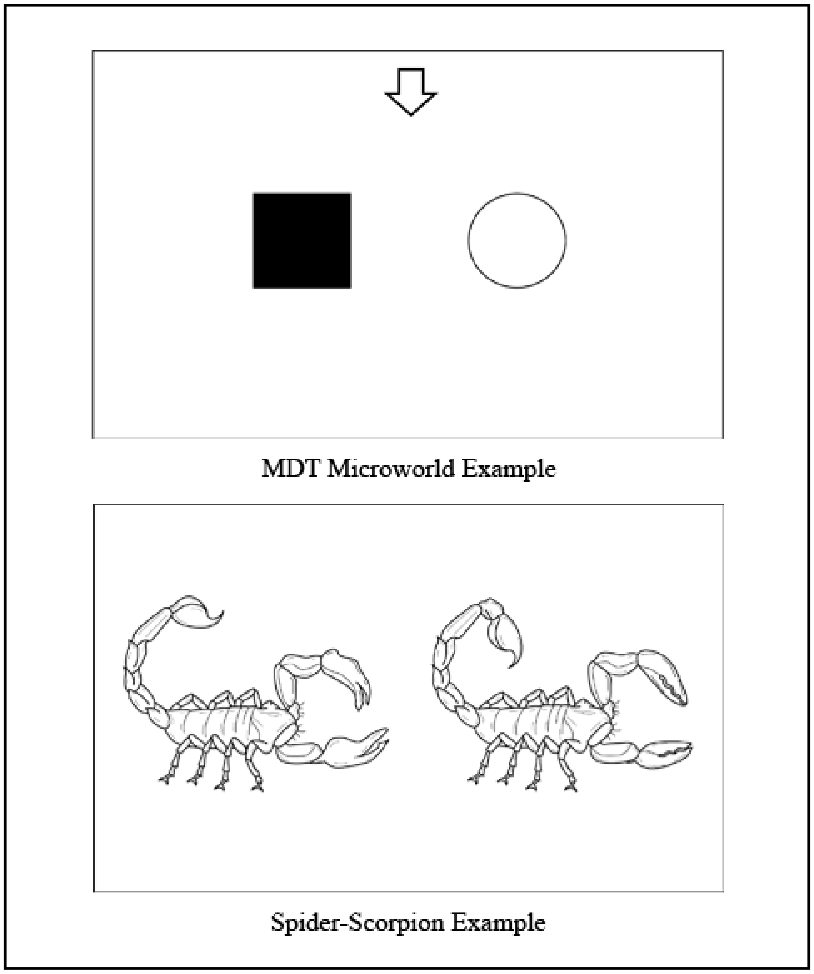


Figure 2.3. The experiment figures given to participants in the scenario instructions (Top: MDT Microworld scenario, Bottom: Scorpion-Spider scenario). The example experiment figures for the scenarios did not include a +H instance.

The basic event prediction questions provide a probe of participants’ understanding of the task and the given hypothesis. The questions are simple tests of fundamental awareness of the features of the figures and which features are hypothesized to lead to a positive event. When participants are unable to respond accurately to the basic event prediction questions, it is likely that the participants either do not understand the materials or do not understand the given hypothesis. Participant responses to these simple questions provide an indication of their level of understanding of the materials and the hypothesis. There were five basic event prediction questions in the packet.

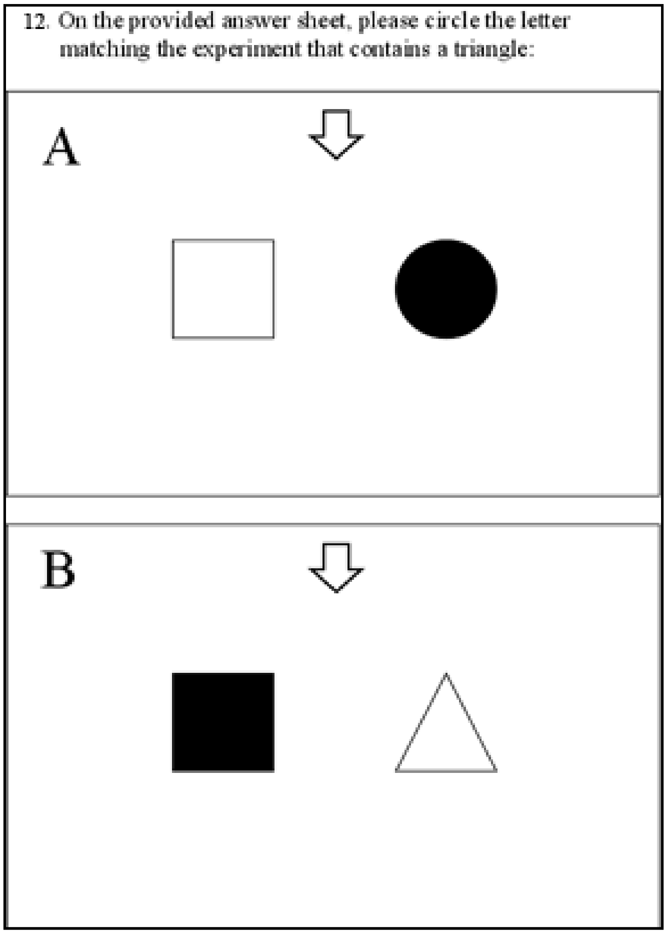
The positivity assessments required participants to consider a single experiment figure and estimate the likelihood that the phenomenon of interest would occur if the experiment were performed. Participants were instructed to respond by making a mark on a visual analog scale (VAS) on the answer sheet. The visual analog scale was presented as a horizontal line 127 mm in length with a “0” anchor above the left end of the line and a “100” anchor above the right end of the line.

The positivity assessment questions require participants to explicitly state the likelihood of a positive event occurring for each figure in the questions. By requiring participants to provide these quantitative, figure by figure assessments, I may be able to determine what factors participants use to assess the positivity of a given figure. This will provide a deeper understanding of the process participants use to assess possible tests for selection. Participants estimated the likelihood of a positive event for six experiment figures.

The catch trials in the third section of questions asked participants to consider two experiment figures and report which figure included an object with a particular feature (MDT Microworld: object with a triangular shape, Scorpion-Spider: scorpion with pincher claws). In the catch trials, the feature used to define the hypothesized rule was the target feature participants were asked to find in the figures. Figure 2.4 is an example of an MDT Microworld catch trial. The two tests were labeled “A” and “B” and participants were instructed to circle the letter on the answer sheet that matched the selected test from the question.

The catch trials are extremely simple assessments of participants’ basic understanding of the fundamentals of the task: comprehension of the figures. Participants are told explicitly by name to search for a specific feature in the two figures and identify which figure contains the feature. Performance on these questions provides a measure that can be used to determine if a participant completely misunderstood the task or is not fully engaged in the experiment. There were two catch trial questions in the questionnaire.

The explicit positive selection questions asked participants to consider two experiment figures and select the figure that was most likely to result in a positive event based on the given hypothesis. In other words, participants selected the experiment that was most likely to result in the particle stopping or the spider being eaten. The two figures were labeled “A” and “B” and participants were instructed to circle the letter on the answer sheet that matched the selected figure for the question.



12. On the provided answer sheet, please circle the letter matching the experiment that contains a triangle:

Figure 2.4. An example catch trial from the third section of questions in the MDT Microworld scenario questionnaire packet. The participant is asked to identify the experiment that contains a triangle. The correct response is 'B.'

The explicit positive test selection questions are included for comparison to two other measures: First, the positivity assessment for individual figures can be used to predict figure selection based on positivity. If the individual positivity assessment is not predictive of explicit positive test selection, then participants may be using one process to assess positivity of individual figures and another to choose between two figures. Second, the test selections in the explicit positive test selection questions can be compared to the hypothesis test selection questions to assess whether participants are making the same decisions for explicit positive test selection and for hypothesis test selection. Participants were asked to perform five explicit positive selections.

The hypothesis test selection questions asked participants to consider two possible experiment figures and select the figure that represented the most effective test of the given hypothesis. The two tests were labeled “A” and “B” and participants were instructed to circle the letter on the answer sheet that matched the selected figure.

The hypothesis test selection questions are the key questions for investigating positive test selection when attempting to test a particular hypothesis. I expected the proportion positive test selection to be affected by scenario in the pilot study and other treatments in the full study. If behavior on hypothesis test selection is modified by one of our treatments, the other questions will probe participants comprehension of the task at different levels and provide a deeper understanding of the process of hypothesis test selection than in other methodologies. Participants were asked to respond to five hypothesis test selection questions.

Personality Inventory Procedure

The personality inventory consisted of a screen of instructions and a listing of 100 items (see Appendix A) taken from the International Personality Item Pool (“International Personality Item Pool,” 2007). The International Personality Item Pool is an effort to develop a public-domain personality measure (Goldberg et al., 2006). Before the items were presented, participants were presented with a screen of instructions:

> “You will see a series of phrases describing people’s behavior. For each statement, you will need to select from five alternative buttons to indicate how accurately the statement describes you. Describe yourself as you generally are now, not as you wish to be in the future. Be honest: how do you feel you compare to other people you know of the same sex as you are, and roughly your same age?
>
> Please read each statement carefully, and then click on the button that best describes your choice.
> Click on the ‘Continue’ button below to begin the personality inventory.”

After clicking on the ‘Continue’ button, inventory items were presented in the IPIP application window with a row of five response buttons along the bottom of the window. The response buttons were labeled from left to right: “Very Inaccurate, Inaccurate, Neither, Accurate, Very Accurate”. Participants were instructed to respond by clicking the button that described how accurately the item described themselves.

Personality inventory items. Participants responded to 100 items. As described in Chapter I, the 100 items were sorted into 10 scales based on scales found in the International Personality Inventory Pool (IPIP). Five of the scales were 10-item scales similar to the Big Five personality scales in the NEO PI-R (Costa & McCrae, 1992). The

10-item scales organized (“International Personality Item Pool,” 2007) by the International Personality Item Pool for Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness were included in the personality inventory.

The other five scales were scales that were face valid for scientific discovery: cautiousness, creativity, inquisitiveness, insight and intellect. In the IPIP scale index (“International Personality Item Pool,” 2007) each named scale lists components from other constructs. For example, the IPIP scale for cautiousness links to two scales for cautiousness: one scale is similar to the cautiousness facet of the Abridged Big Five-Dimensional Circumplex (AB5C; Hofstee, de Raad, & Goldberg, 1992) and one scale is similar to the cautiousness facet of the NEO-PI-R domain (Costa & McCrae, 1992). In this case, I selected items from the NEO-PI-R facet. For the creativity scale, I selected items from each of the scales listed on the IPIP index for creativity/originality. These scales crossed four constructs: AB5C, Hogan Personality Inventory (HPI; Hogan Assessment Systems, 2007), HEXACO Personality Inventory (Ashton, Lee, & Goldberg, 2006), and Values in Action (VIA; Peterson & Seligman, 2004). The inquisitiveness scale was similar only to constructs from the HEXACO personality inventory. The insight scale was similar to constructs in Gough’s California Psychological Inventory (CPI; Gough & Bradley, 1996). The intellect scale was similar to constructs in AB5C, NEO-PI-R domains (Costa & McCrae, 1992), CPI (Gough & Bradley, 1996), and the 16 Personality Factor Questionnaire (16PF, Cattell & Schuerger, 2003).

All of the items from the selected IPIP scales were inserted into the personality inventory list. Duplicate items were removed. To bring the total number of items to 100, items that were considered less relevant (e.g., “Do crazy things”) were also removed.

As discussed in Chapter I, previous research has indicated that individual differences may predict the success of participants on scientific discovery tasks. The previous research reviewed has focused on cognitive ability, education and experience, or divergent thinking. The current research is investigating the impact of personality traits on performance in scientific discovery. The NEO-PI-R domains provide a broad assessment of personality whereas the face valid scales selected from other IPIP scales provide an assessment of specific personality traits that I expected would be associated with performance on a scientific discovery task. A full listing of the inventory items is provided in Appendix A.

The complete listing of IPIP items was randomly permuted to mix the presentation of items from the selected scales. The IPIP items were presented to all participants in the same order. After participants responded to the last item, the IPIP application instructed participants to begin questionnaire packet #2.

Questionnaire packet #2 contained the Scorpion-Spider scenario for all participants. The structure, content, and procedures for the questionnaire packet have been previously described. After completing questionnaire packet #2, participants were given the opportunity to ask the experimenter questions and were given a debriefing sheet that described the purpose of the research.

Results

All 26 participants completed both questionnaires and the IPIP personality inventory.

Response Accuracy

Response accuracy was calculated for the participants for basic event prediction questions and catch trials. Response accuracy was calculated by determining the number of correct responses for a question type divided by the total number of responses for a question type.

For basic event prediction questions, a correct response was defined as a “yes” response when the experiment figure in the question represented a +H test (if the experiment was performed, a positive event would be expected according to the given hypothesis) and “no” otherwise. For the MDT Microworld scenario, a correct response was “yes” if a triangle was depicted in the experiment figure and “no” otherwise. For the Scorpion-Spider scenario, a correct response was “yes” if a scorpion with pincher claws was depicted in the experiment figure and “no” otherwise.

For the catch trials, a correct response was defined as selection of the one experiment figure from the two given figures that included the feature requested in the question. For the MDT Microworld scenario, participants were asked to identify the figure containing a triangle. The correct response was selection of the experiment figure that contained a triangle. For the Scorpion-Spider scenario, participants were asked to identify the figure containing a scorpion with pincher claws. The correct response was selection of the figure that contained the pincher claws.

A paired-samples t test was used to determine if accuracy differed across the two scenarios. The results for response accuracy on basic event prediction questions and catch trials are presented in Table 2.1.

Table 2.1

Mean Response Accuracy for Basic Event Prediction Questions and Catch Trials for MDT Microworld and Scorpion-Spider Scenarios (N = 26)

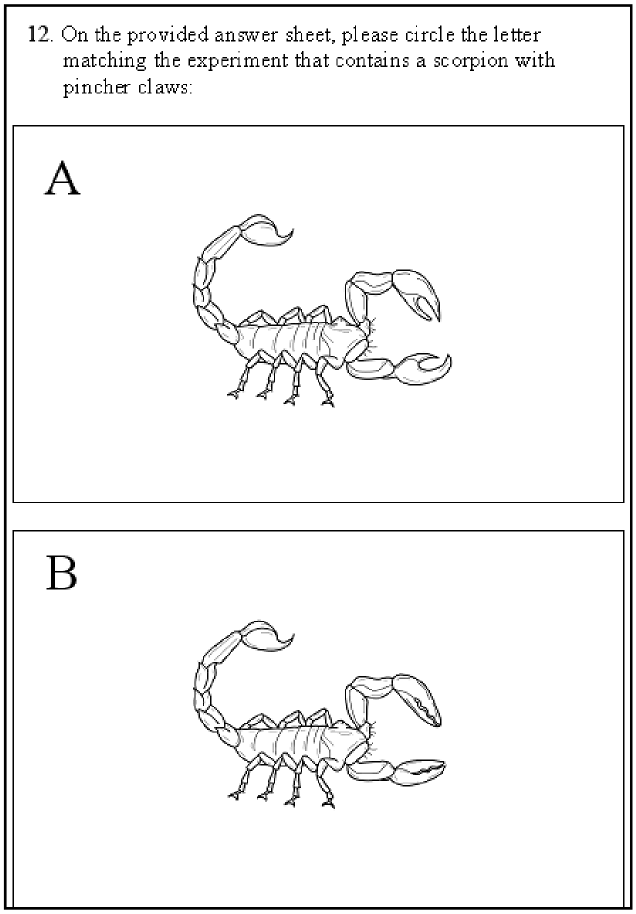
	MDT Microworld		Scorpion-Spider		Paired-Samples t-test	
	Accuracy	*SD*	Accuracy	*SD*	Diff	*t*
Basic event prediction	.9308	.1692	.8385	.2467	.0923	1.59
Catch trials	.9808	.0981	.8077	.3762	.1731	2.214*

* *p* < .05.

There was no statistical difference in response accuracy for the MDT Microworld scenario (*M* = 93%, *SD* = .17) and the Scorpion-Spider scenario (*M* = 83.9%, *SD* = .25), *t* (25) = 1.594, *p* = .123 (two-tailed), on the basic event prediction questions. Both groups performed reasonably well on the basic event prediction questions.

For the catch trials in the MDT Microworld scenario, participants were again highly accurate in their responses (*M* = 98.1%, *SD* = .09). However, participants were significantly less accurate in their responses to the catch trials for the Scorpion-Spider scenario (*M* = 80.7%, *SD* = .37), *t* (25) = 2.214, *p* = 0.036 (two-tailed), *d* = 0.63.

The catch trials (questions #12 and #13) were designed to determine if participants understood the most basic aspects of the scenarios. If participants failed to respond accurately to the catch trials, they either misunderstood fundamental elements of the task or were not fully engaged in the experiment.



12. On the provided answer sheet, please circle the letter matching the experiment that contains a scorpion with pincher claws:

Figure 2.5. Scorpion-Spider catch trial (question #12). Experiment figure A is a single scorpion with an up-turned tail and pincher claws. Experiment figure B is a single scorpion with an up-turned tail and serrated claws.

In the MDT Microworld scenario, only 1 participant out of the 26 total participants responded incorrectly to a catch trial. All other participants responded correctly to both catch trials. It appears that, in the MDT Microworld scenario, participants were able to correctly identify the requested feature by name. This is not surprising as the target feature was the triangle shape.

In the Scorpion-Spider scenario, 15% of the participants responded incorrectly to both catch trials and 23% of participants incorrectly responded to at least one trial when asked to identify which test contained a scorpion with a pincher claw. These inaccurate participants are the source of the difference in response accuracy on catch trials for the Scorpion-Spider scenario in comparison to the MDT Microworld.

Figure 2.5 shows the two figures used in one of the catch trials (question #12) in the Scorpion-Spider scenario. The upper figure (A) depicts a single scorpion with an up-turned tail and pincher claws. The lower figure (B) depicts a single scorpion with an up-turned tail and serrated claws. The catch trial question asked the participant: "On the provided answer sheet, please circle the letter matching the experiment that contains a scorpion with **pincher claws**" (emphasis added). The correct response was to select figure A because the scorpion in that figure has pincher claws. However, 19% of the participants incorrectly reported that figure B contained the pincher claw.

In order to understand why a number of participants failed to accurately respond to the catch trials, I examined participants' performance on the Scorpion-Spider basic event prediction questions. The basic event prediction questions are similar to the catch trial questions because participants are asked to determine whether the figure contains the feature that leads to a positive event according to the hypothesis. I expected inaccurate participants in the catch trials to show a similar pattern in the basic event prediction questions.

Figure 2.6 shows the figure given in one of the basic event prediction questions (question #1). Participants were asked to predict whether the spider would be eaten given the experiment. The scorpion depicted in Figure 2.6 has an up-turned tail and pincher claws. The up-turned tail is irrelevant given the hypothesis (‘spiders are being eaten by scorpions with pincher claws’). However, the scorpion in Figure 2.6 has pincher claws indicating that it is likely, given the hypothesis, that the scorpion would eat the spider. Participants should have responded “yes” indicating that the spider would be eaten given the hypothesis. 77% of the participants responded correctly to question #1. Five of the six participants that incorrectly responded to question #1 also incorrectly responded to the Scorpion-Spider catch trials.

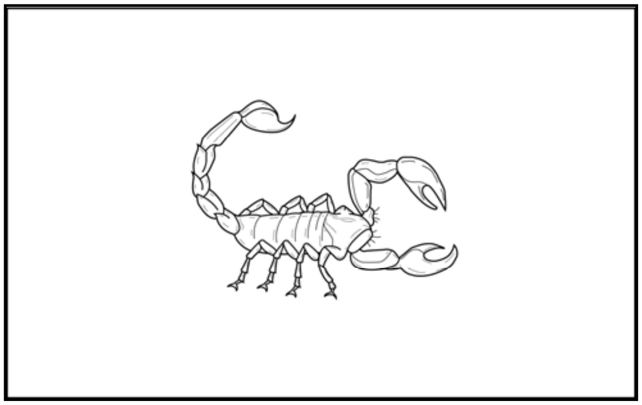


Figure 2.6. Scorpion-Spider basic event prediction figure (question #1) representing a possible experimental test. The figure contains a single scorpion with an up-turned tail and pincher claws. Participants that responded incorrectly to the catch trials also responded incorrectly to this trial.

Figure 2.7 shows another figure given in a basic event prediction question (question #2). Neither of the two scorpions in the figure has the pincher claws identified in the hypothesis. The correct response to the question is “no” indicating that the spider is not likely to be eaten in the test represented by the figure given the pincher claw hypothesis. Three of the five participants that missed both of the catch trials responded incorrectly to question #2.

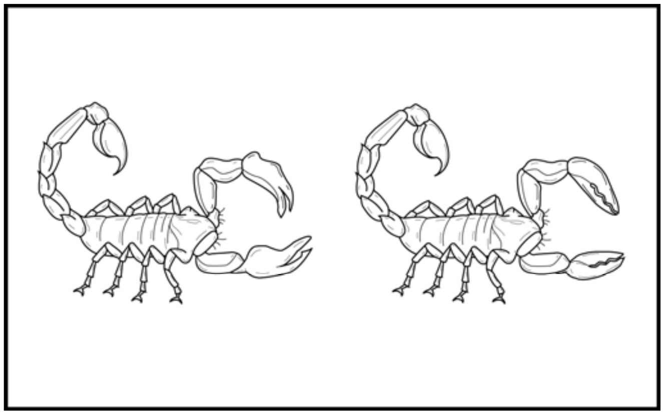


Figure 2.7. Scorpion-Spider basic event prediction figure (question #2) representing a possible experimental test. The figure contains two scorpions. The left scorpion has a down-turned tail and thick claws. The right scorpion has a down-turned tail and serrated claws. Half of the inaccurate participants incorrectly responded that this figure is likely to result in a positive event.

The responses of the inaccurate participants on the catch trials and the basic event prediction questions suggested that they were not correctly identifying scorpions with pincher claws. Based on the pattern of results, it appears that the inaccurate participants confused the serrated claws (as seen on the right scorpion in Figure 2.7) with the pincher claws. Thus, the participants believed that the scorpion shown in Figure 2.6 did not have the claws referred to by the hypothesis and specified in the catch trial questions. Instead, these participants responded to the basic event prediction questions and the catch trials questions as if the serrated claws on the scorpion on the right of Figure 2.7 were the target feature.

The six participants that missed catch trials in the Scorpion-Spider scenario were dropped and analyses performed on the revised data set. The results are given in Table 2.2. After removing the inaccurate participants from the data set, response accuracy on the catch trials in the Scorpion-Spider scenario (*M* = 100%, *SD* = .00) was not significantly different from response accuracy on the catch trials in the MDT Microworld scenario (*M* = 97.5%, *SD* = .11), *t* (19) = -.237, *p* = .815. Similarly, response accuracy on the basic event prediction questions in the Scorpion-Spider scenario (*M* = 94%, *SD* = .11) was not significantly different from accuracy on basic event prediction questions in the MDT Microworld scenario (*M* = 93%, *SD* = .175), *t* (19) = -1.0, *p* = .815, *d* = .07, following removal of the inaccurate participants.

Table 2.2

Mean Response Accuracy for Basic Event Prediction Questions and Catch Trials for MDT Microworld and Scorpion-Spider Scenarios, Excluding Inaccurate Participants (n = 20)

	MDT Microworld		Scorpion-Spider		Paired-Samples t-test	
	Accuracy	*SD*	Accuracy	*SD*	Diff	*t*
Basic event prediction	.9300	.1750	.9400	.1143	-.0100	-.237
Catch trials	.9750	.1118	1.000	.0000	-.0250	-1.0

Although it is trivial that by dropping all of the participants with inaccurate responses in the Scorpion-Spider scenario there was now no difference between groups in the catch trials and there was no significant change in the relationship for the basic event prediction questions. Excluding the inaccurate participants, the basic event prediction questions revealed reasonably high accuracy rates (93% and 94%) and no difference between the scenarios. This suggests that the remaining participants understood the fundamentals of both scenarios: the content of the figures and the given hypotheses. Additionally, dropping these participants did affect the analyses of the other question types.

Positivity Assessment Questions

In the positivity assessment questions, participants were presented with a single figure and asked to estimate the likelihood that the experimental test would result in a positive event. Participants were instructed to indicate their estimate by making a vertical mark on a visual analog scale (VAS), a 126 mm line with a 0 anchor on the left and a 100 anchor on the right. However, 39% of the participants did not respond correctly to the VAS questions. Some participants wrote in a fixed value over the VAS line, other participants circled the 0 or 100 anchors, and other participants marked multiple points on the VAS line. Given the high rate of error and difficulty in interpreting the intention of the responses on the VAS questions, the data was not analyzed for the pilot study. The key result was that participants were not familiar with VAS questions and additional instructions related to the VAS questions would be needed for the full study.

Positive Test Selection

In the explicit positive selection and hypothesis test selection questions, participants were presented with a pair of experiment figures and asked to select one of the figures. In explicit positive selection, participants were asked to select the figure with the highest positivity, or highest likelihood, of a positive event. In the hypothesis test selection questions, participants were instructed to select the figure that represented the experimental test that would most effectively test the given hypothesis. Positive test selection is defined as the proportion of +H tests selected when the other potential selection was a -H test. For both the explicit positive selections and the hypothesis test selections, positive test selection was calculated by counting the number +H test selections and dividing by the total number of questions that included a +H test and a –H test. There were three questions in each section that presented participants with a choice between a +H and a –H test.

For the explicit positive selection questions, I expect to find no difference between the scenarios because the selection is explicitly positive. However, for the hypothesis test selection questions, if the thematic content provided additional memory cues or a mental framework that allowed participants to better understand the task and to select appropriate tests, I would expect an increase in selection of negative tests in the Scorpion-Spider scenario compared to the MDT Microworld.

Given the inaccuracy in the catch trials, the inaccurate participants were dropped from the following analyses. The results of the analyses after dropping the inaccurate participants are shown in Table 2.3.

As expected, there was no significant difference in positive test selection for the explicit positive selection questions between the MDT Microworld scenario (*M* = 98.3%, *SD* = .07) and the Scorpion-Spider scenario (*M* = 98.3%, *SD* = .07), *t* (19) = .00, *p* = 1.0, *d* = 0. There was also no significant difference in positive test selection for the hypothesis test selection questions between the MDT Microworld scenario (*M* = 91.6%, *SD* = .23) and the Scorpion-Spider scenario (*M* = 83.3%, *SD* = .29), *t* (19) = 1.0, *p* = .330. Based on this result, there was no effect of scenario on positive test selection in the pilot study.

Table 2.3

Positive Test Selection for Explicit Positive Test Selection and Hypothesis Test Selection Questions for MDT Microworld and Scorpion-Spider Scenarios, Excluding Inaccurate Participants (n = 20)

	MDT Microworld		Scorpion-Spider		Paired-Samples t-test	
	Bias	*SD*	Bias	*SD*	Diff	*t*
Explicit Positive Test Selection	.9833	.0745	.9833	.0745	.00	.000
Hypothesis Test Selection	.9167	.2388	.8333	.2962	.0833	1.00

Discussion

The results of the pilot study provided three main points of information: First, participants were clearly having problems understanding the materials. This was especially apparent in the Scorpion-Spider scenario when 23% of the participants responded incorrectly and in the positivity assessment questions when 39% of participants failed to appropriately use the visual analog scale. Second, there was no effect of thematic content on positive test selection for hypothesis test selection. Third, participants were completing two packets of questions and a 100-item personality inventory in less than half of the time allotted for the experiment.

Based on these three points, modifications were made to the materials and procedures for the full study. First, the labels and renderings of the scorpion claws in the experiment figures were modified to make it easier to visually differentiate between the three types of claws and easier to match the label to the rendering. Second, new video-based instructions were created that standardized the verbal instructions, demonstrated proper response using the visual analog scale, and explicitly named the three types of claws. Third, instructions were inserted between each section of questions and included restatements of the given hypothesis and the instruction condition manipulation. Finally, the number of questions in each question packet was also expanded from 23 questions to 63 questions to provide a more thorough coverage of the experimental figures because participants were completing the pilot study questionnaires so quickly.

Despite the lack of a result for the effect of thematic content, the overall theme of the Scorpion-Spider scenario was not modified significantly. The absence of an effect of thematic content was believed to be due to the general instructions and the issues with the materials (figures and questions) rather than poor scenario selection.

CHAPTER III

METHOD

The purpose of the experiment was to probe a number of factors that, based on previous research, I believe may impact positive test selection in a realistic scientific discovery task. The experimental design used in the full study further extended and modified the original Mynatt, Doherty and Tweney (1977) microworld. In addition to modifications based on the results of the pilot study, three novel instruction conditions were created to investigate the impact of high miss costs, of a violation of an assumption underlying the positive test strategy, and of an alternative hypothesis. The experiment is a 2x2x6 factorial design with one within-subjects factor (scenario) and two between-subjects factors (presentation order and instruction condition).

The two scenarios used in the pilot study (MDT Microworld and Scorpion-Spider scenario; see Chapter II) were used again in the full study to investigate the effect of thematic content on test selection. Modifications were made to the scenarios based on the results of the pilot study and are described in detail in following sections. Despite the nonsignificant result of thematic content in the pilot study, I still expected participants to select more –H tests in the Scorpion-Spider scenario than in the MDT Microworld given the improved materials and the addition of instruction conditions that were expected to work in synergy with the thematic content.

Participants were assigned to one of six instruction conditions designed to investigate the effect of various treatments (described in detail in following sections) on test selection. Three of the instruction conditions are replications from the original Mynatt, Doherty, and Tweney (1977) research. In the original study, the general instructions to *test*, *confirm*, or *disconfirm* the hypothesis did not lead to significant differences in positive test selection. I expected to replicate the Mynatt, Doherty, and Tweney (1977) results for the original instruction conditions.

However, I added three instruction conditions designed to increase –H test selection. First, I included a *high miss cost* instruction condition that, based on Klayman's (1995; Klayman & Ha, 1987) and Friedrich's (1993) claims, was expected to increase –H test selection. Second, I included a *common event* instruction condition that informed participants that one of the assumptions underlying the positive test strategy (Klayman, 1995; Klayman & Ha, 1987) had been violated. If participants' use of the positive test strategy was adaptable, a possibility that has apparently not been empirically studied, participants were expected to use more –H tests than used in the original MDT instruction conditions. Third, I included an *alternative hypothesis* instruction condition that provided participants with the same primary hypothesis as the other instruction conditions and a second, alternative hypothesis. The use of alternative hypotheses has been shown to increase –H test generation and improve performance (e.g., Dual-Goal task; Tweney et al., 1980). I expected participants given an alternative hypothesis to select more –H tests of the primary hypothesis than the other instruction conditions. However, I also expected that the –H tests selected by these participants were selected only because those tests were +H tests of the alternative hypothesis.

Participants

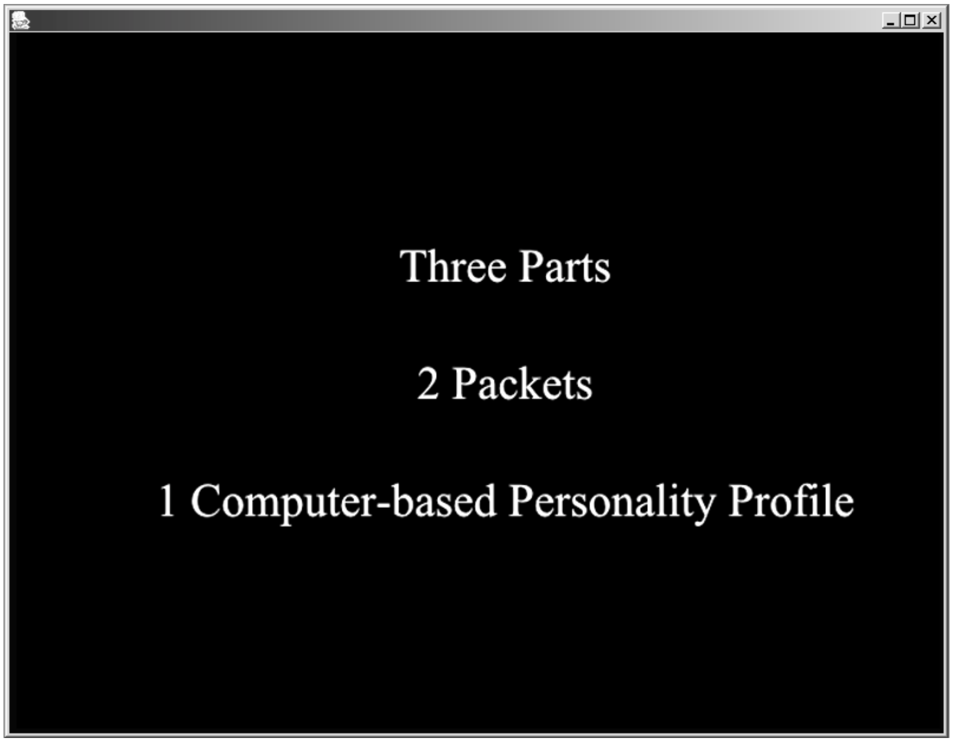
One hundred and eighty-three undergraduate students were recruited from the Mississippi State University Psychology Department research pool. Participants received credit as partial fulfillment of classroom requirements. Participants were randomly assigned to one of six instruction conditions: test (*n*=31), confirm (*n* =29)*,* disconfirm (*n* =28)*,* high miss cost (*n* =31)*,* common event (*n* =32)*,* and alternative hypothesis (*n* =32). Data was collected in experimental sessions that included no more than 9 participants. All participants completed the experiment and required no longer than 1 hour and 30 minutes. All procedures were reviewed and approved by the Mississippi State University Institutional Review Board (IRB Docket #07-154; see Appendix F for copy of IRB approval letter).

Apparatus and Materials

The study was presented in three parts: two paper questionnaire packets and one computer-based personality inventory.

Experiment Application

In the pilot study, the IPIP application was used to present the IPIP personality inventory to participants. The original application was extended to create an experiment application that provided embedded audio/video instructions that covered general instructions and specific instructions for each questionnaire in addition to the personality inventory. The audio/video instructions were embedded Quicktime movies generated in Quicktime PRO. Quicktime movies were generated for the general instructions and for each instruction condition. See Appendix D for transcripts and screenshots taken from the audio/video instructions. The movies contained text slides, photographs, video clips, and audio narration. The photographs and video clips were recorded using a Kodak C613 digital camera. The audio narration was recorded via a microphone connected to a Dell Inspiron E1405 laptop running Windows Vista. The Quicktime API for Java was used to load and play the audio/video instructions for the appropriate instruction conditions.



Three Parts

2 Packets

1 Computer-based Personality Profile

Figure 3.1. Screenshot of experiment application presenting introductory instructions. Participants were presented with audio-video recordings of general instructions and scenario-specific instructions before completing the questionnaire packets.

The IPIP application was presented on different personal computer platforms for different participants. The platforms included two iMac G3s, 4 PowerMac, 2 PowerMac G4s and 1 Dell PC. Participants were required to wear headphones during the presentation of recorded instructions. The experiment application was developed in Java 1.4 and was similar in presentation across the different hardware platforms. Figure 3.1 is a screenshot of the application during the presentation of the introductory instructions.

Questionnaire Materials

The basic format for the questionnaires was unchanged from the pilot study (see Chapter II for a detailed description of the pilot study questionnaires). Each questionnaire for the full study included a question packet and a separate answer sheet (See Appendix C for copies of the study questionnaires and answer sheets). Each question packet began with one page of instructions that described the scenario, explained how the figures represented possible experiments, gave the initial hypothesis, and provided basic guidance for responding to the questions. Following the instructions, participants were asked to consider a series of questions regarding possible experimental tests. The question packet was split into sections for the five different question types described in Chapter II: basic event prediction, positivity assessment, catch trials, explicit positive selection and hypothesis test selection.

Although the basic format of the questionnaires was the same as the pilot study questionnaires, details of the content were changed in an effort to overcome some of the issues identified in the pilot study and to fully investigate all of the factors of interest. The total number of questions per questionnaire was increased from 23 to 63 because the previous number of questions did not provide enough data for a thorough analysis and participants were able to complete the study in less than half of the time allotted. New versions of the drawings and labels of the scorpion claws were created to reduce the confusion that participants exhibited in the Scorpion-Spider scenario in the pilot study. Also, new video-based instructions were recorded to reduce confusion and ensure that participants responded correctly to all of the question types.

Beyond the modifications made based on the results of the pilot study, the six instruction condition treatments were added to fully evaluate the effect of different treatments on test selection behavior. Three of the instruction conditions were replications of the original Mynatt, Doherty, and Tweney study (1977) and three were new treatments expected to modify test selection behavior based on previous research in scientific discovery.

Additional questions. In the pilot study, participants were given two hours to complete the two questionnaires and the IPIP personality inventory. All of the participants easily completed the entire experiment within one hour. Additionally, our analysis of the pilot data revealed that there were too few questions of each type for a thorough analysis. For example, with only two catch trials, if a participant responded incorrectly to one trial, it was difficult to discern whether the participant simply made a mistake on one trial or was actually a poorly performing participant. Also, the five questions in the explicit positive selection and the hypothesis test selection question types covered only five of the 351 possible combinations of experiment figures that participants could be tested.

Given these issues, I took the opportunity to increase the number of questions. The basic event prediction questions (participants predicted whether an experiment would result in a positive or negative event) were increased from five to eight questions. The positivity assessment questions (participants rated the likelihood that the event would happen for a given figure on a visual analog scale) were increased from six to 18 positivity assessment questions. This increase allowed participants to rate the positivity for almost every individual figure they would be asked about in other sections of the questionnaire. The number of catch trials (participants identified which figure from a pair of figures included an object or scorpion with a specific feature) was doubled from two to four. The explicit positive selection questions (participants selected the experiment that was most likely to result in a positive event) were increased from five to 13 questions. The number of hypothesis test selection questions (participants selected the experiment that would most effectively test the given hypothesis) increased from 5 to 20 questions in the revised questionnaires.

In the pilot study, the five question types were blocked together in sections but there was no explicit separation between the different question types. In other words, there was nothing in the question packet informing the participant that they had completed a section of questions and were about to begin another section of questions.

Given the proximity and similarity of the explicit positive selection questions and the hypothesis test selection questions, the revised question packets included a page between each section that explicitly notified the participant that the question type was about to change. The section break pages also included a restatement of the given hypothesis and the instruction condition.

Modifications to the figures. In the pilot study, participants appeared to be confused by the rendering and the labeling of the scorpion claws (see Figure 2.2 to review the design of the pilot study stimuli). 23% of the pilot study participants were dropped from the analyses. An analysis of the pattern of response by the inaccurate participants indicated that the participants may have confused the pincher claws with the serrated claws.

The labels and the drawing of the scorpion claws were reviewed to identify possible improvements. The pilot study scorpion claws were labeled “thick”, “pincher”, and “serrated”. The “pincher” label described the functionality of the claw. The “serrated” label described the specific shape of the interior edge of the claw. The “thick” label described a general feature of the claw shape. The labels did not all describe a similar dimension and, without all three claws present in a figure, it was difficult to determine which claw the label referred to. In addition to the label problems, the drawings of the claws were deemed to be too similar. There were very few features that defined the difference between the claws. Also, participants were never given instructions that explicitly showed the three claw shapes with their labels.

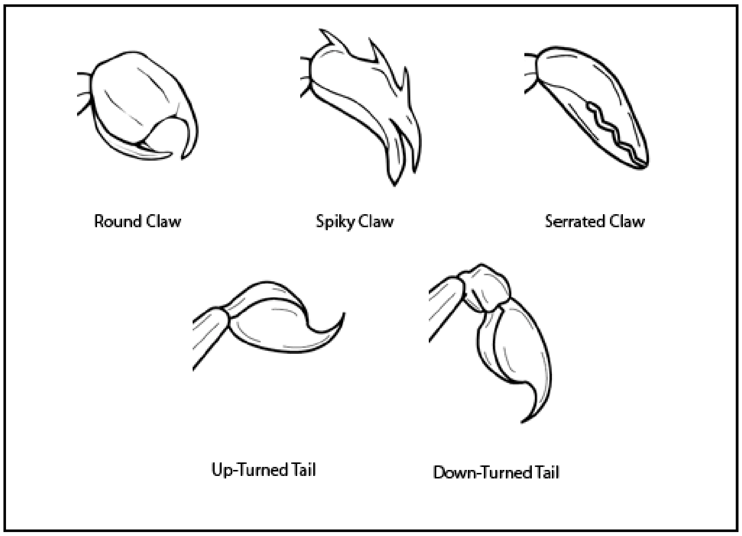


Figure 3.2. Revised designs and labels for the scorpion claws. The distinctive features of the claws were exaggerated and the claws were given more appropriate labels that all described the shape of the claw.

In order to remedy these issues, the renderings of the scorpion claws were modified to create defining features that were expected to make it easier to differentiate between the claws. The pincher claw from the pilot study was given a more rounded shape and a larger gap between the teeth of the claw. The revised claw was labeled the "round" claw. The new label refers to the general shape of the claw rather than the function of the claw. Spikes were added to the thick claw from the pilot study to provide a defining visual feature. The revised claw was labeled the "spiky" claw. Again the label focused on a specific visually defining feature of the claw. The serrated claw was not modified. Figure 3.2 shows the revised scorpion claw renderings and the associated labels.

Instructions. The pilot study provided participants with only basic guidance about how to answer the questions in order to focus on examining participant's basic understanding of the fundamental features of the experiment. In this experiment, participants were assigned to one of six instruction conditions: test, confirm, disconfirm, high miss cost, common event, and alternative hypothesis. The instructions on the first page and on the section break pages of the question packets were specific to each of the instruction conditions. Participants received the same instruction treatment for both the MDT Microworld and the Scorpion-Spider scenario.

The test, confirm, and disconfirm instruction conditions were based on the instruction conditions in the original Mynatt, Doherty, and Tweney (1977) study. The test and confirm conditions acted as control conditions for the experiment. Mynatt, Doherty, and Tweney (1977) reported that participants in the test and confirm instruction conditions selected 71% of the possible confirmatory test selections. In the test condition, participants were encouraged to select tests that would "effectively test theories and hypotheses." In the confirm condition, participants were encouraged to select tests that would "provide supporting evidence in order to confirm their theories and hypotheses." These two conditions provided a baseline to determine whether the treatment instructions had an effect on test selection or estimates of the likelihood of an event occurring in a given test case.

In the disconfirm condition, participants were encouraged to falsify the given hypothesis by selecting test cases that would lead to the falsification of the current hypothesis. Mynatt, Doherty, and Tweney (1977) reported that this condition was not significantly different from the test and confirm instruction conditions.

The remaining instruction conditions were developed based on previous research investigating the nature of confirmation bias. The instructions in these conditions were similar to the test condition. However, additional information that was expected to influence one or more of the factors for confirmation bias was provided to participants. The instructions were tailored for the different scenarios.

The high miss cost condition was designed to inform participants that it was more important to reduce missed targets than false alarms. According to Klayman (1995) and Friedrich (1993), participants in this condition should use more –H tests to ensure that the set of hypothesized tests includes all of the target instances. In the MDT Microworld scenario, the high miss cost instruction condition informed the participant that "if the scientist can determine what is stopping the particles, he can use this knowledge to develop a new treatment for cancer." The intent of the instruction was to suggest to participants that the scientist needs to know all the cases where the particle is stopped. In other words, the participants should do their best to avoid missing a case where the particle will be stopped. In the Scorpion-Spider scenario, participants in the high miss cost instruction condition were informed that the spider was "beneficial to pest control" and the scientist "must discover which scorpions are eating the spiders to avoid significant loss for the farmers." The intent of the instruction was to suggest to participants that if they miss a type of scorpion that eats the spiders then the cost to the farms will be significant.

The common event condition informed participants that one of the assumptions (uncommon event assumption; see Chapter I) underlying the positive test strategy was violated. In both scenarios, participants were told that the events of interest (particle stoppage and spiders being eaten) were common. Specifically, in the MDT Microworld scenario, participants were told that in preliminary experiments the particles were regularly stopped. In the Scorpion-Spider scenario, participants were told that spiders rarely survived long in fields once the scorpions were introduced. If participants were adaptive to violations of the assumptions and were able to interpret the instructions, participants would be expected to select more –H tests.

The alternative hypothesis condition attempted to modify positive test selection in a manner similar to Tweney et al. (1980) by adding a second hypothesis for consideration. Unlike Tweney et al. (1980), the alternative hypothesis was not the complement of the primary hypothesis. There were four combinations of positivity and negativity given the primary and the alternative hypothesis. Some tests were –H for both hypothesis and some tests were +H for both hypotheses. Other tests were +H for one hypothesis and –H for the other. If consideration of alternative hypotheses generally lead to increased -H test selection, participants were expected to select some tests that were –H tests of both hypotheses. If an alternative hypothesis simply encouraged selection of a set of +H tests that happened to be –H tests of the primary hypothesis, participants should select tests that were +H tests for at least one hypothesis but few tests that were –H tests of both hypotheses. In the MDT Microworld scenario, participants were told that there was an alternative hypothesis suggesting that it may be the presence of black objects and not the triangular shape that stopped the particle. In the Scorpion-Spider scenario, participants were told it may be the scorpions with the down-turned tails rather than those with the round claws that were eating the spiders. In this condition, participants were told to approach the test cases with both hypotheses in mind.

Procedure

Group Assignment

Prior to participant arrival, each computer station was randomly assigned a participant ID, instruction condition, and a scenario presentation order. Presentation of the scenarios was counterbalanced. Participant IDs were assigned to an MDT Microworld, Scorpion-Spider (PS) or Scorpion-Spider, MDT Microworld (SP) presentation order. The PS group was given the MDT Microworld scenario first and the SP group was given the Scorpion-Spider scenario first. The appropriate materials for the instruction condition and presentation order were set up at the computer stations. The materials included an informed consent form, two questionnaire packets labeled #1 and #2, two answer sheets, and the experiment application. The participant ID was used to associate the personality inventory data with the responses from the paper questionnaires.

Upon arrival, participants were asked to seat themselves at one of the available computer stations thereby assigning themselves a random participant ID and to an instruction condition and presentation order. After all participants were present (or 5 minutes after the announced start time), participants were asked to provide consent. After participant consent was completed, participants were given basic instructions verbally. The verbal instructions described the basic structure of the experiment (audio-video instructions, first questionnaire, computer-based personality inventory, audio-video instructions, and second questionnaire) and explained the use of the experiment application and headphones for viewing the audio-video instructions. The final verbal instruction directed participants to don their headphones and click “Start” to begin the first audio-video instructions in the experiment application.

General Instructions

The first audio-video instructions were approximately four minutes of general instructions that described the basic structure of the experiment, the contents of the questionnaire packets, and the contents of the IPIP personality inventory (see Appendix D for a transcript and screenshots from the audio-video instructions).

In the pilot study, 39% of participants did not understand how to correctly respond to the visual analog scales used for the positivity assessment questions. In the general instructions video, the five types of questions are described and participants are shown how to respond to the questions on their answer sheets. The general instructions video also described the personality inventory and how to respond to each item. At the end of the general instructions video, participants are given the opportunity to replay the general instructions or press the 'continue' button to continue the experiment.

Questionnaire Procedure

Before beginning the first questionnaire, participants watched audio-video instructions specific to the questionnaire they were about to complete. For example, a participant in the test instruction condition and the SP presentation order watched an instruction video for questionnaire #1 that described the Scorpion-Spider scenario and included the admonition to test the given hypothesis. A participant in the alternative hypothesis instruction condition and the PS presentation order watched an instruction video for questionnaire #1 that described the MDT Microworld scenario, provided the primary hypothesis and the alternative hypothesis, and instructed the participant to keep both hypotheses in mind as they tested the given hypothesis.

As participants watch the audio-video instructions, they were asked to open their packet #1 to the first page. The paper questionnaires included general instructions, a description of the scenario and several pages of questions. See Appendix C to review the questionnaire packets and answer sheets and Appendix D for transcripts and screenshots from the questionnaire audio-video instructions.

Instructions. As in the pilot study, participants were presented with one page of instructions on the first page of the package. The instructions included an introduction to the scenario, a description of the participant's goals for the packet, an example situation that described how images were used to represent possible experiments, a current hypothesis, an example question, and some general guidance on how to complete the packet.

Participants were introduced to the scenario by a description of a scientist investigating a particular phenomenon. In the MDT Microworld scenario, the scientist is attempting to understand which objects stop particles and which do not. In the Scorpion-Spider scenario, the scientist is attempting to understand which scorpions are attacking and consuming a particular species of spider.

The scenario description introduced the basic idea of the scenario, the scientist's goal, and the features that differentiated the different objects in the scenario: object shape and color for the MDT Microworld and scorpion claw and tail shape for the Scorpion-Spider scenario.

The instructions explained that the participant's goal was to assist the scientist by considering certain experiments and by answering different questions about the possible experiments. Participants were given the same example situation as used in the pilot study (see Figure 2.3) that introduced the use of figures to depict possible experiments. The instructions included the example figure and an explanation of how the figure represented two objects with particular features in an experimental setting.

Following the example experiment figure, participants were given a hypothesis (two hypotheses in the alternative hypothesis condition) and were asked to adopt the hypothesis as they answered the questions contained in the questionnaire packet. As described previously, additional comments were added to the instructions depending on the instruction condition.

Questions. Each questionnaire included 63 questions. Each question asked participants to make judgments related to the scenario presented in the instructions. As in the pilot study, participants were presented with five types of questions: basic event prediction questions, positivity assessments, catch trials, explicit positive selections and hypothesis test selections. The questions were explicitly split into sections with a page inserted between sections. The inserted pages informed the participant that the question type was changing, reminded the participant about the given hypothesis (or hypotheses), and restated the instruction condition manipulation. All of the question types from the pilot study were retained. The only change was in the number of questions asked in each section: basic event prediction (8), positivity assessment (18), catch trials (4), explicit positive test selection (13), and hypothesis test selection (20).

Personality Inventory Procedure

After participants completed the first questionnaire, they returned to the computer and the experiment application. On screen, participants were told to press ‘Continue’ once they completed the first questionnaire. When participants selected continue, the IPIP personality inventory was presented. The procedure for the IPIP personality inventory was identical to the procedure used in the pilot study. The first screen of the IPIP personality inventory consisted of short instructions describing how to consider and how to respond to the upcoming IPIP items. Participants read the instructions and then responded to 100 IPIP items taken from the International Personality Item Pool as described in Chapter I (see Appendix A for a listing of the IPIP items). Participants responses ranged from “Very Inaccurate, Inaccurate, Neither, Accurate, Very Accurate”. The response described how accurately the participant felt the IPIP item on-screen described them.

Personality inventory items. As described in Chapter II, the 100 IPIP items were sorted into 10 scales based on scales found in the International Personality Inventory Pool (IPIP). Of the ten scales used in the personality inventory, five of the scales were 10-item scales similar to the Big Five personality scales in the NEO PI-R. The other five scales were scales that were face valid for scientific discovery: cautiousness, creativity, inquisitiveness, insight and intellect. See Appendix A for a full listing of the inventory items.

The complete listing of IPIP items was randomly permuted to mix the presentation of items from the selected scales. The IPIP items were presented to all participants in the same order. After participants responded to the last item, the experiment application instructed participants to begin questionnaire packet #2.

Dependent Measures

Participants were scored based on their responses to the five question types in the questionnaire and on their responses to the personality inventory. From the questionnaire, participants were scored for response accuracy on the basic event prediction questions and the catch trial questions. For the positivity assessment questions, participants' responses were scored for deviation from the expected values given the participants' hypotheses. For the explicit positive selection questions, two measures were calculated: First, for each pair of figures, the expected response was calculated based on the given hypotheses. For example, if figure A was positive on H1 and figure B was negative on H1, the expected response would be to select figure A. Participants' actual selections were compared to the expected selections to assess how well the hypothesis predicted explicit positive test selection. Second, the expected response was calculated based on participant responses to the positivity assessment questions. Participants' actual selections were compared to the expected selections given their positivity assessment to assess how well individual positivity assessment predicted explicit positive test selection. For the hypothesis test selection questions, participants' responses were scored to assess how often participants selected a positive test when given the choice between a positive and a negative test.

For the personality inventory, the inventory items represented 10 scales: neuroticism, extraversion, openness to experience, agreeableness, conscientiousness, cautiousness, creativity, inquisitiveness, insight, and intellect. Participants received a score from 0 to 100 for each scale based on their responses to the personality inventory.

CHAPTER IV

RESULTS

All participants (*N* = 183) completed both questionnaires and the IPIP personality inventory.

Catch Trials

The catch trials were simple questions that asked participants to search two figures representing experimental tests and find the one figure that contained a particular feature. For the MDT Microworld scenario, participants were asked to identify which figure included a triangle-shaped object. For the Scorpion-Spider scenario, participants were asked to identify which figure included a scorpion with a round claw. Participants reported which of the two figures contained the desired feature on their answer sheet. The goal of the catch trial questions was to determine whether participants understood a fundamental element of the task: interpretation of the figures.

Scoring response accuracy. Each question was scored as correct when the participant accurately reported the figure that contained the desired feature. Missing items were scored as incorrect responses. Response accuracy on the catch trials was defined as the proportion correct of all of the catch trial questions (total number of correct responses divided by total number of catch trials). Participants were presented with four catch trials per scenario (twice the number of catch trials used in the pilot study).

For the catch trials, I expected to find no effect of scenario, presentation order, or instruction condition on participant performance. An effect of scenario would indicate a potential problem in the scenario materials as indicated by the catch trial results in the pilot study.

Determining Inaccurate Participants

In the pilot study (see Chapter II), response accuracy on the catch trials was significantly lower in the Scorpion-Spider scenario. Modifications were made to the scenario materials in an attempt to avoid participant confusion in the full study. Twenty-four participants (13%) were identified as inaccurate participants (cutoff at or below 50% accuracy). These participants were marked to be dropped from further analyses of this data. Table 4.1 lists the number of inaccurate participants by presentation order and instruction condition.

In the pilot study, 23% of the participants were dropped. Although the modifications to the study materials reduced the number of inaccurate participants, 13% represents a large number of inaccurate participants. The large number of inaccurate participants suggests that the modifications following the pilot study did not completely eliminate the problems affecting participant performance.

Table 4.1

Number of Participants Dropped From Between-Subject Conditions Using 50% Cutoff on Catch Trial Response Accuracy

Factor	Level	Number of dropped subjects
Order	MDT, Scorpion-Spider	11
	Scorpion-Spider, MDT	13
Instruction	Test	6
	Confirm	5
	Disconfirm	2
	High Miss Cost	4
	Common Event	4
	Alternative Hypothesis	3

ANOVA of mean response accuracy including inaccurate participants. A repeated-measures ANOVA with one within-subjects factor (scenario) and two between-subjects factors (presentation order and instruction condition) was performed to determine which, if any, of the factors affected response accuracy on the catch trials. The sphericity assumption was assessed for each ANOVA and violations will be noted when they were present and affected the results.

Given the nature of the catch trials, I expected no effect of scenario, presentation order, or instruction condition. However, given the issues with the Scorpion-Spider scenario in the pilot study, a scenario effect would suggest that one of the scenarios was more likely than the other to lead to inaccurate responses on the catch trials. Table 4.2 lists the results of the repeated-measures ANOVA.

There was no significant main effect of presentation order or instruction condition. The interaction between presentation order and instruction condition was also not significant. The main within-subjects effect of scenario was significant, *F* (1, 171) = 8.728, *p* = .004, η^2_G = .023. There was no significant interaction between scenario and the other factors.

Table 4.2

Repeated-Measures Analysis of Variance of Mean Response Accuracy on Catch Trials, Including Inaccurate Participants (N = 183)

Source	*df*	*F*	η^2_P	η^2_G [a]	*P*
			Between Subjects		
Order (O)	1	.106	.001	.000	.745
Instruction (I)	5	.347	.010	.006	.884
O × I	5	.613	.018	.010	.690
S within-group error	171	(.047)			
			Within Subjects		
Scenario (S)	1	8.728	.049	.023	.004*
S × O	1	.033	.000	.000	.856
S × I	5	.172	.005	.002	.973
S × O × I	5	.859	.024	.011	.510
S × *S* within-group error	171	(.039)			

Note. Values in parentheses indicate mean square errors. *S* = subjects.
a = Generalized eta squared (η^2_G) is a relatively new measure of effect size proposed for analysis of variance including repeated measures (Olejnik & Algina, 2003; Bakeman, 2005). Bakeman proposes that generalized eta squared effect sizes should be interpreted as follows: .02 is small, .13 is medium, and .26 is large. Partial eta squared (η^2_P) is reported for familiarity and completeness. See Appendix E for a discussion of, and methods for calculating, η^2_G.
*$p < .05$

Catch trial response accuracy was higher in the MDT Microworld scenario (*M* = 96.9%, *SD* = .15) than in the Scorpion-Spider scenario (*M* = 90.9%, *SD* = .24). 18 of the 24 (75%) inaccurate participants fell below the cutoff only in the Scorpion-Spider scenario. Despite the modifications made following the pilot study, some aspect of the Scorpion-Spider scenario still led to a significant level of error in response accuracy for the Scorpion-Spider scenario compared to response accuracy for the MDT Microworld scenario. The inaccurate participants were marked for exclusion from further analyses.

ANOVA of mean response accuracy excluding inaccurate participants. In order to assess the effect of any of the factors on accurate responses to the catch trials, a second repeated measures ANOVA with one within-subjects factor (scenario) and two between-subjects factors (presentation order and instruction condition) was performed on catch trial response accuracy, excluding the inaccurate participants. The results of the repeated-measures ANOVA are listed in Table 4.3.

The between-subjects effects remained insignificant. As expected, the main effect for scenario was no longer significant (F (1, 171) = .413, p = .521) following the removal of the most inaccurate participants. However, in the new analysis, the interaction between scenario and presentation order was significant (F (1, 171) = 4.051, p = .046, η^2_G = .013). Figure 4.1 graphically depicts the Scenario × Presentation Order interaction.

The Scenario × Presentation Order interaction indicates that catch trial performance for the MDT Microworld scenario and the Scorpion-Spider scenario depended on the order of presentation. Pairwise comparisons using the Sidak adjustment for multiple comparisons were performed on the simple effects of the Scenario × Presentation Order interaction. The comparisons indicate that there was an effect of presentation order on catch trial response accuracy for the Scorpion-Spider scenario (p <

.031). Catch trial response accuracy for the Scorpion-Spider scenario was higher for participants in the PS presentation order group (MDT Microworld, Scorpion-Spider scenario; *M* = 100%, *SD* = .00) than for participants in the SP presentation order group (Scorpion-Spider scenario, MDT Microworld; *M* = 98.4%, *SD* = .06). The effect was small but suggests a possible learning effect where participants apply their experience from the MDT Microworld scenario to the apparently more confusing Scorpion-Spider scenario thereby improving their response accuracy on the catch trials.

Table 4.3

Repeated-Measures Analysis of Variance of Mean Response Accuracy on Catch Trials, Excluding Inaccurate Participants (n = 159)

Source	*df*	*F*	η^2_P	η^2_G	*P*
			Between Subjects		
Order (O)	1	1.906	.013	.007	.170
Instruction (I)	5	1.529	.049	.024	.184
O × I	5	.417	.014	.007	.836
S within-group error	147	(.001)			
			Within Subjects		
Scenario (S)	1	.413	.003	.002	.521
S × O	1	4.051	.027	.013	.046*
S × I	5	.714	.024	.011	.614
S × O × I	5	2.230	.070	.037	.054
S × *S* within-group error	147	(.001)			

Note. Values in parentheses indicate mean square errors. *S* = subjects.
**p* < .05

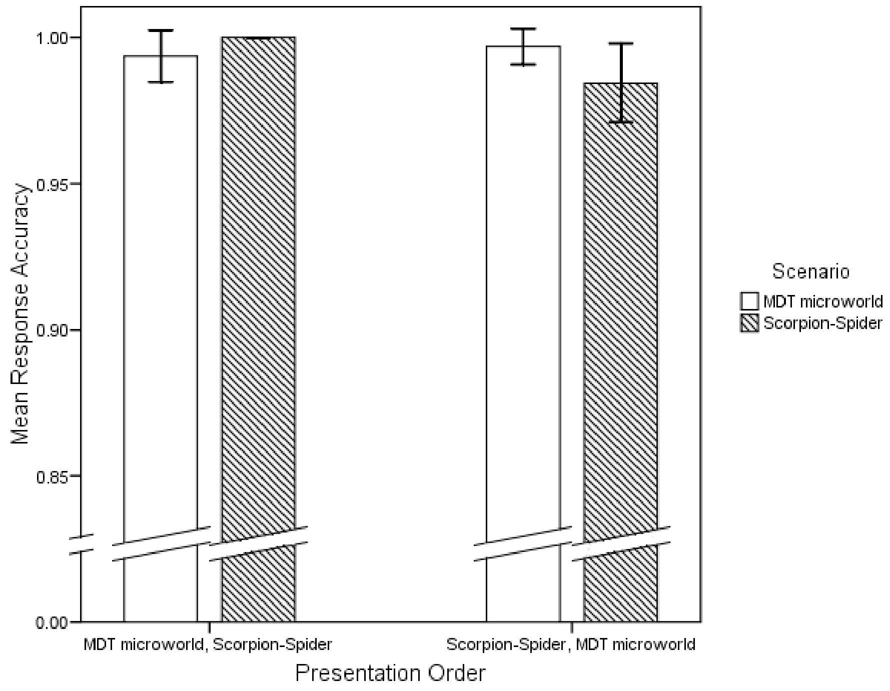


Figure 4.1. Mean catch trial response accuracy for presentation order by scenario interaction, excluding inaccurate participants. Error bars indicate 95% confidence interval.

Summary of Catch Trials

As in the pilot study, a large number (13%) of participants responded incorrectly to the catch trial questions. The catch trial questions were simple questions that tested participants' fundamental understanding of the features in the figures used to represent experimental tests. Failure to accurately respond to these simple questions could have indicated that certain participants were simply not engaged in the task. However, the significantly lower accuracy for the Scorpion-Spider scenario combined with the fact that 75% of the inaccurate participants fell below the cutoff only in the Scorpion-Spider scenario suggests that the reason for low response accuracy was the Scorpion-Spider scenario materials.

After the inaccurate participants were removed from the analysis, the overall scenario effect was not significant. However, the higher mean response accuracy for the Scorpion-Spider scenario for participants in the PS presentation order group compared to participants in the SP presentation order group suggests the previous experience in the MDT Microworld reduces the impact of the issues with the Scorpion-Spider scenario.

Basic Event Prediction

The catch trial questions were designed to reveal participants who were unable to respond accurately to questions based on fundamental elements of the task: recognizing specific features of the figures. The basic event prediction questions assessed participant comprehension of the figures and the relationship between the figures and the hypothesis or hypotheses they were given. If a participant was accurate on the catch trials but inaccurate on the basic event prediction questions, it is likely that they did not understand how the hypothesis (or hypotheses) related to the figures.

The basic event prediction questions required participants to predict, based on the given hypothesis (or hypotheses), whether a positive event would occur if the experiment rendered in the figure were to take place. Participants reported “yes” if the experiment would result in the particle being stopped in the MDT Microworld scenario or in the spider being eaten in the Scorpion-Spider scenario. Participants reported “no” otherwise.

Scoring response accuracy. Each basic event prediction question was scored for correctness depending on the hypothesis given to the participant. Participants were presented with eight basic event prediction questions per scenario. Table 4.4 lists the figures presented in the basic event prediction questions and the positivity for the figures given the primary hypothesis (H1) and the alternative hypothesis (HA). The H1HA column in Table 4.4 also lists the overall positivity given both H1 and HA.

Table 4.4

Basic Event Prediction Questions and Positivity for the Given Hypotheses

Question	MDT Microworld				Scorpion-Spider			
	Figure	H1	HA	H1HA	Figure	H1	HA	H1HA
1	WSWT	P	N	P	DEDS	N	P	P
2	WSWD	N	N	N	DSUE	N	P	P
3	BD	N	P	P	DR	P	P	P
4	WT	P	N	P	URDE	P	P	P
5	WTBT	P	P	P	DRUR	P	P	P
6	WS	N	N	N	DRDS	P	P	P
7	WSBD	N	P	P	DS	N	P	P
8	WDBT	P	P	P	UE	N	N	N

Note. Figure columns describe contents of figures by feature. For MDT Microworld: up to two objects are defined by color and shape where W – white, B – black, S – square, D – disc, and T – triangle. For Scorpion-Spider: up to two objects are defined by tail direction and claw shape where U – Up-turned, D – down-turned, S – spiky, E – serrated, R – round. N indicates a negative test, P indicates a positive test.

Response accuracy on the basic event prediction questions was defined as the proportion correct of their responses to the basic event prediction questions (total number of correct responses divided by total number of questions).Participants' responses were scored as correct when the response for a question matched the appropriate response for the hypothesis or hypotheses they were given. Missing items were not scored.

For participants in most of the instruction conditions, responses to the basic event prediction questions were expected to correspond to the H1 column in Table 4.4. For participants in the alternative hypothesis instruction condition, responses were expected to correspond to the H1HA column in Table 4.4. If participants in the alternative hypothesis instruction condition were scored only with respect to H1, their response accuracy would be inappropriately reduced by failing to account for the difference between the appropriate H1 and H1HA responses. This reduced accuracy would suggest a main effect of instruction condition when there may, in fact, be no effect (See Appendix F for an analysis of unadjusted mean response accuracy on catch trials). The same issue existed for scoring other participants against the appropriate HA responses. By scoring the participants with respect to the hypothesis or hypotheses they were given, an appropriate comparison of the mean response accuracy of the instruction conditions was possible.

For the basic event prediction questions, I expected no significant effect of scenario, presentation order, or instruction condition. The questions required only comparison of the features of the figures to the features specified by the hypothesis and determination of whether the hypothesis predicted a positive result or a negative result. An effect of scenario would indicate a possible problem with the scenario materials especially if the effect indicated lower response accuracy on the Scorpion-Spider scenario given the issues with the scenario in the pilot study and in the catch trial questions.

Dropping Inaccurate Participants

As previously described, 24 participants (13%) were inaccurate on the catch trials. The inaccurate participants were dropped from the analyses of basic event prediction. See Appendix F for an analysis of the basic event prediction questions including the inaccurate participants.

ANOVA of hypothesis adjusted mean response accuracy excluding inaccurate participants. A repeated-measures ANOVA with one within-subjects factor (scenario) and two between-subjects factors (presentation order and instruction condition) was performed on the hypothesis adjusted mean response accuracy for the basic event prediction questions. The inaccurate participants were excluded from the analysis.

With the adjustment for the alternative hypothesis, I expected there to be no main effect of the instruction condition on mean response accuracy. I also expected no main effect of presentation order or scenario and no interactions between factors. The results of the ANOVA are listed in Table 4.5.

The main effects of scenario, instruction condition, and presentation order were not significant. The interaction between presentation order and instruction condition was also not significant.

Table 4.5

Repeated-Measures Analysis of Variance of Hypothesis Adjusted Mean Response Accuracy on Basic Event Prediction, Excluding Inaccurate Participants (n = 159)

Source	*df*	*F*	η^2_P	η^2_G	*p*
			Between Subjects		
Order (O)	1	.324	.002	.001	.570
Instruction (I)	5	.804	.027	.017	.549
O × I	5	.978	.032	.021	.433
S within-group error	147	(.032)			
			Within Subjects		
Scenario (S)	1	1.581	.011	.004	.211
S × O	1	14.099	.088	.034	$.000^*$
S × I	5	11.987	.075	.029	$.040^*$
S × O × I	5	8.705	.056	.021	.129
S × *S* within-group error	147	(.018)			

Note. Values in parentheses indicate mean square errors. *S* = subjects.
**p* < .05

The Scenario × Presentation Order interaction was significant, *F* (1, 147) = 14.099, $p < .001$, $\eta^2_G = .034$, and the Scenario × Instruction Condition interaction was also significant, *F* (5, 147) = 2.397, $p = .040$, $\eta^2_G = .029$. The three-way interaction between scenario, presentation order, and instruction condition was not significant. Figure 4.2 graphically depicts the Scenario × Presentation Order interaction.

The Scenario × Presentation Order interaction suggests a small learning effect as mean response accuracy increased from the first presented scenario to the second presented scenario. Pairwise comparisons using the Sidak adjustment for multiple comparisons were performed on the simple effects of the Scenario × Presentation Order interaction. For the SP group (Scorpion-Spider, MDT Microworld), mean response accuracy increased significantly (*p* = .001) from the first scenario (*M* = 87.8%, *SD* = .18) to the second scenario (*M* = 95.7%, *SD* = .13). For the PS group (MDT Microworld, Scorpion-Spider), mean response accuracy increased from the first scenario (*M* = 91.1%, *SD* = .18) to the second scenario (*M* = 95%, *SD* = .12) but the difference did not rise to the level of significance (*p* = .08).

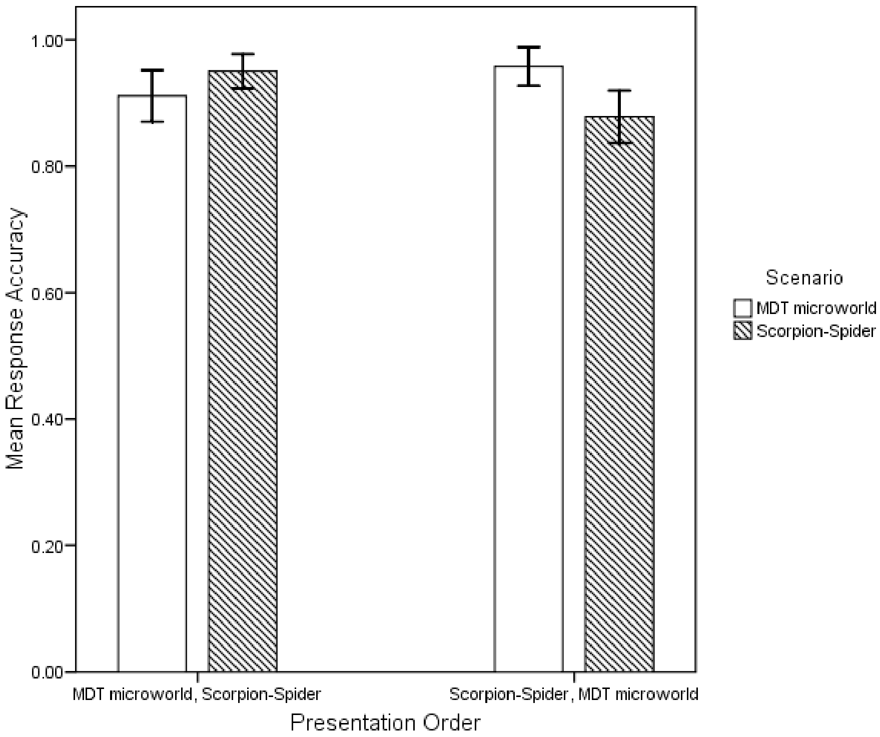


Figure 4.2. Hypothesis adjusted mean response accuracy on basic event prediction questions for presentation order by scenario interaction, excluding inaccurate participants. Error bars indicate 95% confidence interval.

The Scenario × Instruction Condition interaction indicates that basic event prediction for the MDT Microworld scenario and the Scorpion-Spider scenario depended on the instruction condition.

In Figure 4.3, the difference in performance between the MDT Microworld and the Scorpion-Spider scenario for participants in the alternative hypothesis instruction condition appeared to be the source for the interaction. Pairwise comparisons using the Sidak adjustment for multiple comparisons were performed on the simple effects of the Scenario × Instruction Condition interaction. Participants in the alternative hypothesis instruction condition were significantly less accurate (*p* = .001) in the Scorpion-Spider scenario (*M* = 82.3%, *SD* = .23) than in the MDT Microworld scenario (*M* = 94.4%, *SD* = .16).

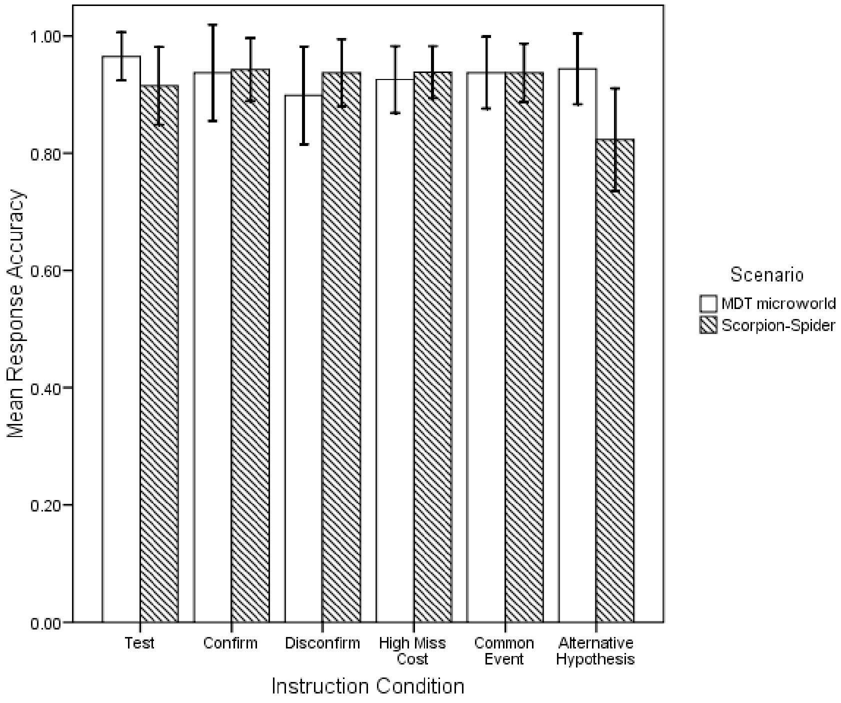


Figure 4.3. Hypothesis adjusted mean response accuracy on basic event prediction questions for instruction condition by scenario interaction, excluding inaccurate participants. Error bars indicate 95% confidence interval.

Participants given the alternative hypothesis appeared to have significant issues interpreting which figures were positive tests of one or both hypothesis. Given the Scenario × Presentation Order interaction and the scenario effect at the alternative hypothesis instruction condition, a further analysis was performed to investigate the possible effect of presentation order on the response accuracy of the alternative hypothesis instruction condition.

ANOVA of mean response accuracy for alternative hypothesis instruction condition. A repeated-measures ANOVA with one within-subjects factor (scenario) and one between-subjects factor (presentation order) was conducted on hypothesis adjusted mean response accuracy for participants in the alternative hypothesis instruction condition. The analysis was performed to determine whether the scenario effect on mean response accuracy for the alternative hypothesis instruction condition was actually due to a Scenario × Presentation Order interaction. Table 4.6 lists the results of the ANOVA.

The interaction between scenario and presentation order was significant, F (1, 27) = 8.996, p = .006, η^2_G = .095. Figure 4.4 graphically depicts the Scenario × Presentation Order interaction for the mean response accuracy of participants in the alternative hypothesis instruction condition. Participants in the alternative hypothesis instruction condition and the SP group (Scorpion-Spider, MDT Microworld) had a significantly lower mean response accuracy on their first scenario (Scorpion-Spider; M = 72.5%, SD = .24) than on their second scenario (MDT Microworld; M = 95.8%, SD = .13) and the SP group's response accuracy on the Scorpion-Spider scenario (M = 72.5%, SD = .24) was significantly lower than participants in the PS group (MDT Microworld, Scorpion-Spider; *M* = 92.8%, *SD* = .16) who completed the scenario following the MDT Microworld. There was no difference in mean response accuracy between groups for the MDT Microworld or in accuracy for the PS group from their first to their second scenario.

Table 4.6

Repeated-Measures Analysis of Variance of Hypothesis Adjusted Mean Response Accuracy on Basic Event Prediction, Alternative Hypothesis Instruction Condition Only and Excluding Inaccurate Subjects (n = 29)

Source	*df*	*F*	η^2_P	η^2_G	*p*
			Between Subjects		
Order (O)	1	2.285	.078	.055	.142
S within-group error	27	(.048)			
			Within Subjects		
Scenario (S)	1	8.996	.250	.095	.006*
S × O	1	8.996	.250	.095	.006*
S × *S* within-group error	27	(.022)			

Note. Values in parentheses indicate mean square errors. *S* = subjects.
$^*p < .05$

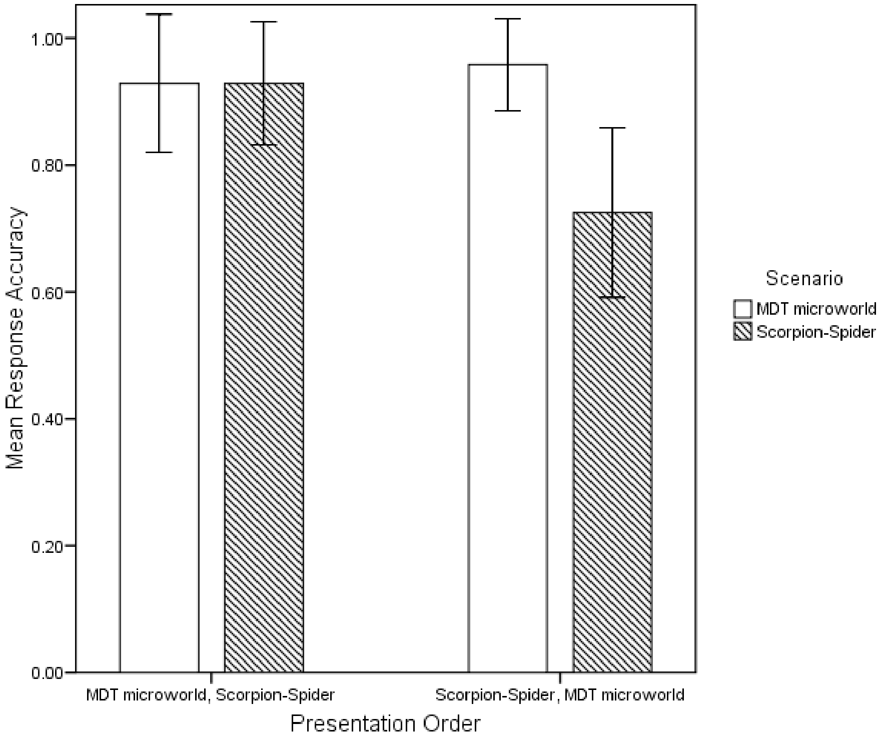


Figure 4.4. Hypothesis adjusted mean response accuracy for basic event prediction questions for the alternative hypothesis instruction condition for Presentation Order × Scenario interaction, excluding inaccurate participants.

Summary of Basic Event Prediction Results

Overall, participants had a relatively high mean response accuracy on the basic event prediction questions for both the MDT Microworld (*M* = 93.4%, *SD* = .16) and the Scorpion-Spider scenario (*M* = 91.4%, *SD* = .16). The interaction between Scenario × Presentation Order on the basic event prediction questions indicated a possible learning effect as the SP group's accuracy improved from their first scenario to the second scenario. Poor performance was isolated to the Scorpion-Spider scenario when it was presented as the first scenario. The Scenario × Instruction Condition interaction indicated that participants in the alternative hypothesis instruction condition had particular issues with the Scorpion-Spider scenario. A further analysis of the alternative hypothesis instruction condition indicated that the low mean response accuracy on the Scorpion-Spider scenario was isolated to the participants in the SP presentation order group. Participants in the alternative hypothesis instruction condition that were presented with the Scorpion-Spider scenario as their second scenario were highly accurate (*M* = 95.8%). The exact nature of the issue affecting the participants in the SP group, especially participants in the SP group *and* the alternative hypothesis instruction condition, remains unclear.

Positivity Assessment

The positivity assessment questions are similar to the basic event prediction questions but, instead of providing a “yes” or “no” response, participants are asked to estimate the likelihood that the phenomenon of interest will occur for a single figure. Participants estimated likelihood using a visual analog scale (VAS) with anchors at 0 and 100. The positivity assessment questions provided a quantitative probe of participants’ assessment of the likelihood that a single test will result in a positive event. Overall, I expected participant positivity assessments to tend toward 0% or 100% with two exceptions: First, the alternative hypothesis instruction condition had two hypotheses to consider and a single figure can be positive for one hypothesis and negative for another hypothesis. Participants may simply rank these figures near or at 100% positivity. However, I expected participants to weigh the two hypotheses and shift away from 0%

and 100%, values that require confidence in the outcome, to assessments closer to 50%. Second, the common event instruction condition informed participants that positive events (stopped particles and eaten spiders) occur regularly. I expected participants in the common event instruction condition to have higher positivity assessments for negative tests than the other instruction conditions.

Scoring deviation from expected positivity values. Participant responses on the VAS were measured and recorded. The VAS was 126.7 mm long. VAS measurements were converted to a positivity score by dividing the recorded measurement by 126.7. The positivity measurements were used to calculate deviation from an expected value by taking the absolute value of the participant's positivity score minus the expected value for the question. Table 4.7 lists the positivity assessment questions and the positivity for the figures given the primary hypothesis (H1) and the alternative hypothesis (HA). Missing items were not scored.

Most of the instruction conditions were given only the primary hypothesis (H1) and the expected values were based on the positivity of the figure given H1. If the figure would lead to a positive event given H1, then the expected value was 100%. If the figure would lead to a negative event given H1, then the expected value was 0%. The alternative hypothesis instruction condition required a different set of expected values that included an adjustment that accounted for the alternative hypothesis. If a screen was positive for either or both of the hypotheses (H1 and HA), the baseline expected value for a participant in the alternative hypothesis condition was 100%. If a screen was negative on both of the hypotheses, the expected value for a participant in the alternative hypothesis condition was 0%.

The positivity assessment questions allow participants to provide quantitative responses. Some of the figures presented in the positivity assessments are positive on the primary hypothesis but negative on the alternative hypothesis or vice versa. In the basic event prediction questions when the two hypotheses are in conflict, participants were forced to simply respond that the event was likely to happen. In the positivity assessment questions, participants may respond closer to chance, 50%, to reflect the uncertainty generated by the conflict between the two hypotheses. If this is the case, I expect to continue to find a main effect of instruction condition because the mean deviation for the figures in conflict will approach .50 raising the overall mean deviation for the alternative hypothesis instruction condition compared to the other instruction conditions. See Appendix F for an analysis of variance of the unadjusted mean deviation from the expected values for positivity assessment.

Dropping Inaccurate Participants

As previously described, 24 participants (13%) were inaccurate on the catch trials. The inaccurate participants were dropped from the analyses of positivity assessment deviation. See Appendix F for analysis of positivity assessment including inaccurate participants.

Table 4.7

Positivity Assessment Questions and Positivity for the Given Hypotheses

Question	MDT Microworld				Scorpion-Spider			
	Figure	H1	HA	H1HA	Figure	H1	HA	H1HA
9	BT	P	P	PP	DR	P	P	PP
10	WSBD	N	P	NP	UR	P	N	PN
11	BD	N	P	NP	URUE	P	N	PN
12	WS	N	N	NN	DRUR	P	P	PP
13	BTBT	P	P	PP	UE	N	N	NN
14	BSBS	N	P	NP	DS	N	P	NP
15	WDWD	N	N	NN	DSUE	N	P	NP
16	WT	P	N	PN	DEDE	P	N	PN
17	BSBD	N	P	NP	UEUS	N	N	NN
18	WSWT	P	N	PN	DEDS	N	P	NP
19	WSBT	P	P	PP	USUS	N	N	NN
20	BSWT	P	P	PP	USDE	N	P	NP
21	BS	N	P	NP	DRDS	P	P	PP
22	WSWD	N	N	NN	DSUR	P	P	PP
23	BSWD	N	P	NP	DRUS	P	P	PP
24	BDBT	P	P	PP	US	N	N	NN
25	WTBT	P	P	PP	URUR	P	N	PN
26	WDBD	N	P	NP	UEDE	N	P	NP

Note. Figure columns describe contents of figures by feature. For MDT Microworld: up to two objects are defined by color and shape where W – white, B – black, S – square, D – disc, and T – triangle. For Scorpion-Spider: up to two objects are defined by tail direction and claw shape where U – Up-turned, D – down-turned, S – spiky, E – serrated, R – round. N indicates a negative test, P indicates a positive test.

ANOVA for mean deviation from expected positivity. A repeated-measures analysis of variance with one within-subjects factor (scenario) and two between-subjects factors (presentation order and instruction condition) was performed on the hypothesis adjusted mean deviation of positivity assessment from expected values. Table 4.8 lists the results of the ANOVA.

Table 4.8

Repeated-Measures Analysis of Variance of Hypothesis Adjusted Mean Deviation of Positivity Assessment from Expected Values, Excluding Inaccurate Participants (n = 159)

Source	*df*	*F*	η^2_P	η^2_G	*p*
			Between Subjects		
Order (O)	1	.182	.001	.001	.670
Instruction (I)	5	8.302	.220	.144	.000*
O × I	5	.729	.024	.015	.602
S within-group error	147	(.023)			
			Within Subjects		
Scenario (S)	1	.915	.006	.002	.340
S × O	1	16.395	.100	.043	.000*
S × I	5	2.025	.064	.027	.078
S × O × I	5	.723	.024	.010	.607
S × *S* within-group error	147	(.016)			

Note. Values in parentheses indicate mean square errors. *S* = subjects.
$^*p < .05$

Despite adjusting for the additional hypothesis given to the alternative hypothesis condition, the main effect of instruction condition was significant, *F* (5,147) = 8.203, $p < .001$, $\eta^2_G = .144$. The main effect of presentation order and the interaction between presentation order and instruction condition were not significant. A Tukey's HSD post-hoc analysis of the main effect of instruction condition revealed that the mean deviation from the expected values for the alternative hypothesis instruction condition (*M* = .31, *SE* = .02) was significantly greater ($p < .004$ for all comparisons) than the mean deviation of the other instruction conditions.

The main effect of scenario was not significant. There was an interaction effect between scenario and presentation order, F (1, 147) = 16.395, p < .001, η^2_G = .043. The interaction between scenario and instruction condition was not significant. The three-way interaction between scenario, instruction condition, and presentation order was also not significant. Figure 4.5 graphically depicts the Scenario × Presentation Order interaction.

Multiple pairwise comparisons were performed using the Sidak adjustment to assess the simple effects of the Scenario × Presentation Order interaction. The hypothesis adjusted mean deviation of participants' positivity assessments from the expected values was lowest for the second presented scenario regardless of the content of the scenario. For the PS presentation order (MDT Microworld, Scorpion-Spider), the positivity assessments became significantly closer (p = .031) to the expected values from the first scenario (MDT Microworld; M = .225, SD = .19) to the second scenario (Scorpion-Spider; M = .181, SD = .12). For the SP presentation order (Scorpion-Spider, MDT Microworld), the positivity assessments likewise became closer (p = .001) to the expected values from the first scenario (Scorpion-Spider; M = .233, SD = .15) to the second scenario (MDT Microworld; M = .161, SD = .11). The results suggest that participants tended to select positivity values closer to 0% and 100% in their second scenario.

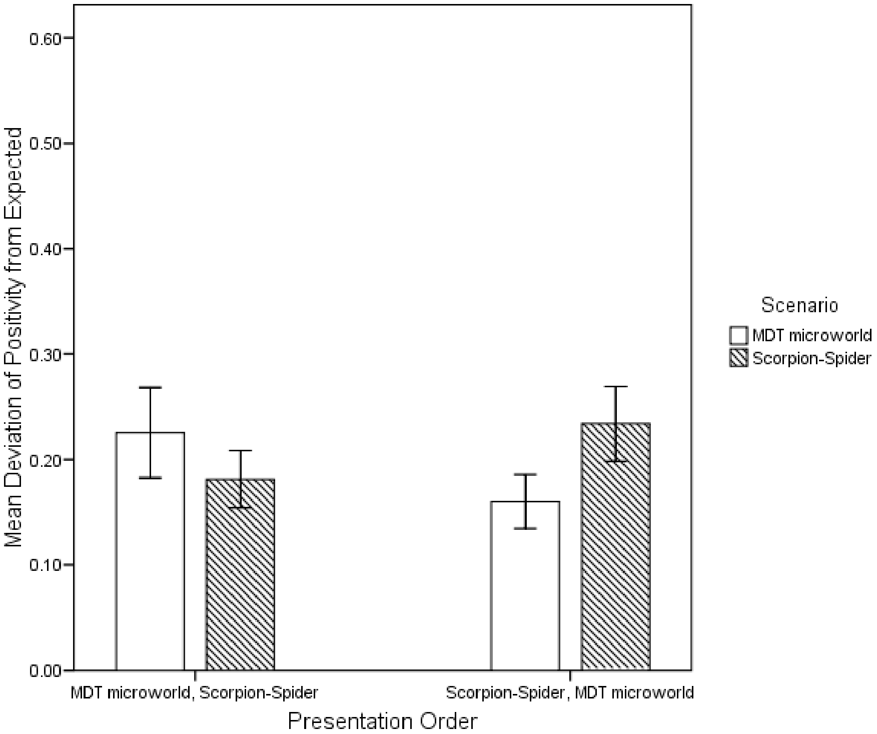


Figure 4.5. Hypothesis adjusted mean deviation of positivity assessment from expected values for scenario by presentation order, excluding inaccurate participants. Error bars indicate 95% confidence interval.

Modified Scoring to Investigate Effect of Question Type

Adjusting the expected positivity assessment values for the additional hypothesis did not fully account for the differences between the alternative hypothesis instruction condition and the other instruction conditions. Because the alternative hypothesis condition participants were judging two hypotheses that predicted conflicting outcomes for some figures, the participants in the alternative hypothesis instruction condition were judging the figures differently from the other instruction conditions. In order to investigate the impact of instruction condition on assessment of different questions, the positivity assessment questions were sorted into four types: negative for both hypotheses (NN), negative for H1 but positive for HA (NP), positive for H1 but negative for HA (PN), and positive for both hypotheses (PP). The mean deviation from the expected values for the four question types was calculated for each participant and a repeated-measures analysis of variance was performed on the resulting scores.

The new scores resulted in a 2 × 4 × 2 × 6 factorial mixed design with two within-subjects variables: scenario and question type. Each level of scenario (MDT Microworld and Scorpion-Spider) had four question types (NN, NP, PN, and PP). The two between-subjects variables were the same as previous analyses: presentation order and instruction condition.

ANOVA of mean deviation by question type. A repeated-measures analysis of variance with two within-subjects factors (scenario and question type) and two between-subjects factors (presentation order and instruction condition) was performed to assess the impact of instruction condition on mean deviation from expected values by question type. Table 4.9 lists the results of the repeated-measures ANOVA. The assumption of sphericity was not met for the repeated-measures ANOVA and the analysis of the effect of question type required application of the Greenhouse-Geisser correction.

The interesting results from this analysis were those related to the question type. There was a main effect for question type, F (2.034, 298.946) = 24.967, $p < .001$, η^2_G = .041, and for the interaction between question type and instruction condition, F (10.168, 298.946) = 6.786, $p < .001$, η^2_G = .055. The interaction between question type and presentation order and the three-way interaction between question type, instruction condition, and presentation order were not significant.

Table 4.9

Repeated-Measures Analysis of Variance of Hypothesis Adjusted Mean Deviation of Positivity Assessment from Expected Values for Four Question Types, Excluding Inaccurate Participants (n = 159)

Source	*df*	*F*	η^2_P	η^2_G	*p*
Between Subjects					
Order (O)	1	7.552	.000	.000	.789
Instruction (I)	5	534.051	.204	.079	.000*
O × I	5	.072	.020	.007	.691
S within-group error	147	(.092)			
Within Subjects					
Scenario (S)	1	.419	.003	.001	.518
S × O	1	15.293	.094	.023	.000*
S × I	5	2.058	.065	.015	.074
S × O × I	5	.621	.021	.005	.684
S × *S* within-group error	147	(.061)			
Question Type (Q)	2.034[a]	24.967	.145	.041	.000*
Q × O	2.034[a]	.784	.005	.001	.460
Q × I	10.168[a]	6.786	.188	.055	.000*
Q × O × I	10.168[a]	1.749	.056	.015	.068
Q × *S* within-group error	298.946[a]	(.034)			
S × Q	2.137[a]	4.986	.033	.006	.006*
S × Q × O	2.137[a]	.516	.003	.001	.609
S × Q × I	10.683[a]	2.566	.080	.016	.004*
S × Q × O × I	10.683[a]	.790	.026	.005	.646
S × Q × *S* within-group error	314.082[a]	(.024)			

Note. Values in parentheses indicate mean square errors. *S* = subjects.
a = The assumption of sphericity was not met. Greenhouse-Geisser correction applied to p value.
**p* < .05

The interaction between scenario and question type was significant, *F* (2.137, 314.082) = 4.986, *p* < .006, η^2_G = .006. The three-way interaction between scenario, question type, and instruction condition was also significant, *F* (10.683, 314.082) = 2.566, *p* < .004, η^2_G = .016. The three-way interaction between scenario, question type, and presentation order and the four-way interaction between scenario, question type, presentation order, and instruction condition were not significant.

The three-way interaction between scenario, question type, and instruction condition incorporates the other significant new effects (question type, Question Type × Instruction Condition, and Scenario × Question Type). The three-way interaction is presented in Figures 4.6 and 4.7. In Figure 4.6, the mean deviation of positivity assessment from the expected values on the MDT Microworld questions for each of the four question types is shown in clusters. Each bar within the four clusters represents one of the instruction conditions. Figure 4.7 depicts the same information for the Scorpion-Spider scenario.

Examining Figures 4.6 and 4.7, it appears that, as expected, the majority of the question type by instruction condition portion of the Scenario × Question Type × Instruction Condition interaction was the result of the increase in mean deviation from the expected positivity values for the alternative hypothesis instruction condition on the NP (MDT; *M* = .35, *SD* = .19; Scorpion-Spider; *M* = .53, *SD* = .18) and PN (MDT; *M* = .33, *SD* = .25; Scorpion-Spider; *M* = .44, *SD* = .17) question types compared to the NN (MDT; *M* = .16, *SD* = .21; Scorpion-Spider; *M* = .15, *SD* = .16) and PP (MDT; *M* = .18, *SD* = .23; Scorpion-Spider; *M* = .27, *SD* = .15) question types in both the MDT Microworld and the Scorpion-Spider scenario.

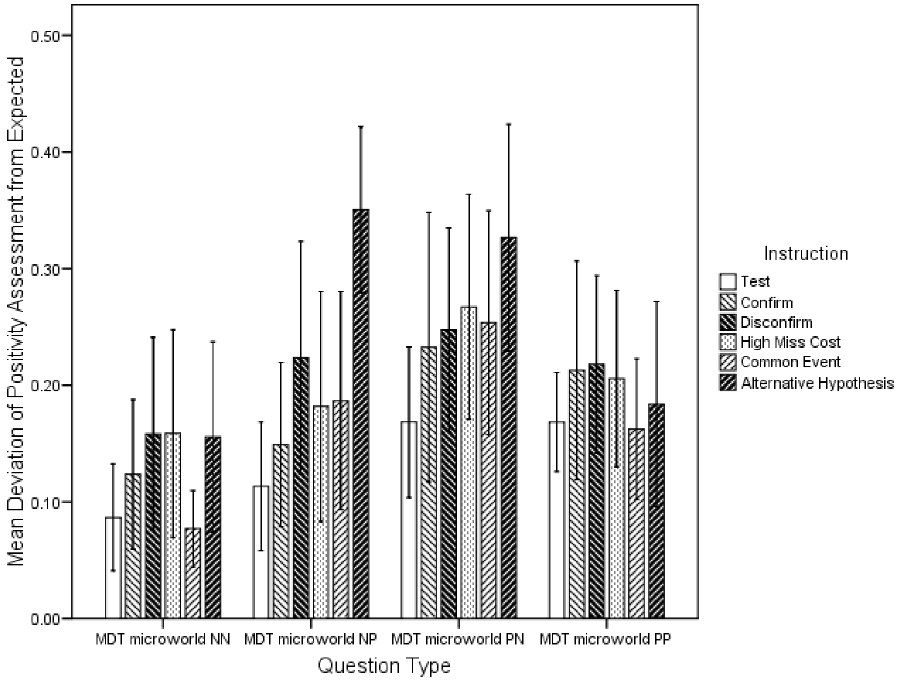


Figure 4.6. Hypothesis adjusted mean deviation of positivity assessment from expected values for instruction condition by question type for MDT Microworld scenario. Error bars indicate 95% confidence interval.

Deviation on the NP, PN and PP question types appear to be larger in the Scorpion-Spider scenario leading to the scenario difference in the interaction. Multiple pairwise comparisons indicate that an increased deviation on the PP questions in the Scorpion-Spider scenario (*M* = .23, *SD* = .16) compared to the MDT Microworld (*M* = .19, *SD* = .18) was the only significant difference (*p* = .027) between the scenarios by question type.

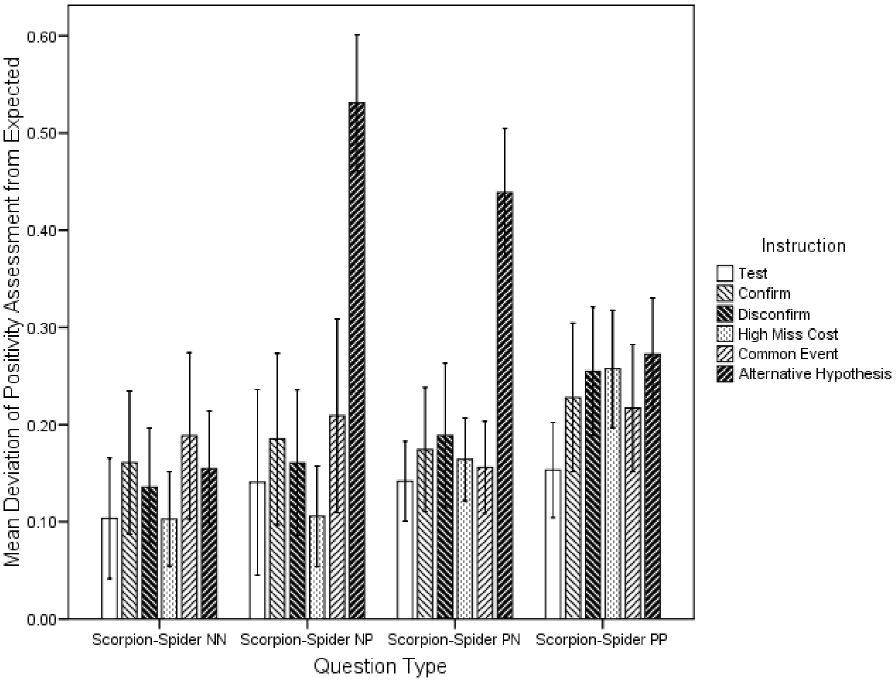


Figure 4.7. Hypothesis adjusted mean deviation of positivity assessment from expected values for instruction condition by question type for Scorpion-Spider scenario. Error bars indicate 95% confidence interval.

Interestingly, the common event instruction condition had a low mean deviation for questions that included figures negative on both hypotheses compared to the other question types. I expected the opposite to be the case given the intended impact of the common event instruction. By informing participants that the positive events were common, I expected participants to demonstrate an inflated assessment of positivity, especially for negative tests. The instructions obviously did not have the desired effect on the assessment of the figures.

The three-way interaction accounts for the main effect of question type indicated in the repeated-measures ANOVA. It also accounts for the interaction between question type and instruction condition: the alternative hypothesis instruction condition has a different pattern of results compared to the other instruction conditions because of the NP and PN questions for which the two hypotheses are in conflict. For the scenario by question type interaction, the positivity assessment for the PP question types is closer to the expected values in the MDT Microworld (*M* = .1914, *SD* = .1852) than in the Scorpion-Spider scenario (*M* = .2139, *SD* = .1594).

Summary of Positivity Assessment

The positivity assessment questions were designed to provide a quantitative perspective on how participants assess individual tests and how the thematic content and instruction conditions impact this assessment.

Throughout the analyses of positivity assessment, there was a main effect of presentation order. Participants in both presentation orders would show a reduced mean deviation from the expected values on the second scenario. Lower deviation from expected values indicates responses closer to 0% or 100% which may indicate an increased confidence in assessment from the first scenario to the second scenario.

The alternative hypothesis instructions significantly affected participant positivity assessments. In order to fully explore the impact of the alternative hypothesis, the questions were split into four question types: the NN group, the NP group, the PN group and the PP group. Participants in the alternative hypothesis instruction condition responded much like other participants to the NN and PP groups. However, the primary and alternative hypotheses provided conflicting predictions for the NP and PN groups of questions and participant responses were accordingly closer to chance (50%) than the 100% used for the baseline expected values when scoring.

The deviation from expected values of the common event instruction condition did not reflect the expected impact of the common event instructions. Rather than reporting increased positivity assessments for negative tests, the participants in the common event instruction condition had one of the lowest mean deviations indicating that their assessments were the closest to the 0% used for the baseline expected values when scoring.

Explicit Positive Selection

The explicit positive selection questions present participants with two figures and require the participant to select which of the two figures are most likely to result in the phenomenon of interest (stopped particle or eaten spider). Participants are expected to assess the likelihood of a positive event occurring based on the given hypothesis or hypotheses for each figure and select the figure they consider the most likely.

Scoring deviation from expected figure selection. Figure selection for each question was predicted based on the given hypothesis or hypotheses. For the instruction conditions given only one hypothesis (H1), there were three possible cases: the figures were both –H tests, one figure was a –H test and another was a +H test, or the figures were both +H tests. If the two figures were both negative or both positive, our prediction for figure selection was chance (.5). If one figure was negative and the other was positive, our prediction was that the participant would select the positive test of the hypothesis.

For the alternative hypothesis instruction condition, participant selection was predicted by determining which figure was most positive given the two hypotheses (H1 and HA). If both figures in the question were matched across the two hypotheses (i.e., positive for both hypotheses, negative for both hypotheses, or negative for one hypothesis and positive for the other hypothesis), then positivity cannot predict which test will be selected and the predicted score was set at chance, .5. Otherwise, the figure that was positive for the most hypotheses should be selected and the predicted score was set appropriately. For example, if the top figure was positive for H1 and HA and the bottom figure was positive on only H1, then the top figure was the expected selection.

Participant test selections were converted to a deviation from expected figure selection by subtracting the actual figure selection from the expected figure selection given H1. A negative deviation indicates that more participants selected the first figure (top figure on the questionnaires) than expected. A positive deviation indicated that more participants selected the second figure (bottom figure on the questionnaires) than expected. A deviation close to zero indicates that participants' selections matched the predicted figure selections. Table 4.10 lists the explicit positive selection questions.

Table 4.10

Explicit Positive Test Selection Questions and Positivity of Figures for the Given Hypotheses

Question	Figures	H1	HA	H1HA
		MDT Microworld		
31	WS / WDWD	N / N	N / N	NN / NN
32	BSBS / WSWD	N / N	P / N	NP / NN
33	BT / BS	P / N	P / P	PP / NP
34	WSWT / BD	P / N	N / P	PN / NP
35	BT / WT	P / P	P / N	PP / PN
36	WSBD / WSWD	N / N	P / N	NP / NN
37	WDBD / WDBT	N / P	P / P	NP / PP
38	WSBD / WSBT	N / P	P / P	NP / PP
39	BDBD / BDBT	N / P	P / P	NP / PP
40	BT / BD	P / N	P / P	PP / NP
41	BSWT / BSWD	P / N	P / P	PP / NP
42	WSWD / WSWS	N / N	N / N	NN / NN
43	WS / WD	N / N	N / N	NN / NN
		Spider- Scorpion		
31	UR / DR	P / P	N / P	PN / PP
32	UR / UE	P / N	N / N	PN / NN
33	UEDE / URDE	N / P	P / P	NP / PP
34	DS / DE	N / N	P / P	NP / NP
35	DSUE / DSUR	N / P	P / P	NP / PP
36	UR / US	P / N	N / N	PN / NN
37	DEDS / DSDS	N / N	P / P	NP / NP
38	USUS / DEDS	N / N	N / P	NN / NP
39	UEUE / URUE	N / P	N / N	NN / PN
40	DRUS / USDE	P / N	P / P	PP / NP
41	DS / DEDE	N / N	P / P	NP / NP
42	DRDS / UE	P / N	P / N	PP / NN
43	DSUE / DEDS	N / N	P / P	NP / NP

Note. Figure columns describe contents of figures by feature. For MDT Microworld: up to two objects are defined by color and shape where W – white, B – black, S – square, D – disc, and T – triangle. For Scorpion-Spider: up to two objects are defined by tail direction and claw shape where U – Up-turned, D – down-turned, S – spiky, E – serrated, R – round. N indicates a negative test, P indicates a positive test.

Dropping Inaccurate Participants

As described in the analysis of catch trial questions, twenty-four participants (13%) were inaccurate on the catch trials, especially in the Scorpion-Spider scenario. The inaccurate participants were dropped from analyses of the explicit positive test selection questions. See Appendix F for the analysis of explicit positive test selection including the inaccurate subjects.

ANOVA of mean deviation from expected selections. A repeated-measures analysis of variance with one within-subjects factors (scenario) and two between-subjects factors (instruction condition and presentation order) was performed on mean deviation from expected figure selection values. The participants that were inaccurate on the catch trials were dropped from this analysis. Table 4.11 lists the results of the ANOVA.

The main effect of presentation order was significant, *F* (1, 147) = 14.346, $p <$ .001, $\eta^2_G = .043$. The main effect for instruction condition was not significant, and the interaction between presentation order and instruction condition was also not significant. The main effect of presentation order reveals that the participants in the PS presentation order (MDT Microworld, Scorpion-Spider) have a very slight tendency to select more top figures (*M* = -.025, *SE* = .006) than expected. The participants in the SP presentation order (Scorpion-Spider, MDT Microworld) have no preference for top or bottom figures (*M* = .008, *SE* = .006).

Table 4.11

Repeated-Measures Analysis of Variance of Hypothesis Adjusted Mean Deviation from Expected Values for Explicit Positive Test Selection Questions, Excluding Inaccurate Participants (n = 159)

Source	*df*	*F*	η^2_P	η^2_G	*p*
			Between Subjects		
Order (O)	1	14.346	.089	.043	.000*
Instruction (I)	5	1.89	.060	.028	.099
O × I	5	1.842	.059	.028	.108
S within-group error	147	(.006)			
			Within Subjects		
Scenario (S)	1	113.787	.436	.295	.000*
S × O	1	.619	.004	.002	.433
S × I	5	1.929	.061	.034	.093
S × O × I	5	.712	.023	.013	.616
S × *S* within-group error	147	(.007)			

Note. Values in parentheses indicate mean square errors. *S* = subjects.
$^*p < .05$

The main effect for the scenario was highly significant, F (1, 147) = 113.787, $p < .001$, $\eta^2_G = .295$. The interaction between scenario and presentation order and the interaction between scenario and instruction condition were not significant. The three-way interaction between scenario, presentation order, and instruction condition was also not significant.

For the MDT Microworld, the mean deviation is negative (*M* = -.058, *SD* = .088) reflecting a tendency to select more top figures than predicted. For the Scorpion-Spider microworld, the mean deviation is positive (*M* = .043, *SD* = .081) reflecting a tendency to select more bottom figures than predicted. The predictions are based on positivity of the figures based purely on the primary hypothesis or the alternative hypothesis. There appears to be another scenario-specific factor related to figure placement leading participants to demonstrate very slight tendencies towards top or bottom figures.

Participant Positivity Assessments

In the positivity assessment questions, participants rated the likelihood of the phenomenon of interest occurring given the hypotheses they were provided. Many of the figures used in the explicit positive test selection questions were assessed in the positivity assessment questions. Given the small differences between the scenarios, I predicted test selection based on each participant's individual positivity assessments to determine if the effects of scenario and presentation order were accounted for.

The new score was calculated by taking each participant's positivity assessment for the two figures in each question and predicting that the participant would select the figure that received the highest positivity assessment by the participant.

ANOVA of mean deviation from values predicted by individual positivity assessment. A repeated-measures ANOVA with one within-subjects factor (scenario) and two between-subjects factors (presentation order and instruction condition) was performed to determine whether the participants' individual positivity assessments accounted for the difference between scenarios in the explicit positive test selection questions. The results of the ANOVA are listed in Table 4.12. The participants that were judged inaccurate on the catch trials were dropped from the analysis.

Table 4.12

Repeated-Measures Analysis of Variance of Mean Deviation from Predicted Values for Explicit Positive Test Selection Questions Given Individual Positivity Assessments, Excluding Inaccurate Participants (n = 159)

Source	*df*	*F*	η^2_P	η^2_G	*p*
			Between subjects		
Order (O)	1	8.4	.054	.026	.004*
Instruction (I)	5	1.281	.042	.020	.275
O × I	5	2.0	.064	.031	.082
S within-group error	147	(.008)			
			Within Subjects		
Scenario (S)	1	93.526	.389	.253	.000*
S × O	1	2.36	.016	.008	.127
S × I	5	1.472	.048	.026	.203
S × O × I	5	1.065	.035	.019	.382
S × *S* within-group error	147	(.009)			

Note. Values in parentheses indicate mean square errors. *S* = subjects.
**p* < .05

For the analysis of deviation from predicted values given individual positivity assessments, there was a main effect of presentation order, *F* (1, 147) = 8.4, *p* = .004, and a main effect of scenario, *F* (1, 147) = 93.526, *p* < .001. The presentation order and scenario results mirrored those of the previous analyses. Using the individual positivity assessments did not provide additional insight into the small differences in test selection between the MDT Microworld and the Scorpion-Spider scenario.

Summary of Explicit Positive Test Selection

In the explicit positive test selection questions, participants were directed to select from two figures the one figure that was most likely to result in a positive event (stopped particles or eaten spider). Overall, the mean deviation from the predicted values was low (less than .07 for the two scenarios) suggesting that participants were generally selecting the figure most likely to result in a positive event according to the hypotheses.

Regardless of predictor (by hypothesis or by individual positivity assessments) the main effect of presentation order and scenario persisted. These effects are difficult to account for. The presentation order effect reveals that participants that were first exposed to the MDT Microworld scenario had a small tendency to select more top figures than predicted given the hypotheses or by individual positivity assessments. The test selections participants first exposed to the Scorpion-Spider scenario indicated no preference for top or bottom figures.

The main effect for scenario also suggests an effect of test placement. When completing the questions for the MDT Microworld, participants overall tended to select more top figures than predicted. When completing the Scorpion-Spider questions, participants overall tended to select more bottom figures than predicted. The tendency doesn't appear to be related to assessment of individual screens given that the same pattern of results appeared when using individual assessments. Identifying the factor involved may reveal more about how participants choose between two screens and how this task differs from the assessment of a single screen.

Hypothesis Test Selection

The hypothesis test selection questions were the core questions of the experimental design. The questions were similar to the explicit positive test selection questions. However, participants weren't asked to judge which figure is most likely to lead to a positive event; participants were asked to judge which figure would most effectively test the hypothesis or hypotheses they were given. I expected positive test selection (selection of a positive test when selecting between a +H and a –H tests) to be reduced in the Scorpion-Spider scenario, the high miss cost instruction condition, the common event instruction condition, and the alternative hypothesis instruction condition.

Scoring hypothesis test selection. Table 4.13 and 4.14 lists the hypothesis test selection questions, the figures included in each question and the positivity of the figures for the given hypotheses. For these questions, participants were scored by how many positive tests of the primary hypothesis (H1) were selected when the question included a positive and a negative test of H1.

Dropping Inaccurate Participants

As in the previous analyses, the inaccurate participants were dropped from analysis of the hypothesis test selection questions. See Appendix F for the analysis of the hypothesis test selection questions including the inaccurate participants.

Overall mean proportion. For the MDT Microworld scenario, the mean percentage of positive tests selected was 86.6% (*SD* = .20). For the Scorpion-Spider scenario, the mean proportion of positive tests selected was 85.8% (*SD* = .23).

Table 4.13

Hypothesis Test Selection Questions for MDT Microworld and Positivity of Figures for the Given Hypotheses

Question	Figures	H1	HA	H1HA
		MDT Microworld		
44	WSWD / WTBT	N / P*	N / P	NN / PP*
45	WSBD / WSWD	N / N	P / N	NP / NN*
46	BSBD / WT	N / P*	P / N	NP / PN
47	WSWD / WSWS	N / N	N / N	NN / NN
48	WDWT / WSWT	P / P	N / N	PN / PN
49	WDWT / BSBD	P / N*	N / P	PN / NP
50	WDBD / WDBT	N / P*	P / P	NP / PP*
51	WS / WD	N / N	N / N	NN / NN
52	WT / BSWD	P / N*	N / P	PN / NP
53	BSWT / BSWD	P / N*	P / P	PP / NP*
54	BT / WT	P / P	P / N	PP / PN*
55	WS / BS	N / N	N / P	NN / NP*
56	BT / BS	P / N*	P / P	PP / NP*
57	WS / WDWD	N / N	N / N	NN / NN
58	BS / WSWD	N / N	P / N	NP / NN*
59	BDBD / BDBT	N / P*	P / P	NP / PP*
60	WT / BSWD	P / N*	N / P	PN / NP
61	WSWD / WDWT	N / P*	N / N	NN / PN*
62	WSBD / WSBT	N / P*	P / P	NP / PP*
63	BT / BD	P / N*	P / P	PP / NP*

Note. Figure columns describe contents of figures by feature. For MDT Microworld: up to two objects are defined by color and shape where W – white, B – black, S – square, D – disc, and T – triangle. N indicates a negative test, P indicates a positive test.

*Indicates question with a negative and a positive test.

Table 4.14

Hypothesis Test Selection Questions for Scorpion-Spider and Positivity of Figures for the Given Hypotheses

		Spider- Scorpion		
44	DSUE / DSUR	N / P*	P / P	NP / PP*
45	UR / US	P / N*	N / N	PN / NN*
46	DEDS / DSDS	N / N	P / P	NP / NP
47	UEUE / URUE	N / P*	N / N	NN / PN*
48	UR / DR	P / P	N / P	PN / PP*
49	DS / DE	N / N	P / P	NP / NP
50	DRUS / USDE	P / N*	P / P	PP / NP*
51	DS / DEDE	N / N	P / P	NP / NP
52	UR / UE	P / N*	N / N	PN / NN*
53	UEDE / URDE	N / P*	P / P	NP / PP*
54	DRDE / UEUS	P / N*	P / N	PP / NN*
55	DS / US	N / N	P / N	NP / NN*
56	DR / USDE	P / N*	P / P	PP / NP*
57	DRDE / DRDS	P / P	P / P	PP / PP
58	US / DEDS	N / N	N / P	NN / NP*
59	DEDS / DRUR	N / P*	P / P	NP / PP*
60	DEDS / DRDE	N / P*	P / P	NP / PP*
61	DSUE / DEDS	N / N	P / P	NP / NP
62	DR / USDE	P / N*	P / P	PP / NP*
63	UEUS / DR	N / P*	N / P	NN / PP*

Note. Figure columns describe contents of figures by feature. For Scorpion-Spider: up to two objects are defined by tail direction and claw shape where U – Up-turned, D – down-turned, S – spiky, E – serrated, R – round. N indicates a negative test, P indicates a positive test.

* Indicates question with a negative and a positive test.

ANOVA of positive test selection for H1. A repeated-measures ANOVA with one within-subjects factor (scenario) and two between-subjects factors (presentation order and instruction condition) was performed excluding the inaccurate participants from the analysis. Table 4.15 lists the results of the ANOVA.

The main effect for presentation order was not significant. The main effect for instruction condition was significant, *F* (5, 147) = 3.267, *p* = .008, η^2_G = .078. The main effect of scenario was not significant. There were no interactions between the factors. Figure 4.9 depicts the mean proportion of positive tests selected by instruction condition.

Table 4.15

Repeated-Measures Analysis of Proportion of Positive Tests Selected for Hypothesis Test Selection Questions, Excluding Inaccurate Participants (n = 159)

Source	*df*	*F*	η^2_P	η^2_G	*p*
			Between Subjects		
Order (O)	1	.116	.001	.001	.734
Instruction (I)	5	3.267	.100	.078	.008*
O × I	5	.871	.029	.022	.502
S within-group error	147	(.068)			
			Within Subjects		
Scenario (S)	1	.351	.003	.001	.554
S × O	1	.325	.002	.001	.570
S × I	5	.846	.028	.007	.519
S × O × I	5	.830	.027	.007	.530
S × *S* within-group error	147	(.022)			

Note. Values in parentheses indicate mean square errors. *S* = subjects.
**p* < .05

A Tukey's HSD post-hoc analysis was used to investigate the main effect of instruction condition. The analysis revealed that the mean proportion of positive tests selected was significantly lower in the alternative hypothesis condition (*M* = 75%) than in the confirm instruction condition (*M* = 92.3%, *p* = .011), the common event condition (*M* = 90.3%, *p* = .025), and the high miss cost condition (*M* = 89.1%, *p* = .048).

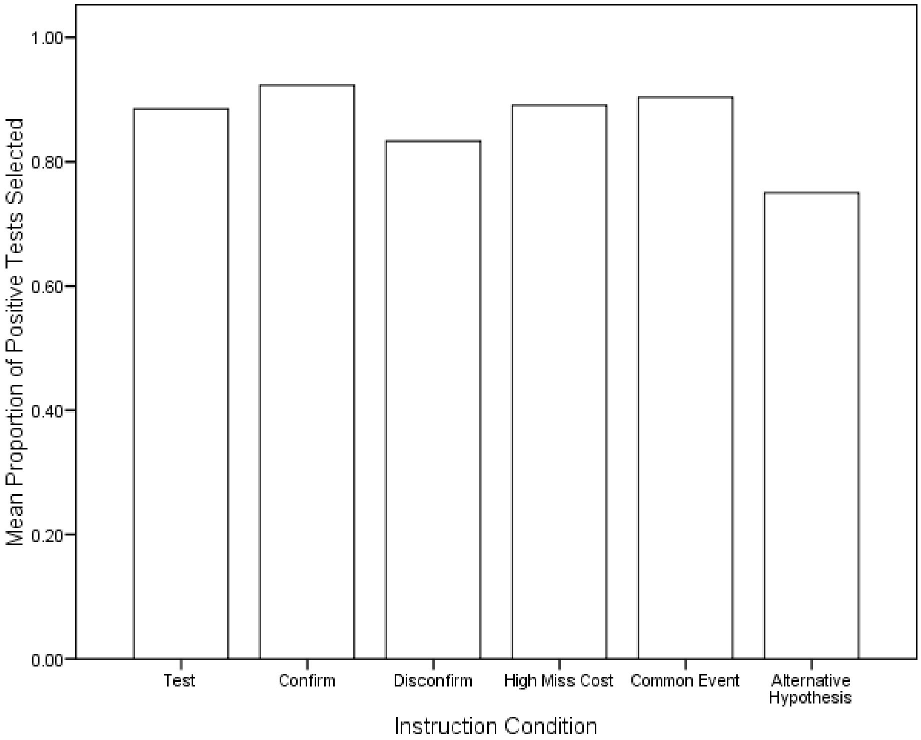


Figure 4.8. Mean proportion of positive tests selected by instruction condition, excluding inaccurate participants.

Adjusting Scoring for Alternative Hypothesis

The previous analysis indicated that providing an alternative hypothesis (HA) did reduce positive test selections with respect to H1. However, it is possible that all of the –H tests that were selected were also positive tests of HA. The scoring for proportion of positive tests selected was adjusted to account for the additional hypothesis provided to the participants in the alternative hypothesis condition. A selection is counted as positive when participants select the *more positive* of the two tests. For example, a participant may be presented with one figure that is positive on H1 but negative on HA and another figure that is positive on both. If they select the figure that is positive on only one hypothesis, it is scored as a negative test selection.

ANOVA excluding inaccurate participants. A repeated-measures ANOVA was performed on the adjusted scores for proportion of positive tests selected. Table 4.16 lists the results of the ANOVA.

Table 4.16

Repeated-Measures Analysis of Hypothesis Adjusted Proportion of Positive Tests Selected for Hypothesis Test Selection Questions, Excluding Inaccurate Participants (n = 159)

Source	*df*	*F*	η^2_P	η^2_G	*p*
			Between Subjects		
Order (O)	1	.156	.001	.001	.693
Instruction (I)	5	1.361	.044	.034	.242
O × I	5	.872	.029	.022	.502
S within-group error	147	(.068)			
			Within Subjects		
Scenario (S)	1	2.893	.019	.005	.091
S × O	1	.499	.004	.001	.481
S × I	5	.846	.035	.009	.380
S × O × I	5	.830	.027	.007	.538
S × *S* within-group error	147	(.022)			

Note. Values in parentheses indicate mean square errors. *S* = subjects.
$^*p < .05$

Following the adjustment for the additional hypothesis given to the alternative hypothesis group, there were no significant effects on the proportion of positive test selections. There was no evidence to suggest that the tendency to select positive tests is reduced when participants are given an alternative hypothesis to test.

The overall mean proportion of positive tests selected for the two scenarios is 88.9% (*SD* = .20) for the MDT Microworld and 86% (*SD* = .22) for the Scorpion-Spider scenario following adjustment for the alternative hypothesis. Although none of the treatments designed to effect positive test selection affected test selection, some participants did make negative test selections. Table 4.17 lists proportions of participants in the two scenarios whose test selections were 100% positive and participants whose test selections were less than 70%.

Table 4.17

Proportion of Participants with 100% Positive Test Selection or <70% Positive Test Selection, Excluding Inaccurate Participants (n = 159)

	MDT Microworld	Scorpion-Spider
100% Positive Test Selection	63.5% (*n* = 101)	58.5% (*n* = 93)
< 70% Positive Test Selection	16.4% (*n* = 26)	20.1% (*n* = 32)

Individual Differences

The participants that selected less than 70% positive tests may have some individual traits that promote negative test selection. The IPIP personality inventory described in Chapter II was included in order to investigate this possibility. Participants were scored on 10 different scales: agreeableness, conscientiousness, creativity, extraversion, inquisitiveness, insight, intellect, neuroticism, and openness to experience.

ANCOVA excluding inaccurate participants. A repeated-measures analysis of covariance with one within-subjects factor (scenario), two between-subjects factors (presentation order and instruction condition), and ten covariates (agreeableness, conscientiousness, creativity, extraversion, inquisitiveness, insight, intellect, neuroticism, and openness to experience) was performed on proportion of positive tests selected, covarying out the effect of personality traits on the proportion of positive tests selected during hypothesis test selection. The inaccurate participants were dropped from the analysis. Table 4.18 lists the results of the ANCOVA.

Of the personality covariates, creativity was significantly related to the proportion of positive tests selected, F (1, 137) = 6.016, $p < .05$, $\eta^2_P = .042$, as was extraversion, F (1, 137) = 7.405, $p < .05$, $\eta^2_P = .051$. The creativity and extraversion scales were significantly correlated, r (183) = .359, $p < .001$. The remaining personality scales were not significantly related to the proportion of positive tests selected.

There was a significant interaction between scenario and the creativity personality scale, F (1, 137) = 7.255, $p < .05$, $\eta^2_P = .045$. An examination of regression coefficients indicated that creativity was negatively correlated (B = -1.084, SE = .292, $p < .001$, η^2_P = .09) with the proportion of positive tests selected in the MDT Microworld scenario. However, in the Scorpion-Spider scenario, creativity was not significantly (B = -.256, SE = .334, p = .444) related to the proportion of positive tests selected. Extraversion regression coefficients indicated a positive correlation with positive test selection (B = .379, SE = .14, p = .005, η^2_P = .056) despite the correlation with creativity.

Table 4.18

Repeated-Measures Analysis of Covariance for Proportion of Positive Tests Selected on the Hypothesis Test Selection Questions, Excluding Inaccurate Participants (n = 159)

Source	*df*	*F*	η^2_P	*p*
		Between Subjects		
Agreeableness (A)	1	.152	.001	.697
Cautiousness (C)	1	.240	.002	.625
Conscientiousness (Co)	1	3.653	.026	.058
Creativity (Cr)	1	6.016	.042	.015*
Extraversion (X)	1	7.405	.051	.007*
Inquisitiveness (Iq)	1	.285	.002	.595
Insight (Is)	1	.776	.006	.380
Intellect (It)	1	1.255	.009	.265
Neuroticism (N)	1	1.398	.010	.239
Openness to Experience (OE)	1	.006	.000	.936
Order (O)	1	.625	.005	.431
Instruction (I)	5	1.636	.056	.155
O × I	5	.063	.035	.426
S within-group error	137	(.064)		
		Within Subjects		
Scenario (S)	1	.009	.000	.924
S × A	1	.898	.007	.345
S × C	1	1.19	.009	.277
S × Co	1	1.085	.008	.299
S × Cr	1	7.255	.050	.008*
S × X	1	.312	.002	.577
S × Iq	1	.197	.001	.658
S × Is	1	2.607	.019	.109
S × It	1	.522	.004	.471
S × N	1	.803	.006	.372
S × OE	1	1.242	.009	.267
S × O	1	.306	.002	.581
S × I	5	1.283	.045	.275
S × O × I	5	.939	.033	.458
S × *S* within-group error	137	(.020)		

Note. Values in parentheses indicate mean square errors. *S* = subjects.
$*p < .05$

Summary of Hypothesis Test Selection Results

In the hypothesis test selection questions, participants were directed to select from two figures the one figure that would most effectively test the hypothesis or hypotheses they were given. Overall, the proportion of positive tests selected during test selection was high: 89% mean positive test selection in the MDT Microworld and 86% mean positive test selection in the Scorpion-Spider scenario. A significant main effect of instruction and post-hoc analysis indicated that the alternative hypothesis instruction condition did reduce positive test selection. However, following an adjustment for the alternative hypothesis, the reduction in positive test selection was no longer significant. This suggests that the –H tests of H1 selected by alternative hypothesis instruction condition participants were also +H tests of the alternative hypothesis. Ultimately, the analyses indicated that none of the treatments (thematic content in Scorpion-Spider scenario, high miss cost instructions, common event instructions, and alternative hypothesis) had a significant effect on positive test selection.

With some participants in the Scorpion-Spider scenario selecting less than 70% of the positive tests, an analysis of 10 personality scales identified significant relationships between two personality scales, creativity and extraversion, and the selection of positive tests. Creativity was negatively correlated with positive test selection but was only significant for the MDT Microworld scenario. Extraversion was overall positively correlated with positive test selection. The remaining personality covariates were not significantly related to positive test selection.

Summary of Results

Effect of Scenario and Presentation Order

Catch trials. In the analysis of catch trials, participants appeared to continue to have problems with the Scorpion-Spider scenario despite the modifications made following similar issues in the pilot study. Overall, twenty-four participants (13%) fell below the 50% or less cutoff on at least one scenario. Of the twenty-four inaccurate participants, eighteen (75%) fell below the cutoff only on the Scorpion-Spider scenario.

Basic event prediction. In the second analysis of fundamental understanding of the task, the basic event prediction questions, there was a significant Scenario × Presentation Order interaction that indicated a possible learning effect where participants that performed the Scorpion-Spider scenario first demonstrated poor accuracy on the Scorpion-Spider scenario, then improved significantly on the MDT Microworld scenario. Participants that performed the MDT Microworld scenario first did not show the same degraded performance on the Scorpion-Spider scenario. This suggests that issues with the Scorpion-Spider scenario were affecting performance of participants that were accurate on the catch trials, and that the issues could be overcome by exposure to the other scenario instructions and materials.

Positivity assessment. In the positivity assessment questions, a more general effect of presentation order was indicated. Participants' positivity assessments tended more towards the limits (0% and 100%) in the second scenario regardless of scenario content. This may indicate that participant's confidence in their assessments increased as they performed the scenarios.

Explicit positive test selection. In the explicit positive test selection questions, there were unexplained effects of scenario and presentation order on test selection: the participants that received the MDT Microworld scenario first had a slight tendency to select more top figures than predicted given the hypotheses. The participants that received the Scorpion-Spider scenario first revealed no preference for top or bottom figures. The small effect doesn't appear to be explained by participants' assessment of individual figures given that the same pattern of results appeared when using individual assessments as the predictors. Identifying the source of the figure placement effect may reveal more about how participants choose between two tests and how test selection differs from the assessment of a single screen.

Hypothesis test selection. There was no effect of scenario or presentation order for the hypothesis test selection. The absence of the expected effect of scenario combined with the issues experienced with the Scorpion-Spider scenario in both the catch trials and the basic event prediction questions indicates that the scenario materials may be flawed.

Effect of Instruction Condition

There were six instruction conditions in this experiment. For almost every question type, the only significantly different instruction condition was the alternative hypothesis instruction condition and, in almost every case, the alternative hypothesis instruction condition was not significantly different from the other instruction conditions once the alternative hypothesis was accounted for.

Catch trials. The catch trials were designed to be an extremely simple test of a participant's fundamental understanding of the figures used to represent experimental tests. The questions had no relationship to the specific details of the instructions and, as expected, there was no effect of instruction condition on response accuracy.

Basic event prediction. In the basic event prediction questions, there was the possibility of an effect of instruction condition for the common event instructions if participants had shown an increased expectation of positive events as expected given the common event instructions. However, the common event instruction condition had no effect on basic event prediction. There was a spurious effect for the alternative hypothesis instruction condition that was corrected by adjusting for the additional positive events expected given the additional hypothesis.

Positivity assessment. In the positivity assessment questions, the common event instruction condition was expected to have a higher mean deviation from the baseline expected values because participants were expected to respond with generally higher positivity assessments than the other instruction conditions, especially for the –H tests. Instead, the common event instruction condition mean deviation from the expected values was very low for the –H tests.

There was an effect of the alternative hypothesis instruction condition on positivity assessment. Participants in the alternative hypothesis instruction condition tended to rate the likelihood that the event would occur closer to chance (50%) when the hypotheses made conflicting predictions. Although an expected result of having two contradictory hypotheses, these results did indicate that the participants were sensitive to the differences between the two hypotheses and attempted to manage their expectations accordingly.

Explicit positive test selection. On the explicit positive test selection questions, there was no expected effect of instruction condition besides the spurious effect of the alternative hypothesis instruction condition prior to adjustment for the additional hypothesis. As expected, no effect of instruction condition was found.

Hypothesis test selection. On the key hypothesis test selection questions, there were expected effects for all three of the novel instruction conditions: high miss cost, common event, and alternative hypothesis. The alternative hypothesis instruction condition was the only condition significantly different from the other instruction conditions for positive test selection. Participants in the alternative hypothesis instruction condition selected fewer positive tests of the primary hypothesis (H1). However, the alternative hypothesis instruction condition was no longer significantly different when positive tests of the alternative hypothesis (HA) were also counted as positive test selections and not as negative test selections. This suggests that the alternative hypothesis only leads to negative test selection of H1 when positive tests of HA are also negative tests of H1.

For the remaining three instruction conditions, our results did replicate the original Mynatt, Doherty, and Tweney (1977) results. The test, confirm, and disconfirm instruction conditions did not have significantly different positive test selections.

Individual Differences

A small number of participants (~20%) selected positive tests in fewer than 70% of the hypothesis test selection questions that included a negative test and a positive test. Previous research has shown that these negative testers may have individual characteristics that differentiate them from the other testers. In this study, I assessed the relationship between 10 different personality scales and positive test selection to determine if personality traits were correlated with positive test selection. Of the 10 personality scales, only two scales were significantly correlated with positive test selection. The extraversion scale, similar to the extraversion scale from the NEO-PI-R personality inventory, was overall positively correlated with positive test selection. The creativity scale, a composite of multiple scales within the International Personality Item Pool, was negatively correlated with positive test selection but only for the MDT Microworld scenario. The two scales, extraversion and creativity, were also positively correlated.

CHAPTER V

DISCUSSION

Confirmation bias, the tendency to select positive tests presumably seeking confirmatory evidence, is a thoroughly studied and difficult to modify test selection behavior. Previous research has indicated certain factors that may mediate the selection of positive tests. Some of these factors (i.e., thematic content) have been empirically explored; other factors (i.e., violation of assumptions of positive test strategy) have not. The current research implemented treatments designed to impact proposed factors and investigate the effect on positive test selection. The factors were selected from different areas of research focused on the processes of scientific discovery, integrated in the current study, and include: the effect of thematic content, the effect of high miss cost, the effect of common positive events, the effect of alternative hypotheses, and the effects of individual differences on positive test selection.

Thematic content is the application of a context to a scientific discovery or reasoning problem. Griggs and Cox (1982) indicated that a familiar context that cues memory and takes advantage of already available knowledge can improve performance on tasks such as the Wason 2-4-6 task , the Wason selection task, and their variants. Cosmides and Tooby (2005) suggest that thematic content benefits are only realized when the content is presented within a context and a perspective that takes advantage of special adaptive processes developed specifically for scenarios involving benefits and costs.

In the current research, a new context was applied to an abstract microworld developed by Mynatt, Doherty, and Tweney (1977). The new context was built around an imagined conflict between predators and prey. The objects defined by brightness and shape in the original microworld were replaced with scorpions defined by tail and claw shapes. The particles that were stopped by some objects were replaced with spiders that were eaten by some scorpions. Participants were given both scenarios and asked to answer questions about possible experiments. According to previous research, if the scenario provides a familiar context, participants may better understand the task and, in the context of positive test selection, select fewer positive tests of the hypothesis.

The results from the analysis of the catch trials and the basic event prediction questions failed to support the idea that thematic content improved participant understanding and accuracy. In fact, the results suggest that the Scorpion-Spider scenario materials led to participant confusion and poor performance on the task compared to the MDT Microworld scenario.

Although these results fail to provide support for this familiar context, the results do not preclude benefits for other possible contexts. The scorpion context may not have been a familiar, memory-cueing context for participants, limiting the benefit of the thematic content. It remains an open question whether a context based on social contract theory would have a significant impact on performance in the MDT Microworld task.

Klayman (1995; Klayman & Ha, 1987) and Friedrich (1993) claim that researchers should be able to adapt to the relative importance of a specific error type and to select tests accordingly. If false alarms are identified as the most important error to avoid, the researchers current tendencies, positive test selection, would be reinforced. However, if missed targets are the most important error to avoid, researchers should adapt and perform additional negative testing of the hypothesis.

In the current research, high miss cost instructions were developed for both the MDT Microworld and the Scorpion-Spider scenario. The intent of the instructions was to inform the participant that missed targets should be avoided. Based on Klayman's (1995) and Friedrich's (1993) claims, if participants understood that they must avoid missed targets, then participants would be expected to select more negative tests.

The results from the analyses of all five question types failed to support any effect of the high miss cost instructions on participant performance. Participants do not appear to modify their test selections when given instructions that should have led them to recognize the importance of missed targets and to select negative tests to better test whether their current hypothesis included all target events. The instructions did not directly inform participants that miss costs were high. Instead the instructions were couched in terms related to the story. Participants were given the instructions multiple times: in the audio-video instructions, in the written instructions, and in the reminders on the pages separating the question types. Despite the multiple exposures to the instructions, either participants did not understand the import of the instructions or participants understood but did not or could not operationalize the instructions as a need for selection of negative testing. If participants did not or could not act on the instructions, this suggests that Klayman and Ha (1987) and Friedrich (1993) are incorrect in assuming that a researcher will consider the relative importance of error type during test selection.

In their proposal of the positive test strategy (Klayman & Ha, 1987) present two assumptions that underlie the applicability of the positive test strategy in a given scenario. First, the target event should be rare. If the target event is rare, it is appropriate to limit your search for the target event to where it is most expected to occur. If the target event is common, it becomes more efficient to ensure that the target event does not appear where it is not expected. However, a researcher may not know how common the target event is. Additionally, Klayman (1995; Klayman & Ha, 1987) makes no claim regarding researcher's ability to adapt to knowledge of a violation of the uncommon event assumption.

The current research developed common event instructions for both scenarios. The common event instructions were intended to inform participants that the target event occurred often, leading participants to recognize that the positive test strategy was not appropriate. As with the high miss cost instructions, the instructions did not directly inform participants that the target event occurred 67% of the time. Instead, they were simply told that the event occurred regularly. If participants understood the implications of a common target event and were able to adapt their test selection behavior, participants in the high miss cost instruction condition should demonstrate increased positivity assessments for negative tests and reduced positive test selection.

The results of the analyses for all five question types failed to support any effect of common event instructions. Specifically, participants' responses to positivity assessment questions indicated that the participants did not demonstrate the expected higher positivity for negative tests compared to the other instruction conditions.

The lack of any effect due to the common event instructions suggests that participants may not modify their assessments of positivity based on a simple one-line report of high base rates. Participants were unable, given the instructions, to recognize the violation of the uncommon event assumption and adapt their test selections accordingly. Klayman (1995) does not claim that participants are aware of the assumptions underlying the positive test strategy. In the current study, the evidence supports the idea that participants are either unaware of the assumptions, unaware of violations of the assumptions, or unable to modify their testing behavior even when aware of a violation of one or more of the assumptions.

Tweney et al. (1980) created a dual-goal version of the Wason 2-4-6 task that demonstrated the potential benefit of considering two complementary hypotheses during test selection. By considering two hypotheses, the researcher searches a broader space, generates more varied tests, and is more likely to run –H tests of the primary hypothesis. In a real-world problem, a complementary and exhaustive alternative hypothesis offers little benefit because the alternative hypothesis covers the entire –H space of tests which may be quite large and impractical to test. In the DAX/MED variant of the Wason 2-4-6 task, all tests of HA or H1 provide information related to both hypotheses. If the alternative hypothesis is not complementary, participants may not receive the same benefits.

In the current research, participants in the alternative hypothesis instruction condition were given a clear alternative hypothesis offered by a second researcher in the scenario description, the audio/video instructions, the separator pages, and both scenarios. The alternative hypothesis was selected to provide a hypothesis that was not complementary to the primary and was not exhaustive. Additionally, the two hypotheses overlap. A single figure could represent a test that was a positive test of both hypotheses, positive for one and negative for the other, or negative for both hypotheses. The researcher must interpret the results of a test with respect to the details of both hypotheses.

Based on previous research, the alternative hypothesis was expected to increase selection of –H tests for H1 in the hypothesis test selection questions. Based on the nature of the additional hypothesis, the alternative hypothesis instruction condition was expected to impact all question types except the catch trials. In most question types, the alternative hypothesis instruction condition required adjusted scores to account for both hypotheses. The correction was expected to remove the effect of the alternative hypothesis for all questions including the hypothesis test selection questions.

The results of the analyses supported the impact of alternative hypotheses on many aspects of test selection. Prior to corrections, the alternative hypothesis instruction condition differed in all analyses except the analysis of catch trial response accuracy. The correction for the additional hypothesis removed the effect of the alternative hypothesis instruction condition from the basic event prediction questions and the explicit positive test selections. Positivity assessments were affected because the adjusted scoring did not account for figures with outcomes that were contradictory depending on the hypothesis. The additional hypothesis did reduce positive test selection for the hypothesis test selection questions, but the effect was not significant once +H tests of HA were counted as positive tests.

Given the results of the positivity assessment questions, the alternative hypothesis instructions had a significant effect on participant understanding of the task and figures. Participants were sensitive to contradictions between the primary and the alternative hypotheses when assessing positivity. The participants given the alternative hypothesis also demonstrated reduced positive test selection for the primary hypothesis. However, the negative hypothesis tests of the primary hypothesis tended to be positive hypothesis tests of the alternative hypothesis. The alternative hypothesis superficially increases negative hypothesis testing of the primary hypothesis not by modifying test selection behavior but by refocusing the tendency to select positive tests on a second overlapping or complementary hypothesis. The effectiveness of the consideration of an alternative hypothesis is dependent on the relationship between the primary hypothesis and the alternative hypothesis.

If the alternative hypothesis is sufficiently different from the primary hypothesis, then more negative tests of the primary hypothesis are likely to be selected. If the alternative hypothesis is too similar to the primary hypothesis, most positive hypothesis tests of the alternative will be positive tests of the primary and few of the positive tests of the alternative will be negative tests of the primary. If the alternative hypothesis is too dissimilar (i.e., a complementary hypothesis), many or all of the positive tests of the alternative hypothesis will be negative tests of the primary hypothesis but, when the problem space matches the assumptions of the positive test strategy, the search space may be too large or include too few tests that would falsify the primary hypothesis to provide a significant benefit. The results of the current study are in line with previous research demonstrating increased negative test selection for the primary hypothesis but the results bring into question the prescriptive benefits of adopting one or more alternative hypotheses.

Previous research on individual differences in scientific discovery and reasoning tasks have focused on traditional differences (gender: no effect, arts and sciences educational background: no effect, Wason, 1960), education and experience (mixed results depending on context, Tweney & Yachanin, 1985; Griggs & Ransdell, 1986), and cognitive ability (SAT scores, Stanovich, 1999). More recently, Vartanian, Martindale, and Kwiatkowski (2003) demonstrated individual difference measures of divergent thinking correlated with reasoning performance. Divergent thinking tests have been correlated to personality traits from the NEO-PI-R including creativity and openness to experience (Costa & McCrae, 1992).

In the current research, ten personality scales (5 NEO-PI-R scales, 5 scales face valid for scientific discovery) were developed using the International Personality Item Pool. The correlation between personality scales for participants and positive test selection were tested to determine if individual differences in personality affected positive test selection.

The results from an analysis of covariance supported the idea that certain personality traits were correlated with positive test selection. Two traits were related to positive test selection: extraversion and creativity. Extraversion was positively correlated overall with positive test selection. Creativity was negatively correlated with positive test selection but only for the Mynatt, Doherty, and Tweney microworld scenario.

Previous research has identified a complex relationship between creativity and extraversion. Introversion has often been linked to creativity (e.g., Feist, 1999). However, extraversion has also been linked to creativity (Eysenck, 1993; Martindale, 2007). Martindale suggested that the relationship between high extraversion and high creativity is based on a tendency towards disinhibited thinking. However, in the current study, introverted participants with high creativity selected more negative tests than extraverted participants, even though extraverted participants were more likely to be high creative types. Feist (1999) identified autonomy and introversion as key components of personality traits associated with scientists. Although extraversion and creativity are correlated, the different effects on test selection are due to other aspects of the personality traits.

The increased positive test selection for high extraverts may be explained by the need for increased arousal in extraverts (Costa & McCrae, 1992). This need for arousal may lead extraverts to prefer positive tests that would lead to interesting events or may lead extraverts to rely on the positive test strategy without deeply considering the tests in order to quickly complete the task. The reduced positive tests selection for high creative types may be due to their increased autonomy and/or curiosity. The different effects for the two scenarios may be due to different opportunities for creative thought. The MDT Microworld is an abstract environment with no constraints on interpretation of the meaning of the objects, particles, and rules. The Scorpion-Spider scenario, by virtue of its familiar context, constrains interpretation within the context provided to the participant. The abstract nature of the MDT Microworld may better allow application of creative tendencies that the Scorpion-Spider scenario does not.

Given these results, creativity and related personality traits appear to be good candidates for consideration in future investigations of individual differences and scientific discovery or reasoning. Individual differences provide a different perspective and avenue of exploration for understanding positive test selection.

Implications for Methodology

Previous research in hypothesis test selection has often asked participants to report their current hypothesis (verbally or in writing), select a test, and state why the test was selected. Although the data provided is undoubtedly helpful in analysis, this procedure seems very unlike the scientific discovery process that is supposedly being scrutinized. Test selection may be done without specific hypotheses in mind (Schunn & Klahr, 1993) and the reports generated by the participants may be just-so stories generated after the fact by participants for the experimenter. Verbal protocols may provide more accurate insight into participants' thought processes but they can be cumbersome for participants, time consuming for researchers, and the analysis is typically constrained to a very low N that will limit investigation of individual differences. The current research uses five simple questions about the figures, the hypotheses, the relationships between the figures and the hypotheses, and the relationship between two tests. The questions probe participant understanding and processes with minimal impact on the task. Additional, carefully selected questions will further illuminate participant understanding and provide new measures of the impact of other instructional treatments.

Summary

In the current research, few of the treatments designed to modify test selection behavior had any effect. The Scorpion-Spider scenario materials proved confusing to some participants and mitigated the benefits of individual differences in personality.

Participants were unable to appropriately respond to the instructions that informed the participants of relevant aspects of the underlying nature of the problem (high miss cost and common event instructions). The alternative hypothesis instruction condition was the only treatment effective at reducing positive test selections but, even in this case, test selection behavior was not modified; it was simply redirected at a second hypothesis that led to negative test selections. The benefits of an alternative hypothesis appear to be limited according to the relationship between the primary hypothesis and the alternative hypothesis. Creativity (and the associated personality trait introversion) leads to reduced positive test selection especially in the more abstract context. Given the issues with the group treatments, the importance of considering individual differences when investigating scientific discovery and reasoning was reinforced by our results.

The current research has implications for the positive test strategy and research methodology for investigating scientific discovery. The Klayman and Ha (1987) analysis of test selection may be an effective analysis for explaining why participants tend to select positive tests: an effective method for testing learned from real-world experience. Although positive test selection may be learned from real-world interaction, the current results indicate that we cannot modify our test selection behavior even when given information that should indicate that positive test selection will be ineffective. Our test selection behavior appears to not be a strategy, but a reasonable, learned, and ingrained bias that is highly resistant to modification.

REFERENCES

Ashton, M.C., Lee, K., & Goldberg, L. R. (2007). The IPIP-HEXACO scales: An alternative, public-domain measure of the personality constructs in the HEXACO model. *Personality and Individual Differences, 42*(8), 1515-1526.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37*(3), 379-384.

Cattell, H. B., & Schuerger, J. M. (2003). *Essentials of 16PF assessment*. Essentials of psychological assessment series. Hoboken, NJ: John Wiley & Sons.

Cheng, P. W., & Holyoak, K. J. (1989). On the natural selection of reasoning theories. *Cognition, 33*, 285-313.

Cosmides, L., & Tooby, J. (2005). Neurocognitive adaptations designed for social exchange. In D. M. Buss (Ed.), *The Handbook of Evolutionary Psychology* (584-627). Hoboken, NJ: John Wiley & Sons.

Costa, P. T., Jr., & McCrae, R. R. (1992). *NEO PI-R professional manual.* Odessa, FL: Psychological Assessment Resources, Inc.

Eysenck, H. J. (1993). Creativity and personality: Suggestions for a theory. *Psychological Inquiry, 4*(3), 147-178.

Feist, G. J. (1998). A meta-analysis of personality in scientific and artistic creativity. *Personality and Social Psychology Review, 2*(4), 290-309.

Friedrich, J. (1993). Primary error detection and minimization (PEDMIN) strategies in social cognition: A reinterpretation of confirmation bias phenomena. *Psychological Review, 100*(2), 298-319.

Gale, M., & Ball, L. J. (2003). Facilitated 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*. Boston, MA: Cognitive Science Society.

Gale, M., & Ball, L. J. (2005). Contrast-class cues and dual goal facilitation in Wason's 2-4-6 task: Evidence for an extended iterative counterfactual model. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition*, *43*, 127-171.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., et al. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84-96.

Gough, H. G., & Bradley, P. (1996). *CPI Manual*. (3rd ed.) Palo Alto, CA: Consulting Psychologists Press.

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology, 73,* 407-420.

Griggs, R. A., & Ransdell, S. E. (1986). Scientists and the selection task. *Social Studies of Science, 16*(2), 319-330.

Hofstee, W. K. B., de Raad, B., & Goldberg, L. R. (1992). Integration of the big five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology, 63,* 146-163.

Hogan Assessment Systems (2007). Hogan Personality Inventory – scales and interpretations. Retrieved August 8, 2007, from http://www.hoganassessment.com/products_services/hpi_scales_interpretations.aspx

International Personality Item Pool: A scientific collaboratory for the development of advanced measures of personality traits and other individual differences. (n.d.). Retrieved August 8, 2007, from http://ipip.ori.org/

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, *12*, 1-48.

Klayman, J. (1995). Varieties of confirmation bias. *The Psychology of Learning and Motivation, 32*, 385-418.

Klayman, J., & Brown, K. (1993). Debias the environment instead of the judge: an alternative approach to reducing error in diagnostic (and other) judgment. *Cognition, 49*, 97-122.

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.

Martindale, C. (2007). Creativity, primordial cognition, and personality. *Personality and individual differences, 43*, 1777-1785.

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

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

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8*(4), 434-447.

Peterson, C., & Seligman, M. E. P. (2004). *Character strengths and virtues: A classification and handbook.* New York: Oxford University Press/Washington, DC: American Psychological Association.

Popper, K. R. (1959). *The logic of scientific discovery.* New York: Basic Books.

Popper, K. R. (1962). *Conjectures and refutations*. New York: Basic Books.

Schunn, C. D., & Klahr, D. (1993). Self vs. other-generated hypotheses in scientific discovery. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Slowiaczek, L. M., Klayman, J., Sherman, S. J., & Skov, R. B. (1992). Information selection and use in hypothesis testing: What is a good question, and what is a good answer? *Memory and Cognition, 20*(4), 392-405.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning.* Mahwah, NJ: Lawrence Erlbaum Associates.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*, 645-726.

Tweney, R. D., Doherty, M. E., Worner, W. J., Pliske, D. B., Mynatt, C. R., Gross, K. A., et al. (1980). Strategies of rule discovery in an inference task. *Quarterly Journal of Experimental Psychology, 32*(1), 109-123.

Tweney, R. D., & Yachanin, S. A. (1985). Can scientists rationally assess conditional inferences? *Social Studies of Science, 15*(1), 155-173.

Vallée-Tourangeau, F., & Penney, A. K. (2005). The impact of external representation in a rule discovery task. *European Journal of Cognitive Psychology, 17*(6), 820-834.

Vallée-Tourangeau, F., Penney, A. K., & Payton, T. (2005). Representational effects in a rule discovery task. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Twenty-seventh Annual Conference of the Cognitive Science Society*, Mahwah, NJ: Lawrence Erlbaum Associates.

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. In W. D. Gray & C. Schunn (Eds.), *Proceedings of the Twenty-fourth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Vartanian, O., Martindale, C., & Kwiatkowski, J., (2003). Creativity and inductive reasoning: The relationship between divergent thinking and performance on Wason's 2-4-6 task. *Quarterly Journal of Experimental Psychology, 56A*(4), 641-655.

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

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology* (pp. 135-151). Harmondsworth, Middlesex, UK: Penguin Books.

APPENDIX A

INTERNATIONAL PERSONALITY ITEM POOL APPLICATION INSTRUCTIONS AND INVENTORY ITEM LISTING

IPIP Application Instructions

> "You will see a series of phrases describing people's behavior. For each statement, you will need to select from five alternative buttons to indicate how accurately the statement describes you. Describe yourself as you generally are now, not as you wish to be in the future. Be honest: how do you feel you compare to other people you know of the same sex as you are, and roughly your same age?
>
> Please read each statement carefully, and then click on the button that best describes your choice.
>
> Click on the 'Continue' button below to begin the personality inventory."

Table A.1

Listing of Items Used for IPIP Personality Inventory, Grouped by Scale

Item	Positive or Negative Key
Neuroticism	
Often feel blue.	Positive
Am often down in the dumps.	Positive
Dislike myself.	Positive
Have frequent mood swings.	Positive
Panic easily.	Positive
Seldom feel blue.	Negative
Rarely get irritated.	Negative
Am not easily bothered by things.	Negative
Feel comfortable with myself.	Negative
Am very pleased with myself.	Negative
Extraversion	
Feel comfortable around people.	Positive
Make friends easily	Positive

Know how to captivate people.	Positive
Am skilled in handling social situations.	Positive
Am the life of the party.	Positive
Don't talk a lot.	Negative
Don't like to draw attention to myself.	Negative
Would describe my experiences as somewhat dull.	Negative
Have little to say.	Negative
Keep in the background.	Negative

Openness to Experience	
Have a vivid imagination.	Positive
Carry the conservation to a higher level.	Positive
Believe in the importance of art.	Positive
Tend to vote for liberal political candidates.	Positive
Enjoy hearing new ideas.	Positive
Avoid philosophical discussions	Negative
Do not like art.	Negative
Am not interested in abstract ideas.	Negative
Do not enjoy going to art museums.	Negative
Tend to vote for conservative political candidates.	Negative

Agreeableness	
Have a good word for everyone.	Positive
Respect others.	Positive
Believe that others have good intentions.	Positive
Accept people as they are.	Positive
Make people feel at ease.	Positive
Cut others to pieces.	Negative
Suspect hidden motives in others.	Negative
Have a sharp tongue.	Negative
Insult people.	Negative
Get back at others.	Negative

Conscientiousness	
Pay attention to details.	Positive
Am always prepared.	Positive
Get chores done right away	Positive
Carry out my plans.	Positive
Make plans and stick to them.	Positive

Do just enough work to get by.	Negative
Waste my time.	Negative
Shirk my duties.	Negative
Find it difficult to get down to work.	Negative
Don't see things through.	Negative

Cautiousness

Avoid mistakes.	Positive
Stick to my chosen path.	Positive
Choose my words with care.	Positive
Make rash decisions.	Negative
Like to act on a whim.	Negative
Often make last-minute plans	Negative
Rush into things.	Negative
Jump into things without thinking.	Negative
Act without thinking.	Negative

Creativity

Have an imagination that stretches beyond that of my friends.	Positive
Can easily link facts together.	Positive
Like to solve complex problems.	Positive
Challenge others' points of view.	Positive
Am able to come up with new and different ideas.	Positive
Have excellent ideas.	Positive
Am an original thinker.	Positive
Ask questions that nobody else does.	Positive
Am full of ideas.	Positive
Like to think of new ways to do things.	Positive
Quickly think up new ideas.	Positive
Come up with new ways to do things.	Positive
Come up with something new.	Positive
Have no special urge to do something original.	Negative
Am not interested in theoretical discussions.	Negative
Have difficulty understanding abstract ideas.	Negative
Try to avoid complex people.	Negative
Do not have a good imagination.	Negative
Am not considered to have new and different ideas.	Negative
Don't pride myself on being original.	Negative

Have difficulty imagining things.	Negative
Seldom experience sudden intuitive insights.	Negative
Have trouble guessing how others will react.	Negative
Consider myself an average person.	Negative
Am not interested in speculating about things.	Negative
Inquisitiveness	
Am interested in science.	Positive
Enjoy intellectual games.	Positive
Would love to explore strange places.	Positive
Don't bother worrying about political and social problems.	Negative
Don't know much about history.	Negative
Insight	
Put a new perspective on things.	Positive
Come up with alternatives.	Positive
Throw a new light on the situation.	Positive
Say nothing new.	Negative
Intellect	
Can handle a lot of information.	Positive
Spend time reflecting on things.	Positive
Am quick to understand things.	Positive
Like to get lost in thought.	Positive
Enjoy thinking about things.	Positive
Know how things work.	Positive
Will not probe deeply into a subject.	Negative

APPENDIX B

PILOT STUDY QUESTIONNAIRES

MDT Microworld Questionnaire

What’s going to happen? Judging image representations of experiments.

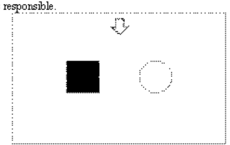
PACKET #1

Please complete this packet first.

A scientist has noticed that certain objects will stop a particle that has been fired in the direction of the object, while other objects allow the particle to pass by. The scientist is trying to understand which objects block particles and which don’t. The objects vary in both shape and color. Possible shapes include *triangles, squares,* and *disks,* while the colors are *black* and *white*. Thus, there are a total of six different objects the scientist must consider.

We would like you to consider certain experiments that the scientist might perform to help determine what characteristic or characteristics of the object stop particles. Unfortunately, it is not always possible to isolate an object, so sometimes the scientist must experiment with a pair of objects in the environment. You will be asked to help advise the scientist about various experiments that might be performed. The specific tasks will differ slightly throughout the session and you will be given specific instructions for each question.

Before we proceed to the questions, let us consider an example situation that will help you to understand more about the task. Notice in the figure below there are two objects, a *black square* and a *white disc*. The particles are emitted from the arrow at the top of the figure and will move straight down towards the objects. The particle may be stopped, or it may pass through the object or objects. Because the object stops before it reaches either object, when you have a pair of objects that stop the particle, you cannot determine which object is responsible.

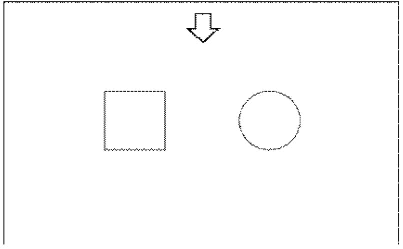


After conducting some preliminary research, the scientist believes that particles are stopped by *triangular objects regardless of their color*. We ask that you adopt this hypothesis as you answer questions about experiments.

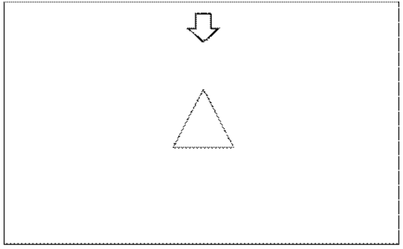
For example, you might be asked “Will the particle to be stopped by the objects in this experiment?” You should answer “No” because there is no triangle in this experiment, and by our hypothesis, only triangles (of any color) will block particles.

We want to give the best possible advice to the scientist about each question, so please carefully consider the question and your answer before moving on to the next one. Do not worry about answering quickly: you will have plenty of time to complete the experiment in the allotted time. If you have any questions before beginning the experiment, please raise your hand and the experimenter will come over and help you. Otherwise you may begin the experiment.

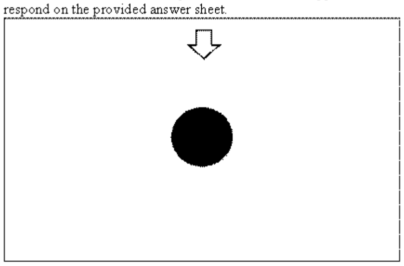
1. Given the test below, will the particle’s motion be stopped? Please respond on the provided answer sheet.



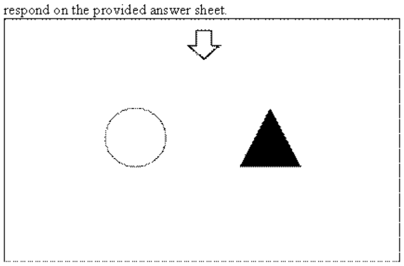
2. Given the test below, will the particle’s motion be stopped? Please respond on the provided answer sheet.



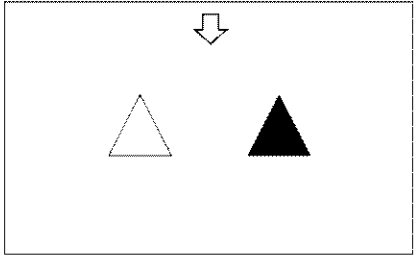
3. Given the test below, will the particle’s motion be stopped? Please respond on the provided answer sheet.



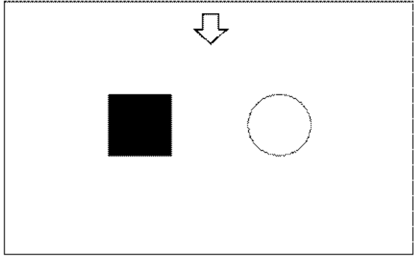
4. Given the test below, will the particle’s motion be stopped? Please respond on the provided answer sheet.



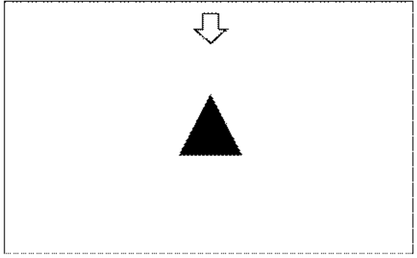
5. Given the test below, will the particle's motion be stopped? Please respond on the provided answer sheet.



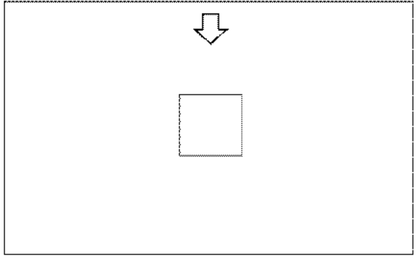
6. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.



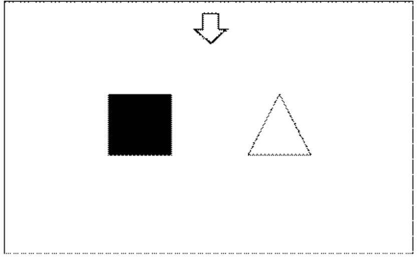
7. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.



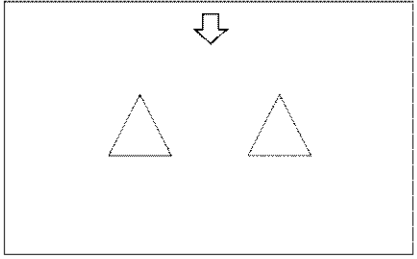
8. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.



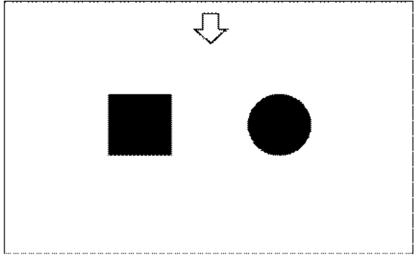
9. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.



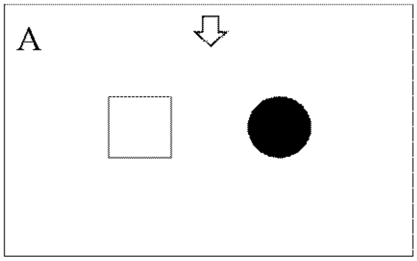
10. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.

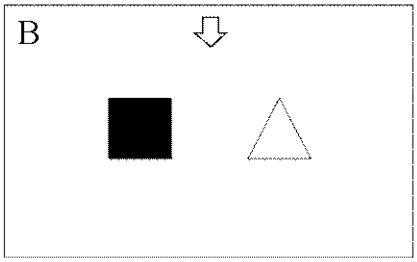


11. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.

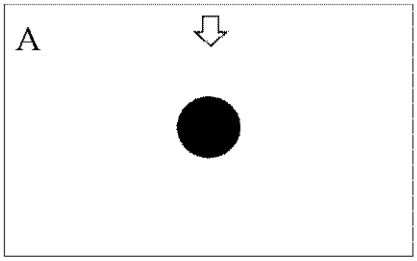


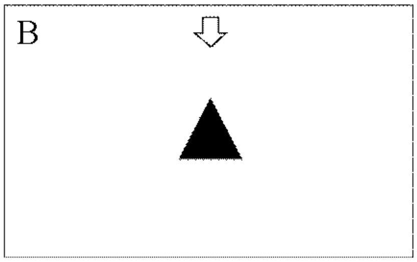
12. On the provided answer sheet, please circle the letter matching the experiment that contains a triangle:



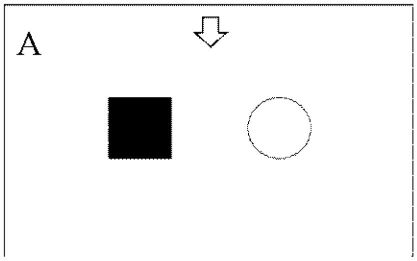


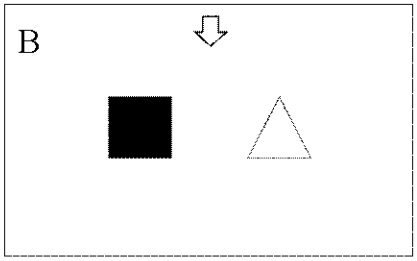
13. On the provided answer sheet, please circle the letter matching the experiment that contains a triangle:



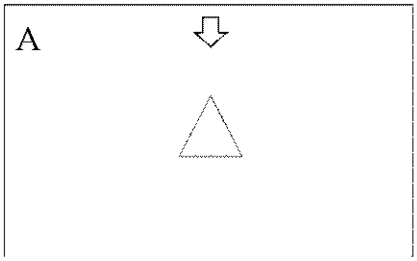


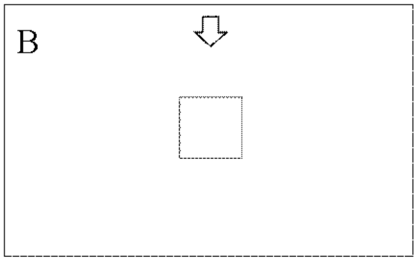
14. On the provided answer sheet, please circle the letter matching the experiment that is most likely to result in the particle's motion being stopped:



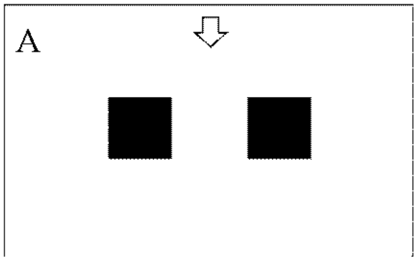


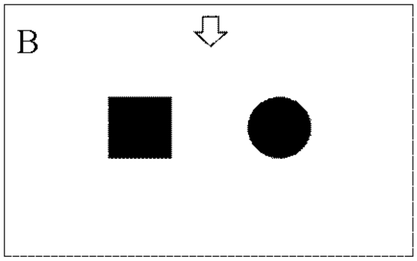
15. On the provided answer sheet, please circle the letter matching the experiment that is most likely to result in the particle's motion being stopped:



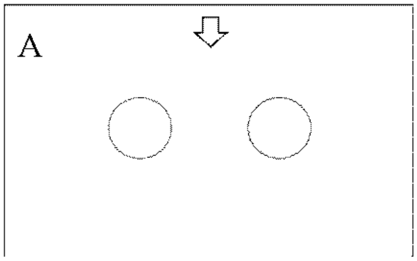


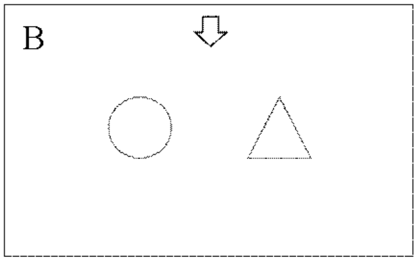
16. On the provided answer sheet, please circle the letter matching the experiment that is most likely to result in the particle's motion being stopped:



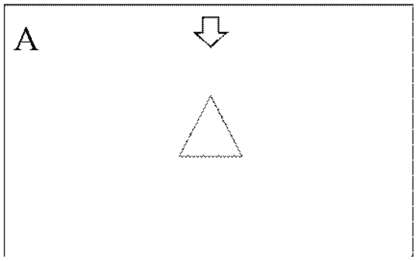


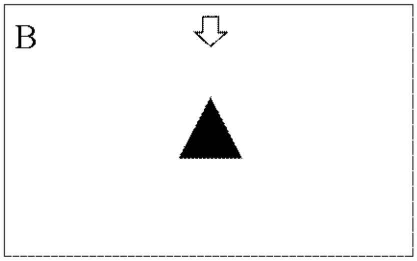
17. On the provided answer sheet, please circle the letter matching the experiment that is most likely to result in the particle's motion being stopped:



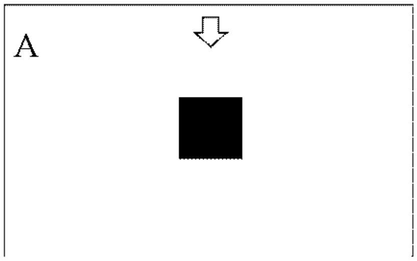


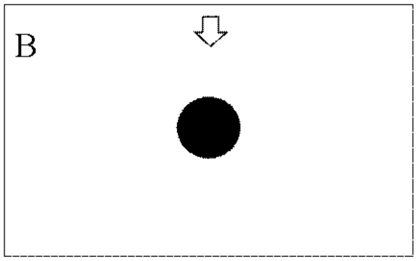
18. On the provided answer sheet, please circle the letter matching the experiment that is most likely to result in the particle's motion being stopped:



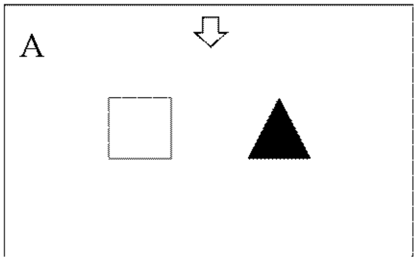


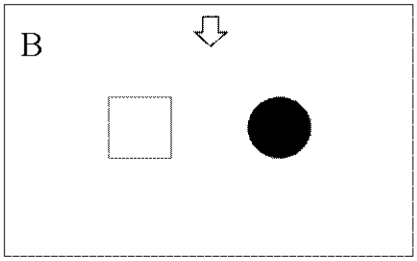
19. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:



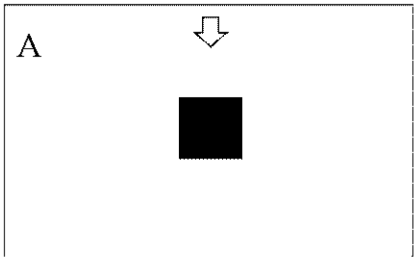


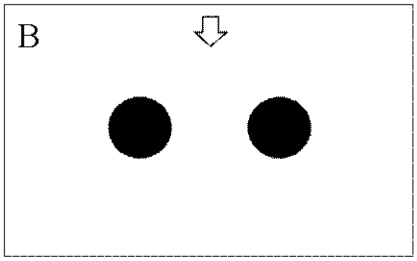
20. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:



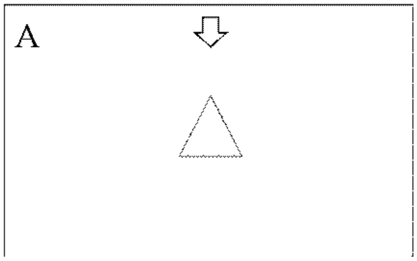


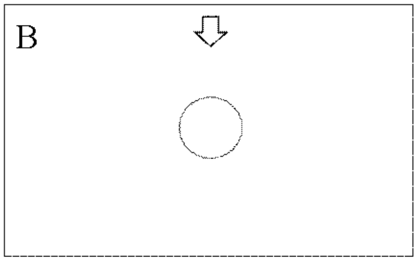
21. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:



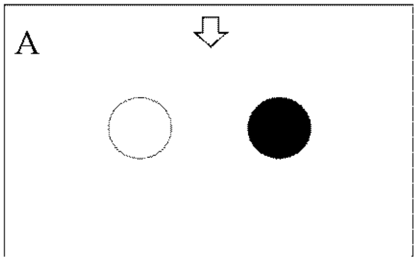


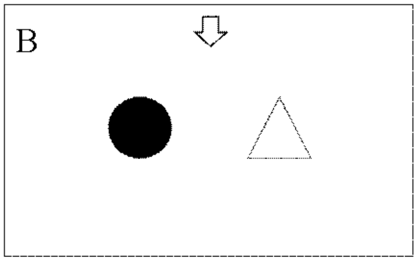
22. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:





23. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:





You have completed Packet #1.

Please complete the personality inventory on the computer. If you have any questions, please raise your hand and the experimenter will assist you.

Scorpion- Spider Questionnaire

What's going to happen? Judging image representations of experiments.

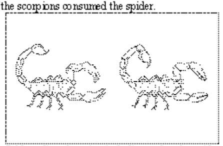
PACKET #2

Please complete this packet second.

Farmers recently introduced 6 types of scorpions to their farmland in an attempt to deal with a pest problem. Unfortunately, the scorpions have had the opposite effect as the number of pest insects has actually increased. A scientist has noticed that the scorpions are attacking and consuming a particular species of spider that was previously eating many of the insects. The scientist wants to understand what determines which types of scorpion eat the spiders and which don't. The scorpions can be identified by the shape of their tail and the shape of their claws. The scorpion tail's can be curled *up* or curled *down* while the claws' can be *serrated, thick,* or *pincher* claws. There are a total of six different scorpions the scientist must consider.

We would like you to consider certain experiments that the scientist might perform to help understand what features of the scorpions determine whether or not they will eat the spiders. Unfortunately, it is not always possible to isolate one type of scorpion, so sometimes the scientist must experiment with a pair of scorpions in the environment. You will be asked to help advise the scientist about various experiments that might be performed. The specific tasks will differ slightly throughout the session and you will be given specific instructions for each question.

Before we proceed to the questions, let us consider an example situation that will help you to understand more about the task. In the figure below, there are two scorpions, one with a tail that curls *up* and *thick* claws and the other with a tail that curls *down* and *serrated* claws. In an experiment, a spider will be introduced to the enclosed environment for a week or until the spider is eaten and consumed. The spider may be eaten by the scorpion or scorpions, or it may be ignored by the scorpions. The scorpions cannot be observed 24 hours, so when you have two scorpions in the environment and the spider is eaten, you cannot determine which of the scorpions consumed the spider.

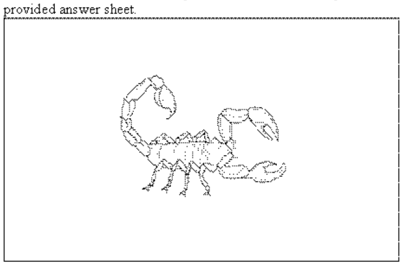


Based on some preliminary field studies, the scientist believes that the spiders are being attacked and consumed by *scorpions with pincher claws regardless of their tail shape.* We ask that you adopt this hypothesis as you answer questions about experiments.

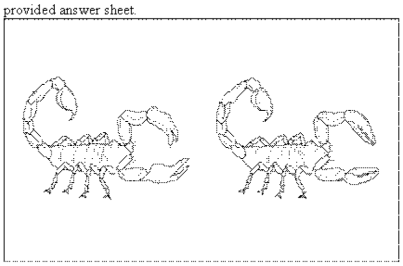
For example, you might be asked "Will the spider be eaten in this experiment?" You should answer "No" because none of the scorpions in this experiment have pincher claws, and by our hypothesis, scorpions must have pincher claws (and any tail shape) to attack and consume the spiders.

We want to give the best possible advice to the scientist about each question, so please carefully consider the question and your answer before moving on to the next one. Do not worry about answering quickly: you will have plenty of time to complete the experiment in the allotted time. If you have any questions before beginning the experiment, please raise your hand and the experimenter will come over and help you. Otherwise you may begin the experiment.

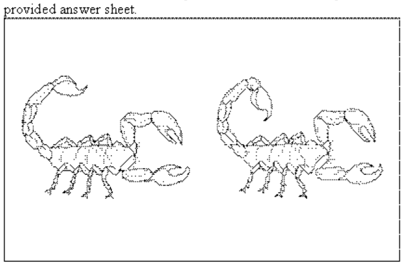
1. Given the test below, will the spider be eaten? Please respond on the provided answer sheet.



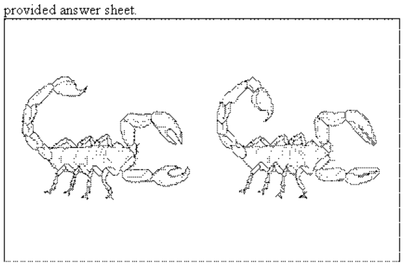
2. Given the test below, will the spider be eaten? Please respond on the provided answer sheet.



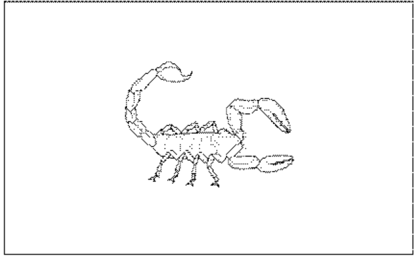
3. Given the test below, will the spider be eaten? Please respond on the provided answer sheet.



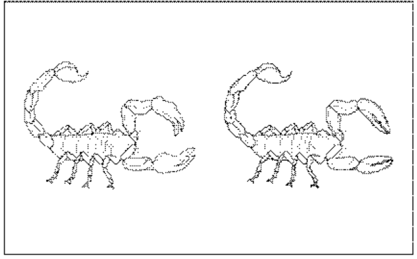
4. Given the test below, will the spider be eaten? Please respond on the provided answer sheet.



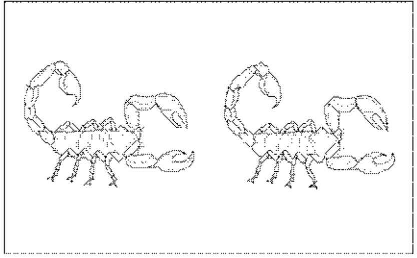
5. Given the test below, will the spider be eaten? Please respond on the provided answer sheet.



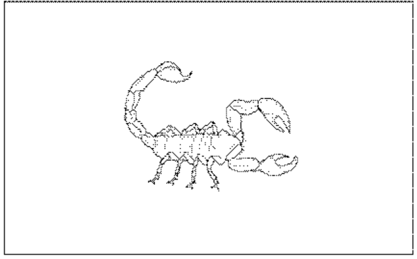
6. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.



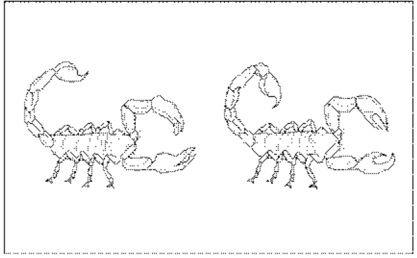
7. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.



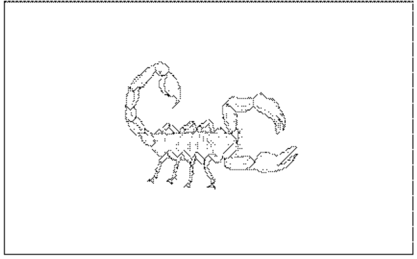
8. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.



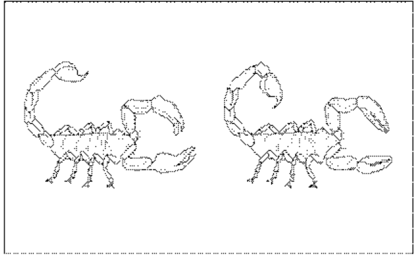
9. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.



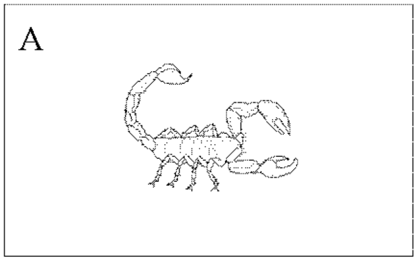
10. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.

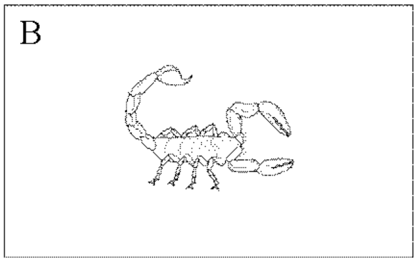


11. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.

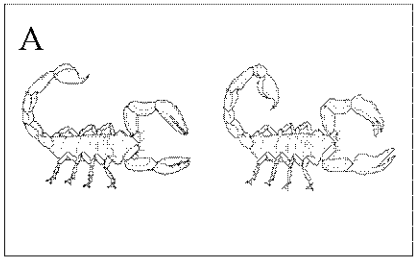


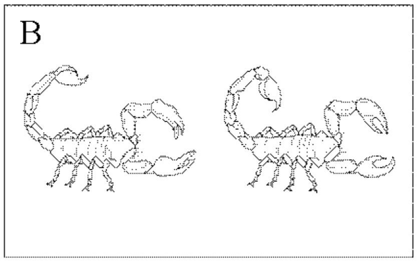
12. On the provided answer sheet, please circle the letter matching the experiment that contains a scorpion with pincher claws:



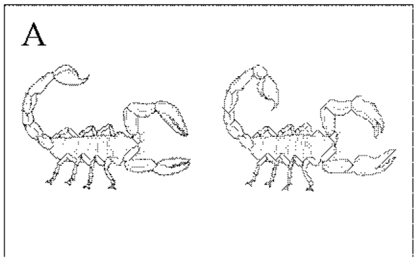


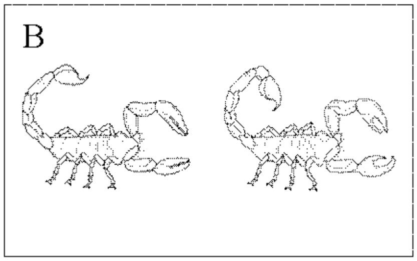
13. On the provided answer sheet, please circle the letter matching the experiment that contains a scorpion with pincher claws:



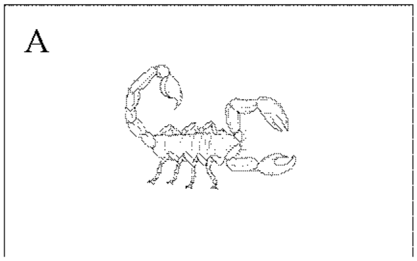


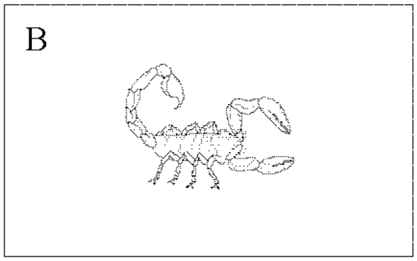
14. On the provided answer sheet, please circle the letter matching the experiment that is most likely to result in the spider being eaten:



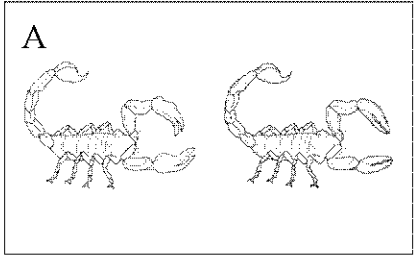


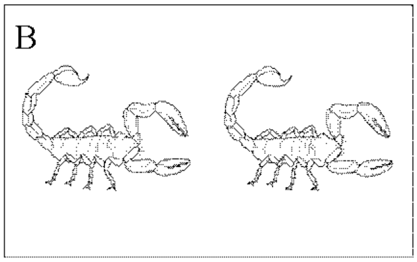
15. On the provided answer sheet, please circle the letter matching the experiment that that is most likely to result in the spider being eaten:



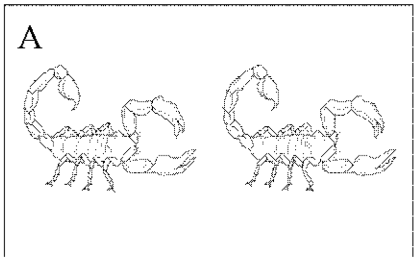


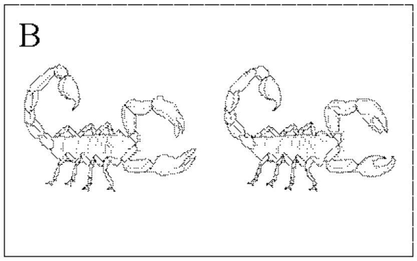
16. On the provided answer sheet, please circle the letter matching the experiment that that is most likely to result in the spider being eaten:



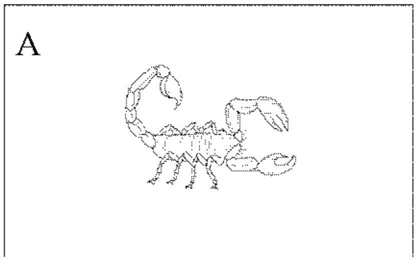


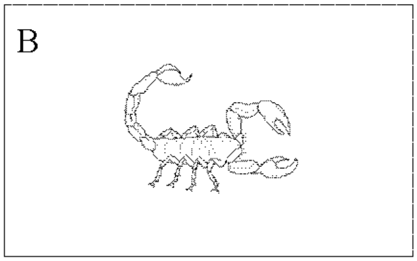
17. On the provided answer sheet, please circle the letter matching the experiment that that is most likely to result in the spider being eaten:



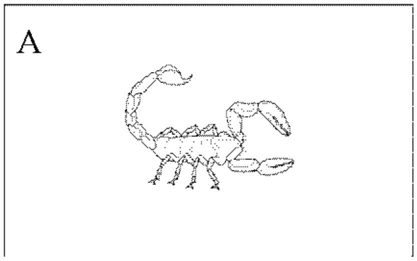


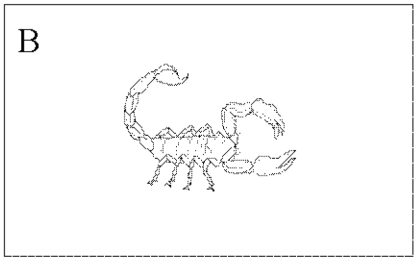
18. On the provided answer sheet, please circle the letter matching the experiment that that is most likely to result in the spider being eaten:



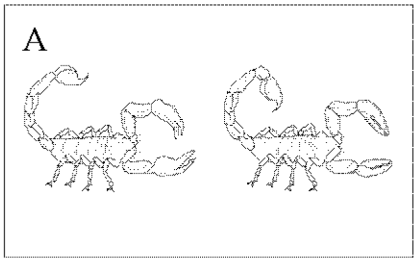


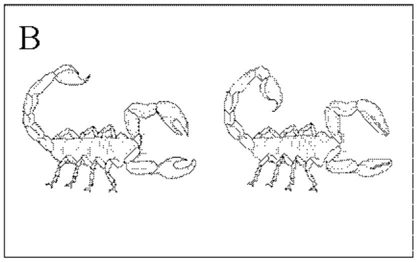
19. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:



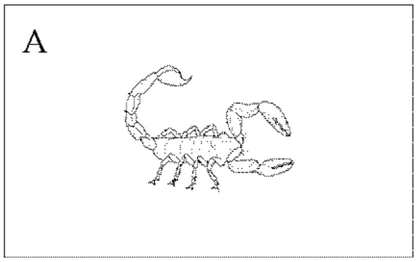


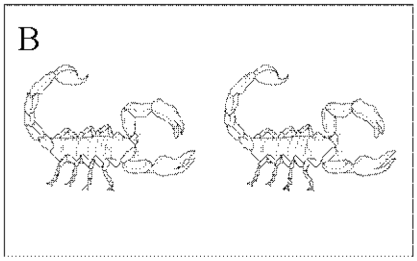
20. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:



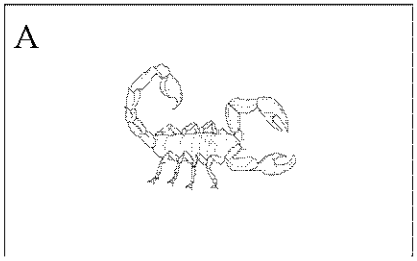


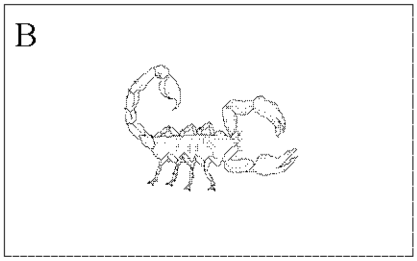
21. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:



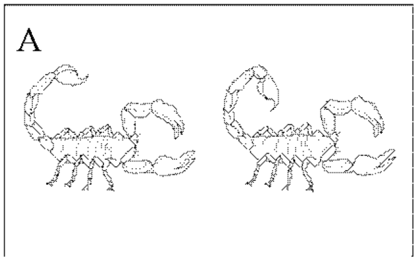


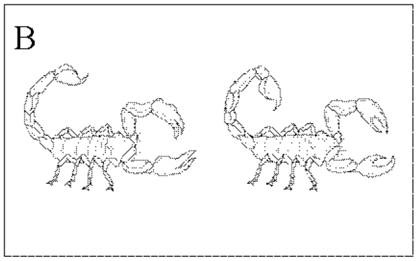
22. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:





23. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:





You have completed Packet #2.

Thank you for your participation in this experiment.

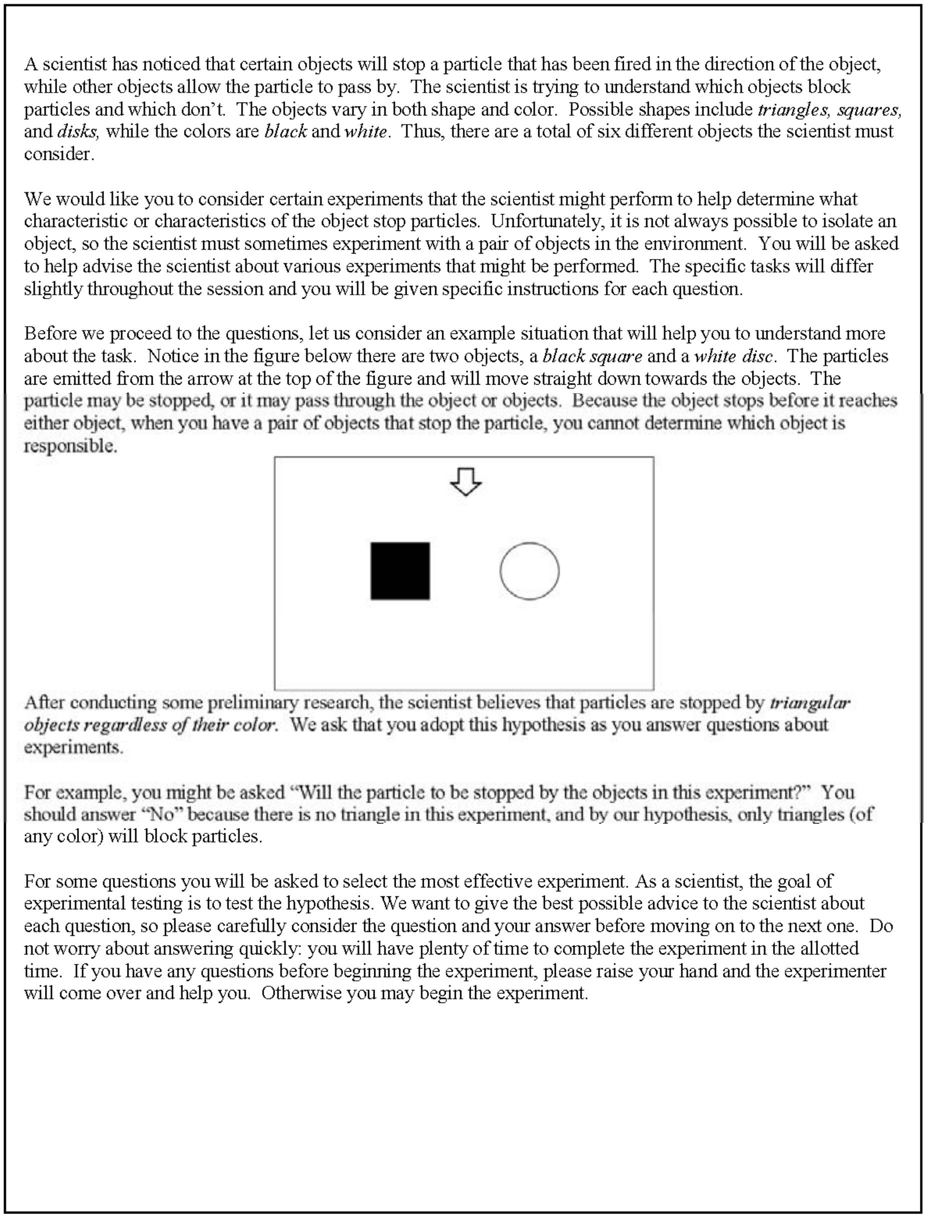
Please leave the packets and answer sheets at your station and raise your hand to let the experimenter know that you have finished.

APPENDIX C

QUESTIONNAIRES

MDT Microworld Scenario

Test Instruction Condition



A scientist has noticed that certain objects will stop a particle that has been fired in the direction of the object, while other objects allow the particle to pass by. The scientist is trying to understand which objects block particles and which don't. The objects vary in both shape and color. Possible shapes include *triangles, squares,* and *disks,* while the colors are *black* and *white*. Thus, there are a total of six different objects the scientist must consider.

We would like you to consider certain experiments that the scientist might perform to help determine what characteristic or characteristics of the object stop particles. Unfortunately, it is not always possible to isolate an object, so the scientist must sometimes experiment with a pair of objects in the environment. You will be asked to help advise the scientist about various experiments that might be performed. The specific tasks will differ slightly throughout the session and you will be given specific instructions for each question.

Before we proceed to the questions, let us consider an example situation that will help you to understand more about the task. Notice in the figure below there are two objects, a *black square* and a *white disc*. The particles are emitted from the arrow at the top of the figure and will move straight down towards the objects. The particle may be stopped, or it may pass through the object or objects. Because the object stops before it reaches either object, when you have a pair of objects that stop the particle, you cannot determine which object is responsible.

After conducting some preliminary research, the scientist believes that particles are stopped by ***triangular objects regardless of their color***. We ask that you adopt this hypothesis as you answer questions about experiments.

For example, you might be asked "Will the particle to be stopped by the objects in this experiment?" You should answer "No" because there is no triangle in this experiment, and by our hypothesis, only triangles (of any color) will block particles.

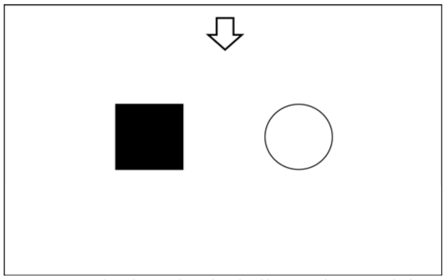
For some questions you will be asked to select the most effective experiment. As a scientist, the goal of experimental testing is to test the hypothesis. We want to give the best possible advice to the scientist about each question, so please carefully consider the question and your answer before moving on to the next one. Do not worry about answering quickly: you will have plenty of time to complete the experiment in the allotted time. If you have any questions before beginning the experiment, please raise your hand and the experimenter will come over and help you. Otherwise you may begin the experiment.

Confirm Instruction Condition

A scientist has noticed that certain objects will stop a particle that has been fired in the direction of the object, while other objects allow the particle to pass by. The scientist is trying to understand which objects block particles and which don't. The objects vary in both shape and color. Possible shapes include *triangles, squares,* and *disks,* while the colors are *black* and *white*. Thus, there are a total of six different objects the scientist must consider.

We would like you to consider certain experiments that the scientist might perform to help determine what characteristic or characteristics of the object stop particles. Unfortunately, it is not always possible to isolate an object, so the scientist must sometimes experiment with a pair of objects in the environment. You will be asked to help advise the scientist about various experiments that might be performed. The specific tasks will differ slightly throughout the session and you will be given specific instructions for each question.

Before we proceed to the questions, let us consider an example situation that will help you to understand more about the task. Notice in the figure below there are two objects, a *black square* and a *white disc*. The particles are emitted from the arrow at the top of the figure and will move straight down towards the objects. The particle may be stopped, or it may pass through the object or objects. Because the object stops before it reaches either object, when you have a pair of objects that stop the particle, you cannot determine which object is responsible.



After conducting some preliminary research, the scientist believes that particles are stopped by *triangular objects regardless of their color*. We ask that you adopt this hypothesis as you answer questions about experiments.

For example, you might be asked "Will the particle to be stopped by the objects in this experiment?" You should answer "No" because there is no triangle in this experiment, and by our hypothesis, only triangles (of any color) will block particles.

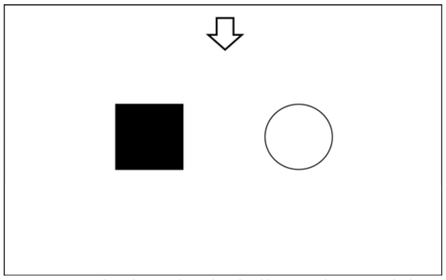
For some questions you will be asked to select the most effective experiment. As a scientist, the goal of experimental testing is to prove the hypothesis. We want to give the best possible advice to the scientist about each question, so please carefully consider the question and your answer before moving on to the next one. Do not worry about answering quickly: you will have plenty of time to complete the experiment in the allotted time. If you have any questions before beginning the experiment, please raise your hand and the experimenter will come over and help you. Otherwise you may begin the experiment.

Disconfirm Instruction Condition

A scientist has noticed that certain objects will stop a particle that has been fired in the direction of the object, while other objects allow the particle to pass by. The scientist is trying to understand which objects block particles and which don't. The objects vary in both shape and color. Possible shapes include *triangles, squares,* and *disks,* while the colors are *black* and *white*. Thus, there are a total of six different objects the scientist must consider.

We would like you to consider certain experiments that the scientist might perform to help determine what characteristic or characteristics of the object stop particles. Unfortunately, it is not always possible to isolate an object, so the scientist must sometimes experiment with a pair of objects in the environment. You will be asked to help advise the scientist about various experiments that might be performed. The specific tasks will differ slightly throughout the session and you will be given specific instructions for each question.

Before we proceed to the questions, let us consider an example situation that will help you to understand more about the task. Notice in the figure below there are two objects, a *black square* and a *white disc*. The particles are emitted from the arrow at the top of the figure and will move straight down towards the objects. The particle may be stopped, or it may pass through the object or objects. Because the object stops before it reaches either object, when you have a pair of objects that stop the particle, you cannot determine which object is responsible.



After conducting some preliminary research, the scientist believes that particles are stopped by *triangular objects regardless of their color*. We ask that you adopt this hypothesis as you answer questions about experiments.

For example, you might be asked "Will the particle to be stopped by the objects in this experiment?" You should answer "No" because there is no triangle in this experiment, and by our hypothesis, only triangles (of any color) will block particles.

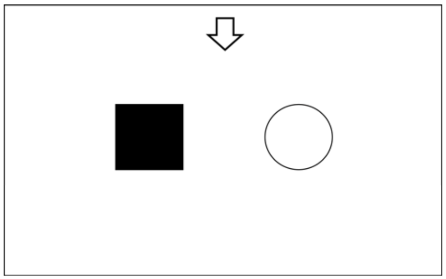
For some questions you will be asked to select the most effective experiment. As a scientist, the goal of experimental testing is to falsify the hypothesis. We want to give the best possible advice to the scientist about each question, so please carefully consider the question and your answer before moving on to the next one. Do not worry about answering quickly: you will have plenty of time to complete the experiment in the allotted time. If you have any questions before beginning the experiment, please raise your hand and the experimenter will come over and help you. Otherwise you may begin the experiment.

High Miss Cost Instruction Condition

A scientist has noticed that certain objects will stop a particle that has been fired in the direction of the object, while other objects allow the particle to pass by. The scientist is trying to understand which objects block particles and which don't. If the scientist can determine what is stopping the particles, he can use this knowledge to develop a new treatment for cancer. The objects vary in both shape and color. Possible shapes include *triangles, squares,* and *disks,* while the colors are *black* and *white*. Thus, there are a total of six different objects the scientist must consider.

We would like you to consider certain experiments that the scientist might perform to help determine what characteristic or characteristics of the object stop particles. Unfortunately, it is not always possible to isolate an object, so the scientist must sometimes experiment with a pair of objects in the environment. You will be asked to help advise the scientist about various experiments that might be performed. The specific tasks will differ slightly throughout the session and you will be given specific instructions for each question.

Before we proceed to the questions, let us consider an example situation that will help you to understand more about the task. Notice in the figure below there are two objects, a *black square* and a *white disc*. The particles are emitted from the arrow at the top of the figure and will move straight down towards the objects. The particle may be stopped, or it may pass through the object or objects. Because the object stops before it reaches either object, when you have a pair of objects that stop the particle, you cannot determine which object is responsible.



After conducting some preliminary research, the scientist believes that particles are stopped by *triangular objects regardless of their color*. We ask that you adopt this hypothesis as you answer questions about experiments.

For example, you might be asked "Will the particle to be stopped by the objects in this experiment?" You should answer "No" because there is no triangle in this experiment, and by our hypothesis, only triangles (of any color) will block particles.

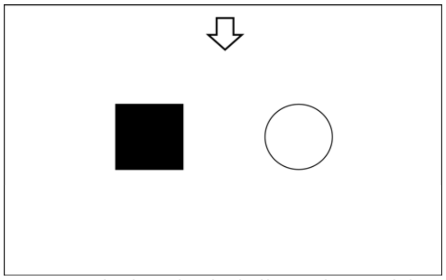
We want to give the best possible advice to the scientist about each question, so please carefully consider the question and your answer before moving on to the next one. Do not worry about answering quickly: you will have plenty of time to complete the experiment in the allotted time. If you have any questions before beginning the experiment, please raise your hand and the experimenter will come over and help you. Otherwise you may begin the experiment.

Common Event Instruction Condition

A scientist has noticed that certain objects will stop a particle that has been fired in the direction of the object, while other objects allow the particle to pass by. The scientist is trying to understand which objects block particles and which don't. The objects vary in both shape and color. Possible shapes include *triangles, squares,* and *disks,* while the colors are *black* and *white*. Thus, there are a total of six different objects the scientist must consider.

We would like you to consider certain experiments that the scientist might perform to help determine what characteristic or characteristics of the object stop particles. Unfortunately, it is not always possible to isolate an object, so the scientist must sometimes experiment with a pair of objects in the environment. You will be asked to help advise the scientist about various experiments that might be performed. The specific tasks will differ slightly throughout the session and you will be given specific instructions for each question.

Before we proceed to the questions, let us consider an example situation that will help you to understand more about the task. Notice in the figure below there are two objects, a *black square* and a *white disc*. The particles are emitted from the arrow at the top of the figure and will move straight down towards the objects. The particle may be stopped, or it may pass through the object or objects. Because the object stops before it reaches either object, when you have a pair of objects that stop the particle, you cannot determine which object is responsible.



After conducting some preliminary research, the scientist believes that particles are stopped by *triangular objects regardless of their color*. The particles were regularly stopped during the preliminary research. We ask that you adopt this hypothesis as you answer questions about experiments.

For example, you might be asked "Will the particle to be stopped by the objects in this experiment?" You should answer "No" because there is no triangle in this experiment, and by our hypothesis, only triangles (of any color) will block particles.

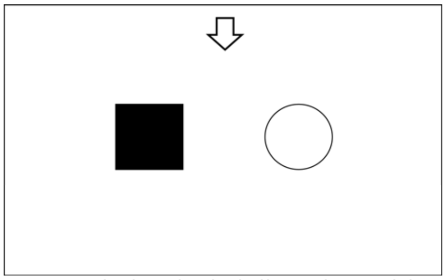
We want to give the best possible advice to the scientist about each question, so please carefully consider the question and your answer before moving on to the next one. Do not worry about answering quickly: you will have plenty of time to complete the experiment in the allotted time. If you have any questions before beginning the experiment, please raise your hand and the experimenter will come over and help you. Otherwise you may begin the experiment.

Alternative Hypothesis Instruction Condition

A scientist has noticed that certain objects will stop a particle that has been fired in the direction of the object, while other objects allow the particle to pass by. The scientist is trying to understand which objects block particles and which don't. The objects vary in both shape and color. Possible shapes include *triangles, squares,* and *disks,* while the colors are *black* and *white*. Thus, there are a total of six different objects the scientist must consider.

We would like you to consider certain experiments that the scientist might perform to help determine what characteristic or characteristics of the object stop particles. Unfortunately, it is not always possible to isolate an object, so the scientist must sometimes experiment with a pair of objects in the environment. You will be asked to help advise the scientist about various experiments that might be performed. The specific tasks will differ slightly throughout the session and you will be given specific instructions for each question.

Before we proceed to the questions, let us consider an example situation that will help you to understand more about the task. Notice in the figure below there are two objects, a *black square* and a *white disc*. The particles are emitted from the arrow at the top of the figure and will move straight down towards the objects. The particle may be stopped, or it may pass through the object or objects. Because the object stops before it reaches either object, when you have a pair of objects that stop the particle, you cannot determine which object is responsible.



After conducting some preliminary research, the scientist believes that particles are stopped by *triangular objects regardless of their color*. Another researcher working with the same particles believes that the particles are stopped by *black objects regardless of their shape*. We ask that you consider this information as you answer questions about experiments.

For example, you might be asked "Will the particle to be stopped by the objects in this experiment?" You should answer "No" because there is no triangle in this experiment, and by our hypothesis, only triangles (of any color) will block particles.

We want to give the best possible advice to the scientist about each question, so please carefully consider the question and your answer before moving on to the next one. Do not worry about answering quickly: you will have plenty of time to complete the experiment in the allotted time. If you have any questions before beginning the experiment, please raise your hand and the experimenter will come over and help you. Otherwise you may begin the experiment.

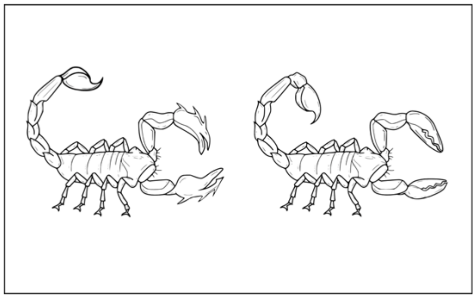
Scorpion – Spider Scenario

Test Instruction Condition

Farmers recently introduced 6 types of scorpions to their farmland in an attempt to deal with a pest problem. A scientist has noticed that the scorpions are attacking and consuming a particular species of spider that is commonly found on the farmland. The scientist wants to understand what determines which types of scorpion eat the spiders and which don't. The scorpions can be identified by the shape of their tail and the shape of their claws. The scorpion tail's can be curled *up* or curled *down* while the claws' can be *serrated, spiky,* or *round* claws. There are a total of six different scorpions that the scientist must consider.

We would like you to consider certain experiments that the scientist might perform to help understand what features of the scorpions determine whether or not they will eat the spiders. Unfortunately, it is not always possible to isolate one type of scorpion, so the scientist must sometimes experiment with a pair of scorpions in the environment. You will be asked to help advise the scientist about various experiments that might be performed. The specific tasks will differ slightly throughout the session and you will be given specific instructions for each question.

Before we proceed to the questions, let us consider an example situation that will help you to understand more about the task. In the figure below, there are two scorpions: one with *spiky* claws and a tail that curls *up;* another with *serrated* claws and a tail that curls *down.* In an experiment, a spider will be introduced to the enclosed environment for a week or until the spider is eaten and consumed. The spider may be eaten by the scorpions or it may be ignored by the scorpions. Continuous observation of the scorpions is not possible, so when you have two scorpions in the environment and the spider is eaten, you cannot determine which of the scorpions consumed the spider.



Based on some preliminary field studies, the scientist believes that the spiders are being attacked and consumed by *scorpions with round claws regardless of their tail shape.* We ask that you adopt this hypothesis as you answer questions about experiments.

For example, you might be asked "Will the spider be eaten in this experiment?" You should answer "No" because none of the scorpions in this experiment have round claws, and by our hypothesis, scorpions must have round claws (and any tail shape) to attack and consume the spiders.

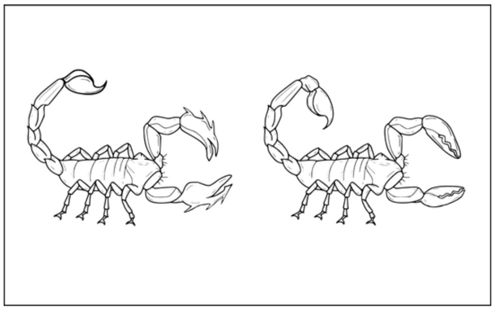
For some questions you will be asked to select the most effective experiment. As a scientist, the goal of experimental testing is to test the current hypothesis. We want to give the best possible advice to the scientist about each question, so please carefully consider the question and your answer before moving on to the next one. Do not worry about answering quickly: you will have plenty of time to complete the experiment in the allotted time. If you have any questions before beginning the experiment, please raise your hand and the experimenter will come over and help you. Otherwise you may begin the experiment.

Confirm Instruction Condition

Farmers recently introduced 6 types of scorpions to their farmland in an attempt to deal with a pest problem. A scientist has noticed that the scorpions are attacking and consuming a particular species of spider that is commonly found on the farmland. The scientist wants to understand what determines which types of scorpion eat the spiders and which don't. The scorpions can be identified by the shape of their tail and the shape of their claws. The scorpion tail's can be curled *up* or curled *down* while the claws' can be *serrated, spiky,* or *round* claws. There are a total of six different scorpions that the scientist must consider.

We would like you to consider certain experiments that the scientist might perform to help understand what features of the scorpions determine whether or not they will eat the spiders. Unfortunately, it is not always possible to isolate one type of scorpion, so the scientist must sometimes experiment with a pair of scorpions in the environment. You will be asked to help advise the scientist about various experiments that might be performed. The specific tasks will differ slightly throughout the session and you will be given specific instructions for each question.

Before we proceed to the questions, let us consider an example situation that will help you to understand more about the task. In the figure below, there are two scorpions: one with *spiky* claws and a tail that curls *up;* another with *serrated* claws and a tail that curls *down.* In an experiment, a spider will be introduced to the enclosed environment for a week or until the spider is eaten and consumed. The spider may be eaten by the scorpions or it may be ignored by the scorpions. Continuous observation of the scorpions is not possible, so when you have two scorpions in the environment and the spider is eaten, you cannot determine which of the scorpions consumed the spider.



Based on some preliminary field studies, the scientist believes that the spiders are being attacked and consumed by *scorpions with round claws regardless of their tail shape.* We ask that you adopt this hypothesis as you answer questions about experiments.

For example, you might be asked "Will the spider be eaten in this experiment?" You should answer "No" because none of the scorpions in this experiment have round claws, and by our hypothesis, scorpions must have round claws (and any tail shape) to attack and consume the spiders.

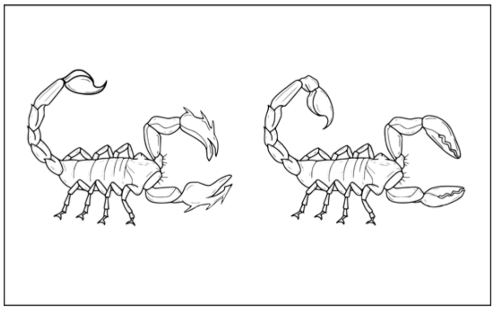
For some questions you will be asked to select the most effective experiment. As a scientist, the goal of experimental testing is to prove the current hypothesis. We want to give the best possible advice to the scientist about each question, so please carefully consider the question and your answer before moving on to the next one. Do not worry about answering quickly: you will have plenty of time to complete the experiment in the allotted time. If you have any questions before beginning the experiment, please raise your hand and the experimenter will come over and help you. Otherwise you may begin the experiment.

Disconfirm Instruction Condition

Farmers recently introduced 6 types of scorpions to their farmland in an attempt to deal with a pest problem. A scientist has noticed that the scorpions are attacking and consuming a particular species of spider that is commonly found on the farmland. The scientist wants to understand what determines which types of scorpion eat the spiders and which don't. The scorpions can be identified by the shape of their tail and the shape of their claws. The scorpion tail's can be curled *up* or curled *down* while the claws' can be *serrated, spiky,* or *round* claws. There are a total of six different scorpions that the scientist must consider.

We would like you to consider certain experiments that the scientist might perform to help understand what features of the scorpions determine whether or not they will eat the spiders. Unfortunately, it is not always possible to isolate one type of scorpion, so the scientist must sometimes experiment with a pair of scorpions in the environment. You will be asked to help advise the scientist about various experiments that might be performed. The specific tasks will differ slightly throughout the session and you will be given specific instructions for each question.

Before we proceed to the questions, let us consider an example situation that will help you to understand more about the task. In the figure below, there are two scorpions: one with *spiky* claws and a tail that curls *up;* another with *serrated* claws and a tail that curls *down.* In an experiment, a spider will be introduced to the enclosed environment for a week or until the spider is eaten and consumed. The spider may be eaten by the scorpions or it may be ignored by the scorpions. Continuous observation of the scorpions is not possible, so when you have two scorpions in the environment and the spider is eaten, you cannot determine which of the scorpions consumed the spider.



Based on some preliminary field studies, the scientist believes that the spiders are being attacked and consumed by *scorpions with round claws regardless of their tail shape.* We ask that you adopt this hypothesis as you answer questions about experiments.

For example, you might be asked "Will the spider be eaten in this experiment?" You should answer "No" because none of the scorpions in this experiment have round claws, and by our hypothesis, scorpions must have round claws (and any tail shape) to attack and consume the spiders.

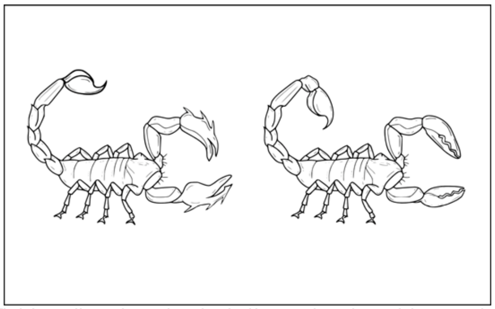
For some questions you will be asked to select the most effective experiment. As a scientist, the goal of experimental testing is to falsify the current hypothesis. We want to give the best possible advice to the scientist about each question, so please carefully consider the question and your answer before moving on to the next one. Do not worry about answering quickly: you will have plenty of time to complete the experiment in the allotted time. If you have any questions before beginning the experiment, please raise your hand and the experimenter will come over and help you. Otherwise you may begin the experiment.

High Miss Cost Instruction Condition

Farmers recently introduced 6 types of scorpions to their farmland in an attempt to deal with a pest problem. A scientist has noticed that the scorpions are attacking and consuming a particular species of spider that is commonly found on the farmland. The spiders are also beneficial to pest control. The scientist must discover which scorpions are eating the spiders to avoid significant loss for the farmers. The scientist wants to understand what determines which types of scorpion eat the spiders and which don't. The scorpions can be identified by the shape of their tail and the shape of their claws. The scorpion tail's can be curled *up* or curled *down* while the claws' can be *serrated, spiky,* or *round* claws. There are a total of six different scorpions that the scientist must consider.

We would like you to consider certain experiments that the scientist might perform to help understand what features of the scorpions determine whether or not they will eat the spiders. Unfortunately, it is not always possible to isolate one type of scorpion, so the scientist must sometimes experiment with a pair of scorpions in the environment. You will be asked to help advise the scientist about various experiments that might be performed. The specific tasks will differ slightly throughout the session and you will be given specific instructions for each question.

Before we proceed to the questions, let us consider an example situation that will help you to understand more about the task. In the figure below, there are two scorpions: one with *spiky* claws and a tail that curls *up;* another with *serrated* claws and a tail that curls *down.* In an experiment, a spider will be introduced to the enclosed environment for a week or until the spider is eaten and consumed. The spider may be eaten by the scorpions or it may be ignored by the scorpions. Continuous observation of the scorpions is not possible, so when you have two scorpions in the environment and the spider is eaten, you cannot determine which of the scorpions consumed the spider.



Based on some preliminary field studies, the scientist believes that the spiders are being attacked and consumed by *scorpions with round claws regardless of their tail shape.* We ask that you adopt this hypothesis as you answer questions about experiments.

For example, you might be asked "Will the spider be eaten in this experiment?" You should answer "No" because none of the scorpions in this experiment have round claws, and by our hypothesis, scorpions must have round claws (and any tail shape) to attack and consume the spiders.

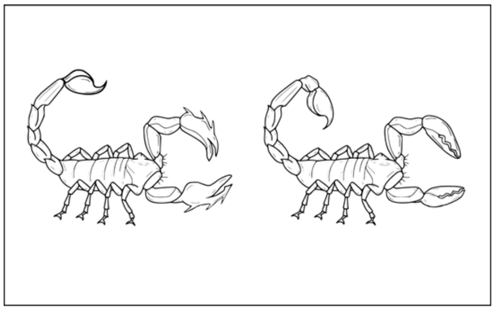
We want to give the best possible advice to the scientist about each question, so please carefully consider the question and your answer before moving on to the next one. Do not worry about answering quickly: you will have plenty of time to complete the experiment in the allotted time. If you have any questions before beginning the experiment, please raise your hand and the experimenter will come over and help you. Otherwise you may begin the experiment.

Common Event Instruction Condition

Farmers recently introduced 6 types of scorpions to their farmland in an attempt to deal with a pest problem. A scientist has noticed that the scorpions are attacking and consuming a particular species of spider that is commonly found on the farmland. The scientist wants to understand what determines which types of scorpion eat the spiders and which don't. The scorpions can be identified by the shape of their tail and the shape of their claws. The scorpion tail's can be curled *up* or curled *down* while the claws' can be *serrated, spiky,* or *round* claws. There are a total of six different scorpions the scientist must consider.

We would like you to consider certain experiments that the scientist might perform to help understand what features of the scorpions determine whether or not they will eat the spiders. Unfortunately, it is not always possible to isolate one type of scorpion, so the scientist must sometimes experiment with a pair of scorpions in the environment. You will be asked to help advise the scientist about various experiments that might be performed. The specific tasks will differ slightly throughout the session and you will be given specific instructions for each question.

Before we proceed to the questions, let us consider an example situation that will help you to understand more about the task. In the figure below, there are two scorpions: one with *spiky* claws and a tail that curls *up;* another with *serrated* claws and a tail that curls *down.* In an experiment, a spider will be introduced to the enclosed environment for a week or until the spider is eaten and consumed. The spider may be eaten by the scorpions or it may be ignored by the scorpions. Continuous observation of the scorpions is not possible, so when you have two scorpions in the environment and the spider is eaten, you cannot determine which of the scorpions consumed the spider.



Based on some preliminary field studies, the scientist believes that the spiders are being regularly attacked and consumed by *scorpions with round claws regardless of their tail shape.* The farmers have noticed that spiders seldom survive once the scorpions are introduced to their fields. We ask that you adopt this hypothesis as you answer questions about experiments.

For example, you might be asked "Will the spider be eaten in this experiment?" You should answer "No" because none of the scorpions in this experiment have round claws, and by our hypothesis, scorpions must have round claws (and any tail shape) to attack and consume the spiders.

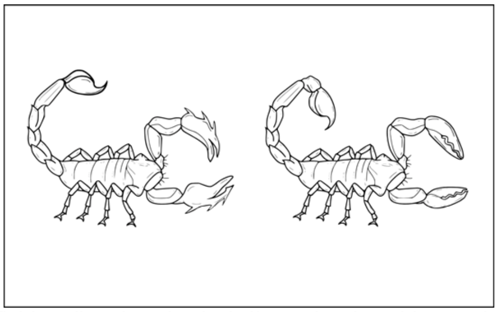
We want to give the best possible advice to the scientist about each question, so please carefully consider the question and your answer before moving on to the next one. Do not worry about answering quickly: you will have plenty of time to complete the experiment in the allotted time. If you have any questions before beginning the experiment, please raise your hand and the experimenter will come over and help you. Otherwise you may begin the experiment.

Alternative Hypothesis Instruction Condition

Farmers recently introduced 6 types of scorpions to their farmland in an attempt to deal with a pest problem. A scientist has noticed that the scorpions are attacking and consuming a particular species of spider that is commonly found on the farmland. The scientist wants to understand what determines which types of scorpion eat the spiders and which don't. The scorpions can be identified by the shape of their tail and the shape of their claws. The scorpion tail's can be curled *up* or curled *down* while the claws' can be *serrated, spiky,* or *round* claws. There are a total of six different scorpions the scientist must consider.

We would like you to consider certain experiments that the scientist might perform to help understand what features of the scorpions determine whether or not they will eat the spiders. Unfortunately, it is not always possible to isolate one type of scorpion, so the scientist must sometimes experiment with a pair of scorpions in the environment. You will be asked to help advise the scientist about various experiments that might be performed. The specific tasks will differ slightly throughout the session and you will be given specific instructions for each question.

Before we proceed to the questions, let us consider an example situation that will help you to understand more about the task. In the figure below, there are two scorpions: one with *spiky* claws and a tail that curls *up;* another with *serrated* claws and a tail that curls *down.* In an experiment, a spider will be introduced to the enclosed environment for a week or until the spider is eaten and consumed. The spider may be eaten by the scorpions or it may be ignored by the scorpions. Continuous observation of the scorpions is not possible, so when you have two scorpions in the environment and the spider is eaten, you cannot determine which of the scorpions consumed the spider.

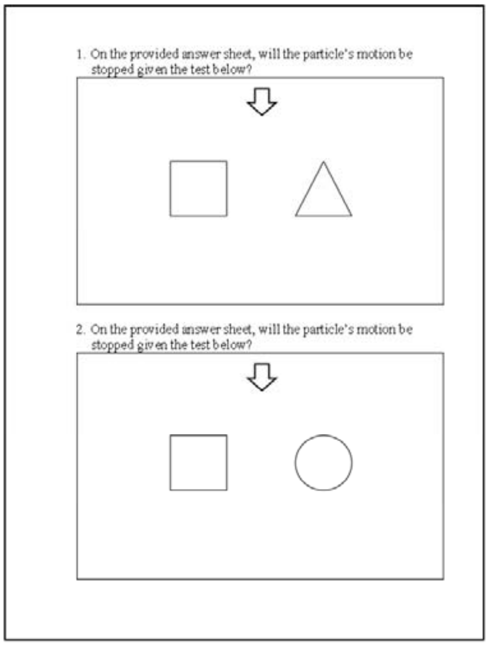


Based on some preliminary field studies, the scientist believes that the spiders are being attacked and consumed by *scorpions with round claws regardless of their tail shape.* Other scientists have suggested instead that the *scorpions with tails that curl down regardless of their claw shape* are responsible for the deaths of the spiders. We ask that you consider this information as you answer questions about experiments.

For example, you might be asked "Will the spider be eaten in this experiment?" You should answer "No" because none of the scorpions in this experiment have round claws, and by our hypothesis, scorpions must have round claws (and any tail shape) to attack and consume the spiders.

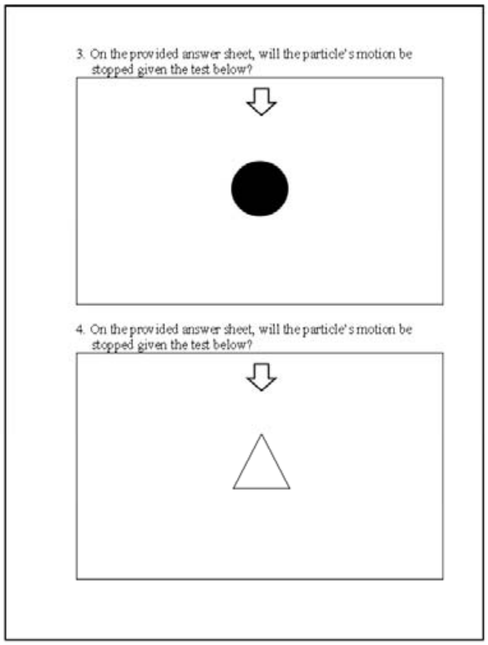
We want to give the best possible advice to the scientist about each question, so please carefully consider the question and your answer before moving on to the next one. Do not worry about answering quickly: you will have plenty of time to complete the experiment in the allotted time. If you have any questions before beginning the experiment, please raise your hand and the experimenter will come over and help you. Otherwise you may begin the experiment.

MDT Microworld Scenario



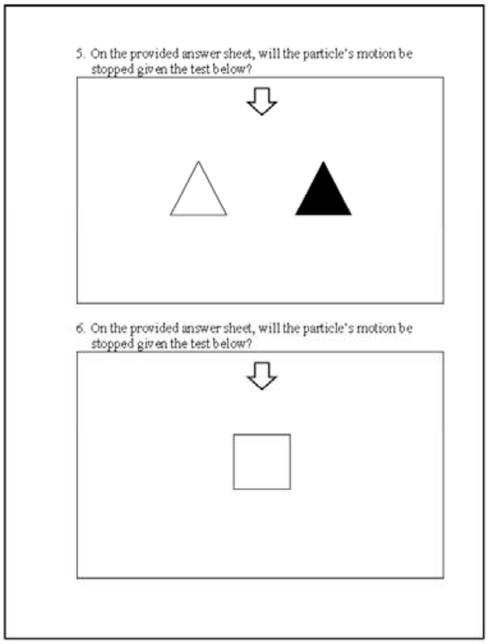
1. On the provided answer sheet, will the particle's motion be stopped given the test below?

2. On the provided answer sheet, will the particle's motion be stopped given the test below?



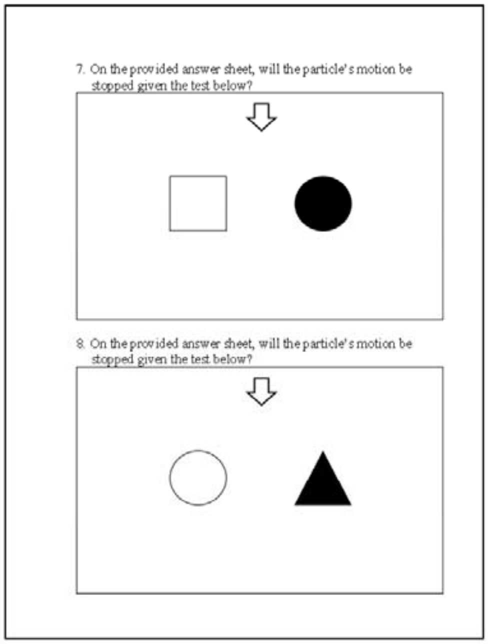
3. On the provided answer sheet, will the particle's motion be stopped given the test below?

4. On the provided answer sheet, will the particle's motion be stopped given the test below?



5. On the provided answer sheet, will the particle's motion be stopped given the test below?

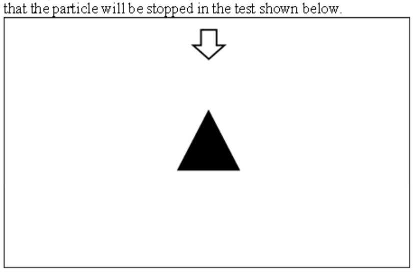
6. On the provided answer sheet, will the particle's motion be stopped given the test below?



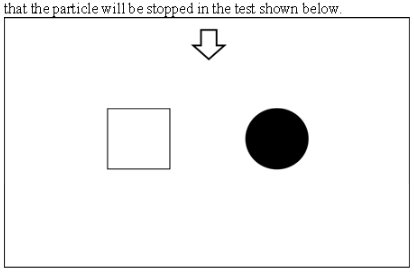
7. On the provided answer sheet, will the particle's motion be stopped given the test below?

8. On the provided answer sheet, will the particle's motion be stopped given the test below?

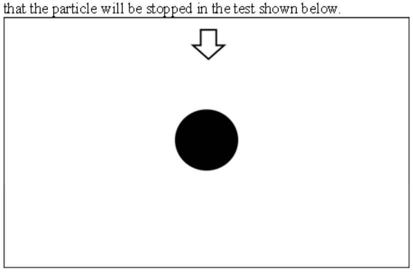
9. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.



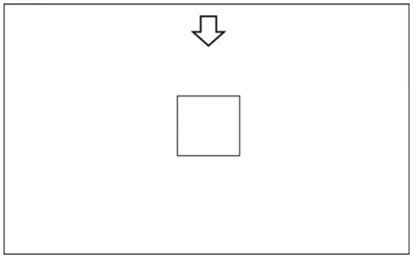
10. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.



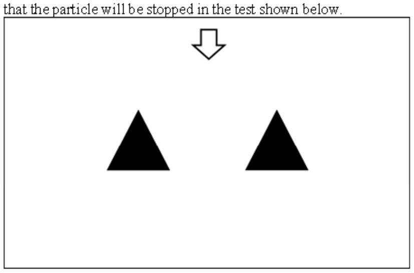
11. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.



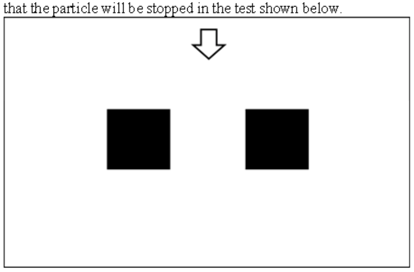
12. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.



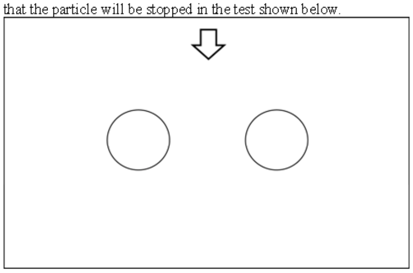
13. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.



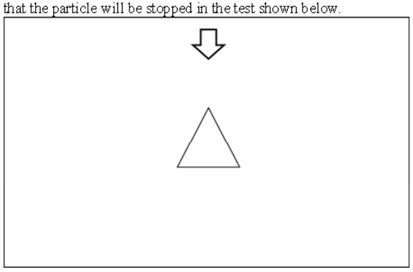
14. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.

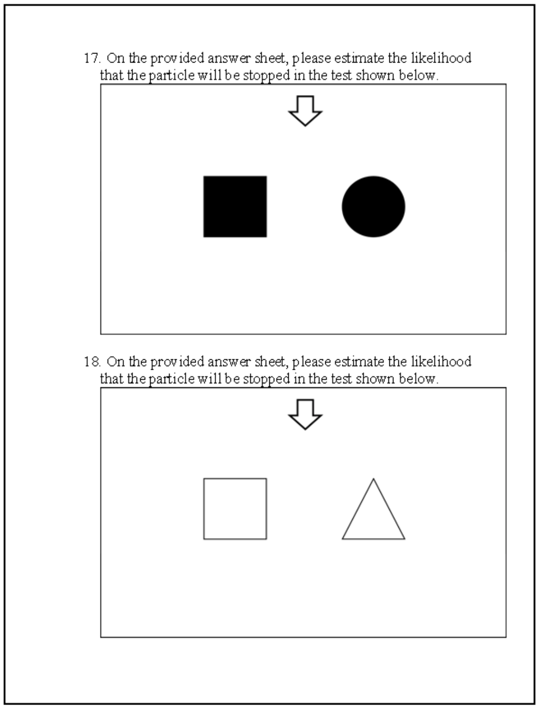


15. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.



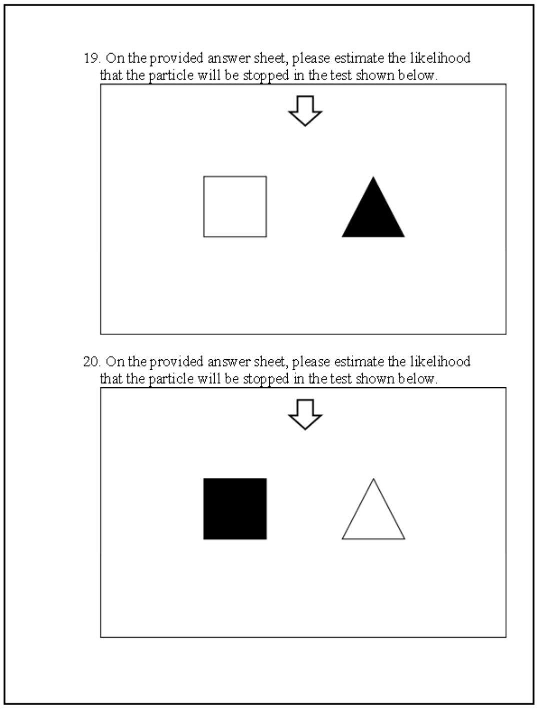
16. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.





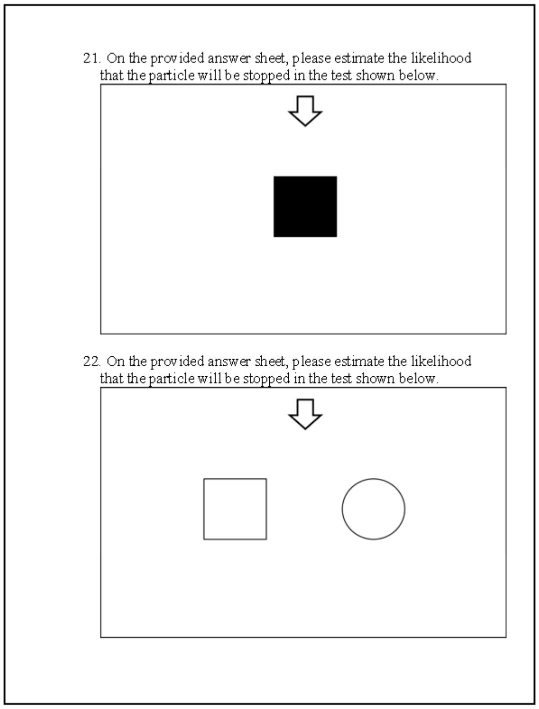
17. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.

18. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.



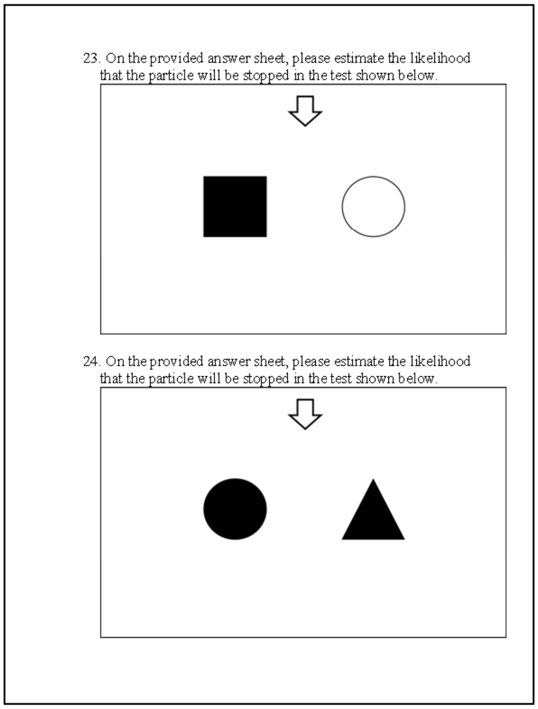
19. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.

20. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.



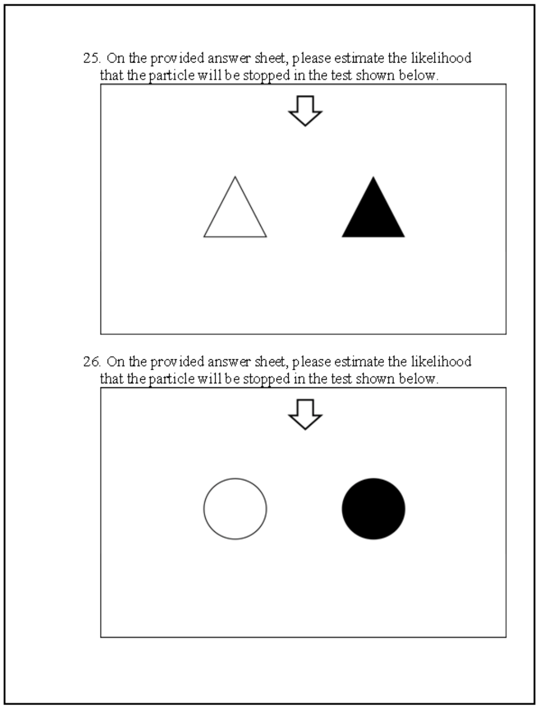
21. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.

22. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.



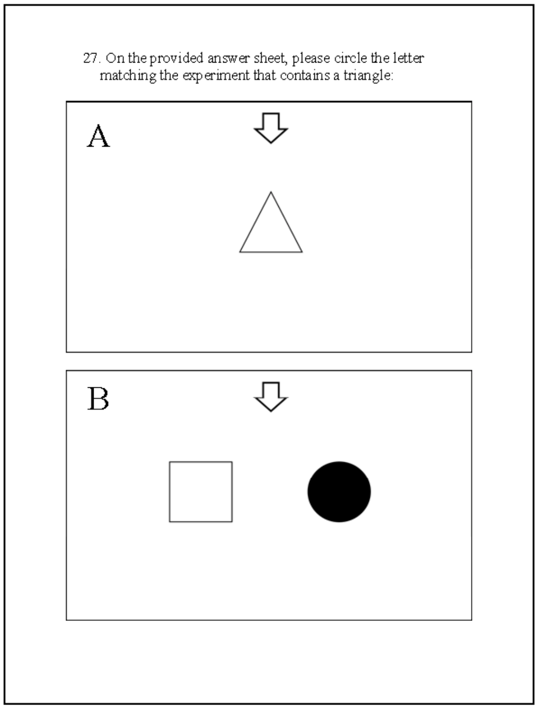
23. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.

24. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.

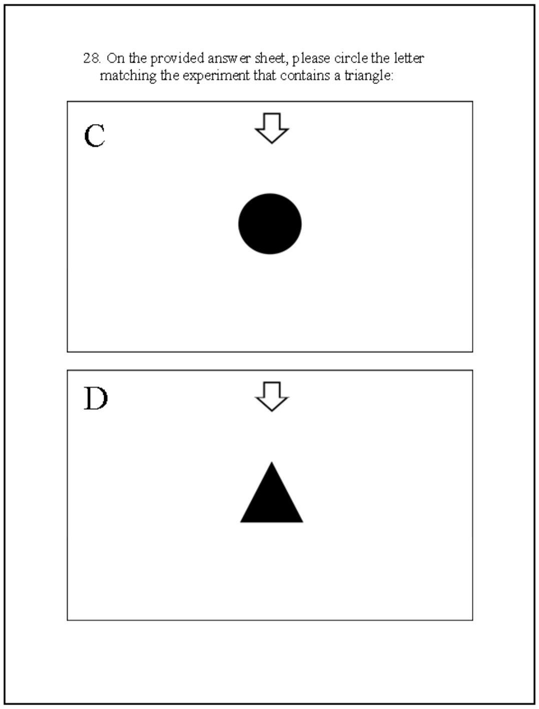


25. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.

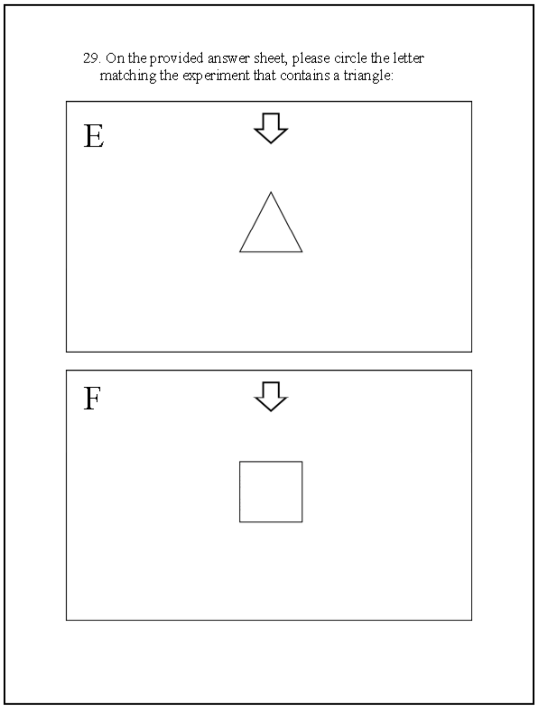
26. On the provided answer sheet, please estimate the likelihood that the particle will be stopped in the test shown below.



27. On the provided answer sheet, please circle the letter matching the experiment that contains a triangle:

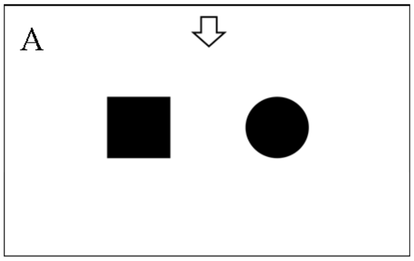


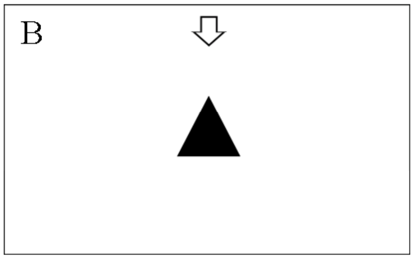
28. On the provided answer sheet, please circle the letter matching the experiment that contains a triangle:



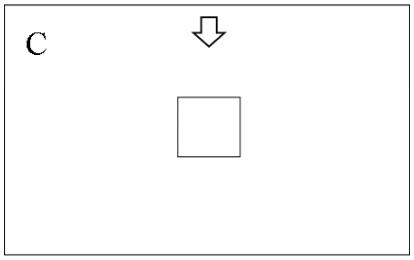
29. On the provided answer sheet, please circle the letter matching the experiment that contains a triangle:

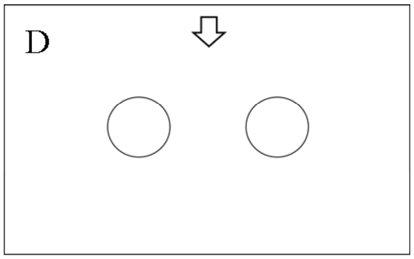
30. On the provided answer sheet, please circle the letter matching the experiment that contains a triangle:



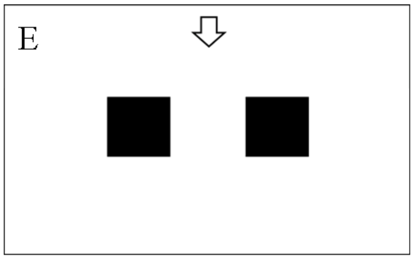


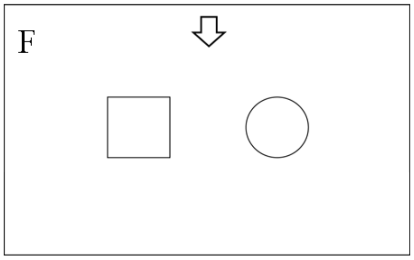
31. On the provided answer sheet, please circle the letter matching the experiment that is most likely to result in the particle's motion being stopped:



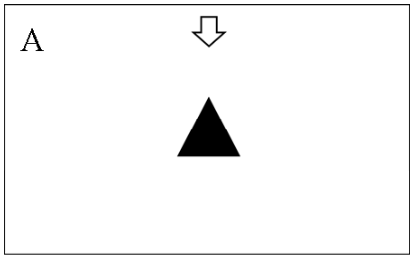


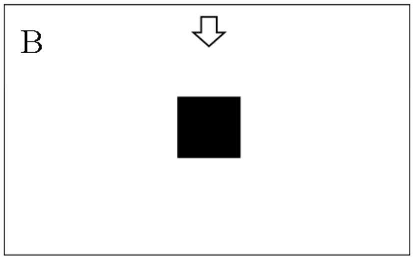
32. On the provided answer sheet, please circle the letter matching the experiment that is most likely to result in the particle's motion being stopped:



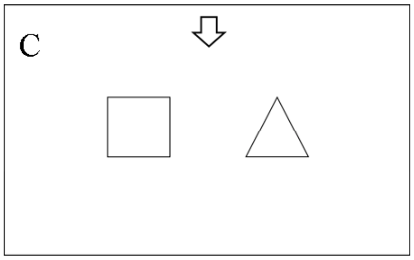


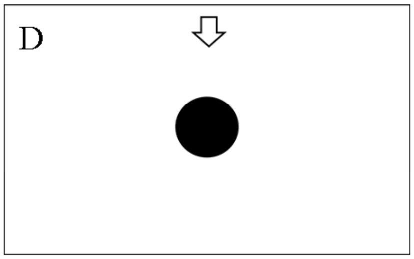
33. On the provided answer sheet, please circle the letter matching the experiment that is most likely to result in the particle's motion being stopped:



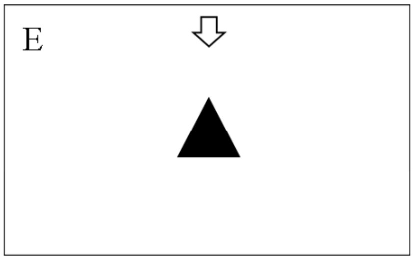


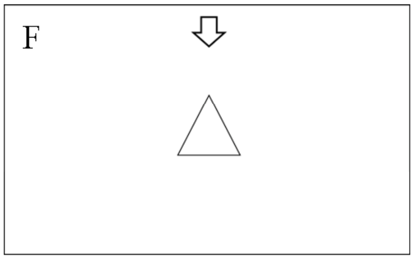
34. On the provided answer sheet, please circle the letter matching the experiment that is most likely to result in the particle's motion being stopped:



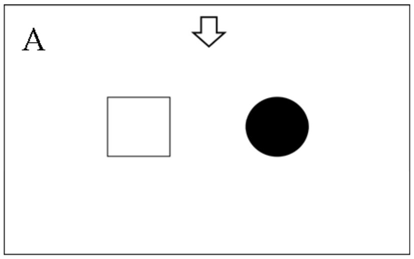


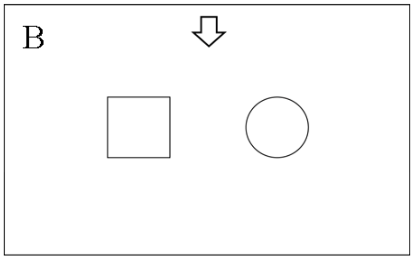
35. On the provided answer sheet, please circle the letter matching the experiment that is most likely to result in the particle's motion being stopped:



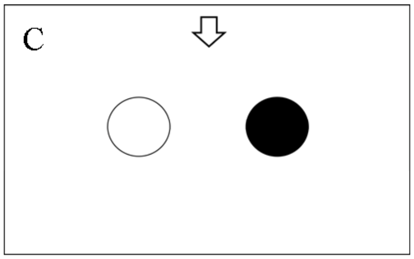


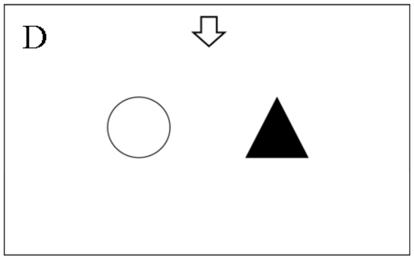
36. On the provided answer sheet, please circle the letter matching the experiment that is most likely to result in the particle's motion being stopped:



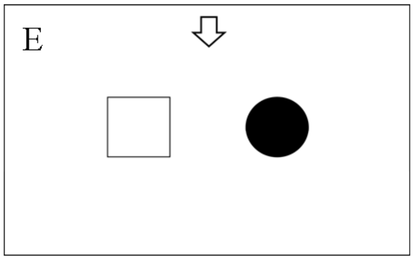


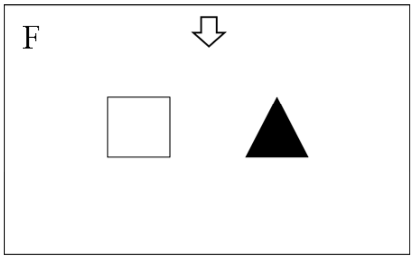
37. On the provided answer sheet, please circle the letter matching the experiment that is most likely to result in the particle's motion being stopped:



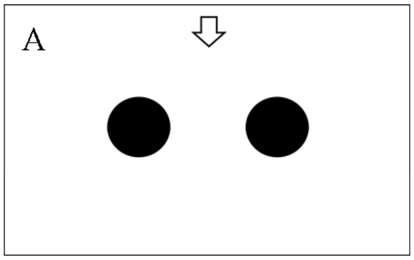


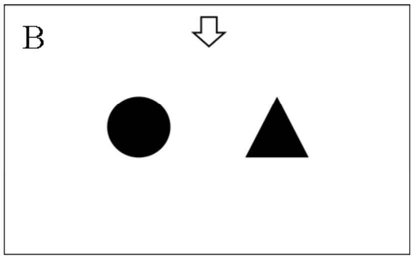
38. On the provided answer sheet, please circle the letter matching the experiment that is most likely to result in the particle's motion being stopped:



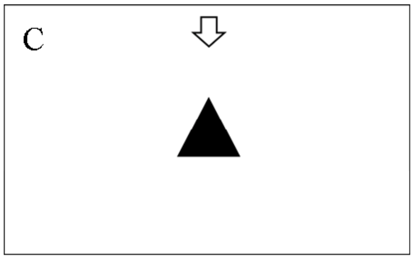


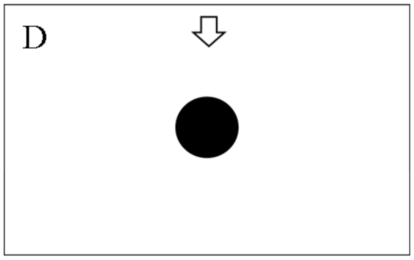
39. On the provided answer sheet, please circle the letter matching the experiment that is most likely to result in the particle's motion being stopped:



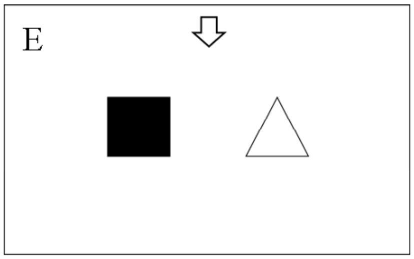


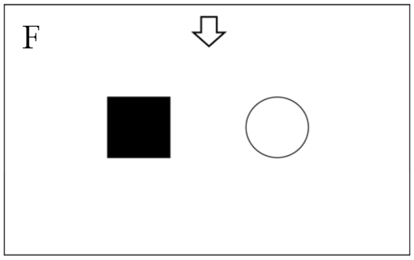
40. On the provided answer sheet, please circle the letter matching the experiment that is most likely to result in the particle's motion being stopped:



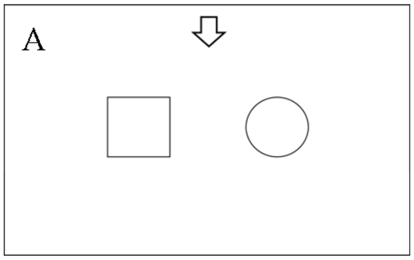


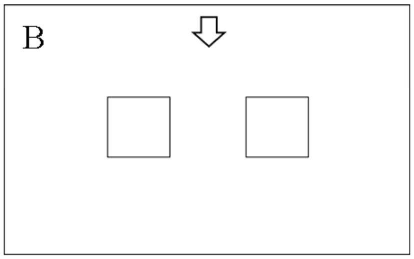
41. On the provided answer sheet, please circle the letter matching the experiment that is most likely to result in the particle's motion being stopped:



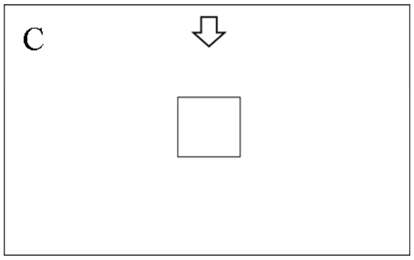


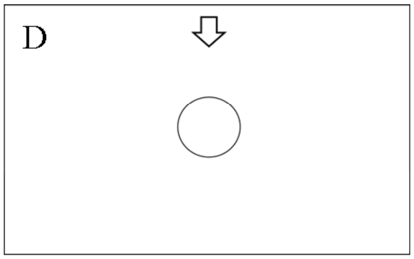
42. On the provided answer sheet, please circle the letter matching the experiment that is most likely to result in the particle's motion being stopped:



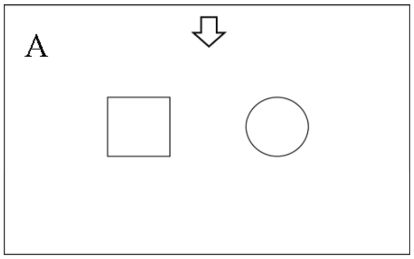


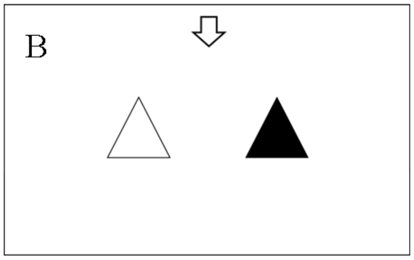
43. On the provided answer sheet, please circle the letter matching the experiment that is most likely to result in the particle's motion being stopped:



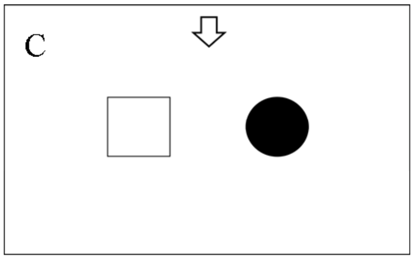


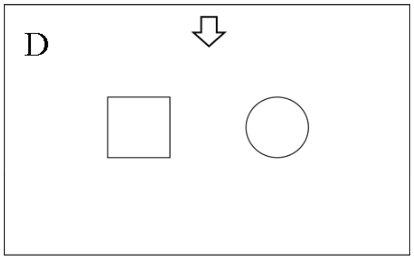
44. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:





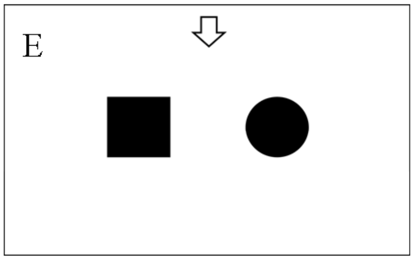
45. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:



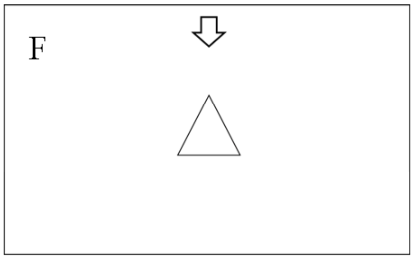


46. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:

E

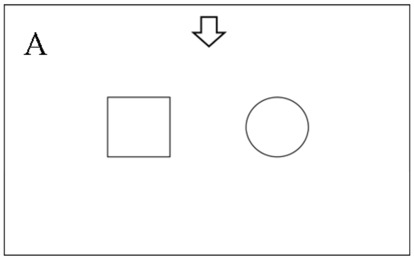


F

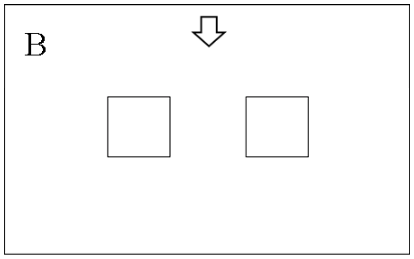


47. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:

A

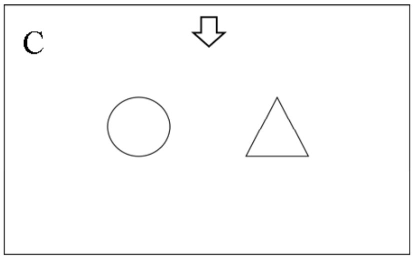


B

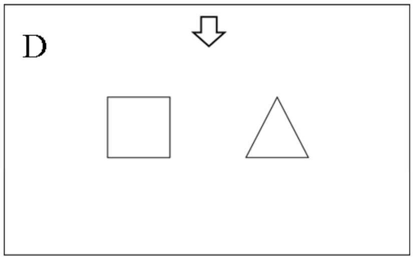


48. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:

C

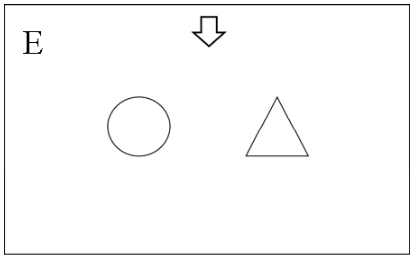


D

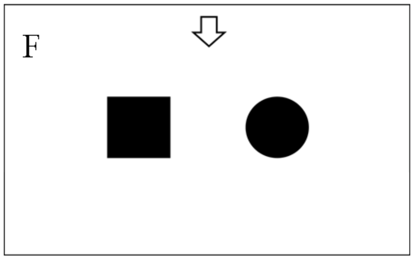


49. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:

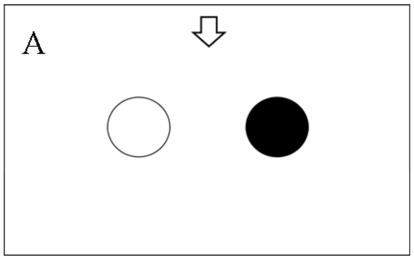
E

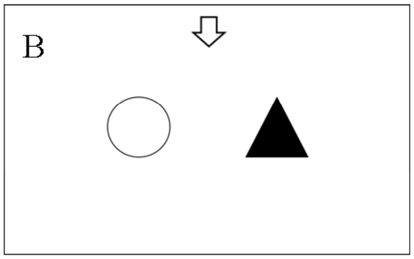


F

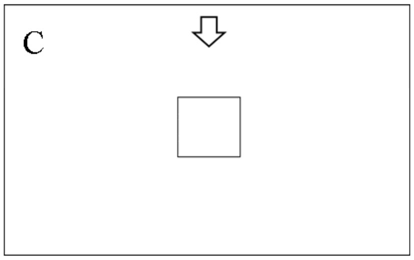


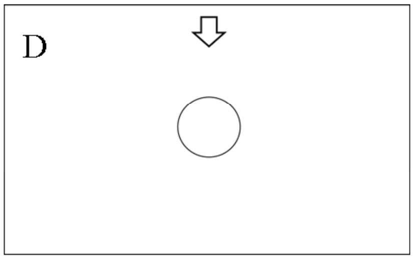
50. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:



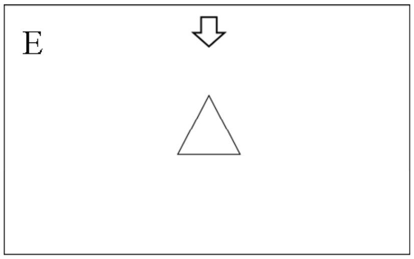


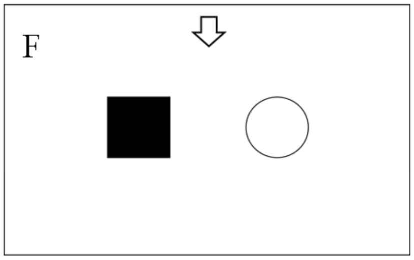
51. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:



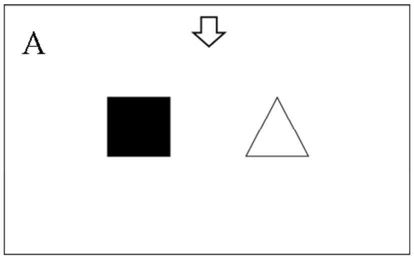


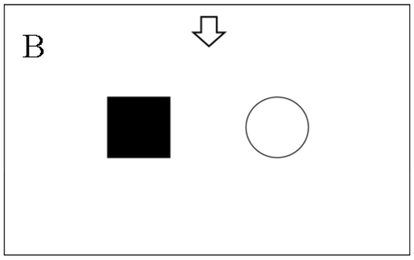
52. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:



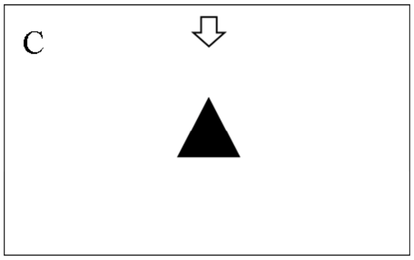


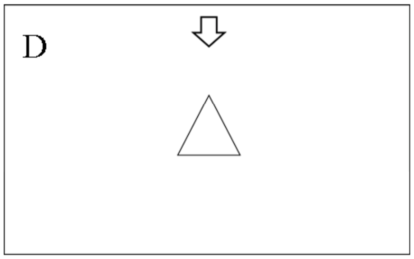
53. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:



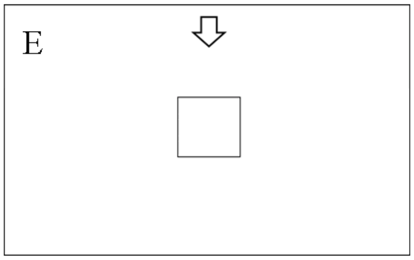


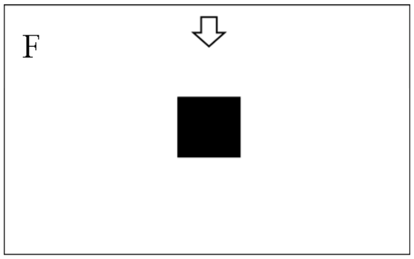
54. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:



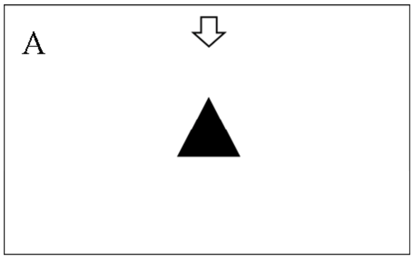


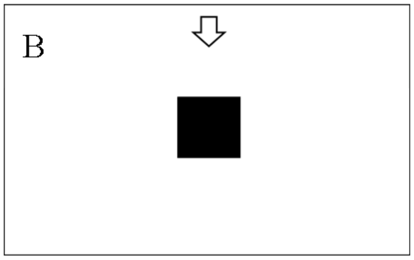
55. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:



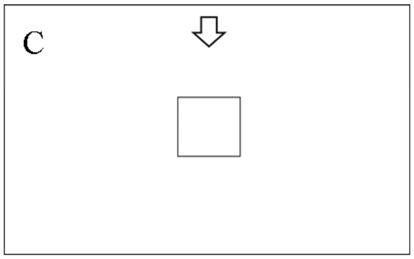


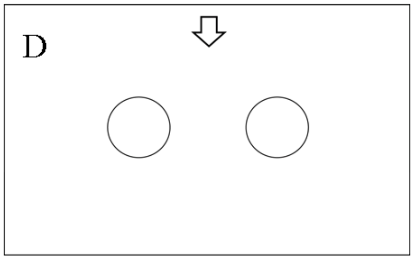
56. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:



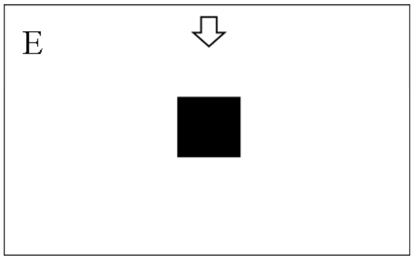


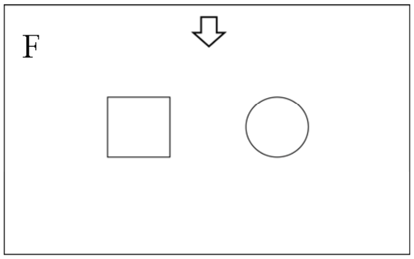
57. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:



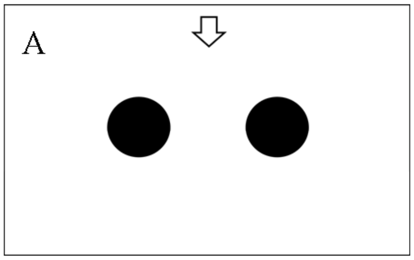


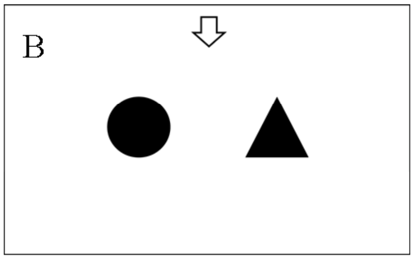
58. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:



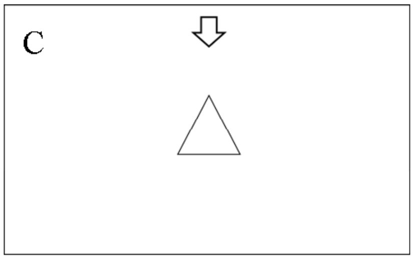


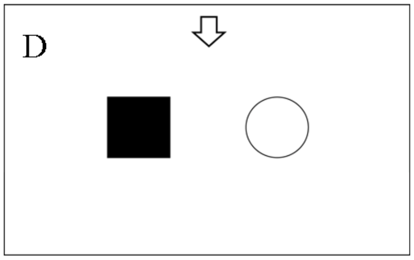
59. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:



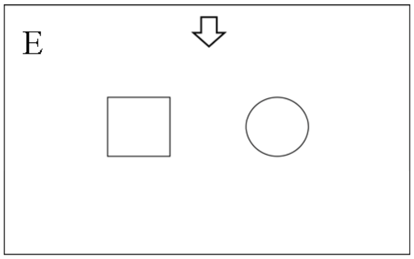


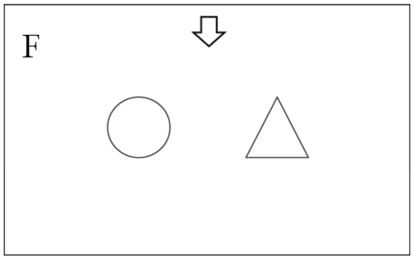
60. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:



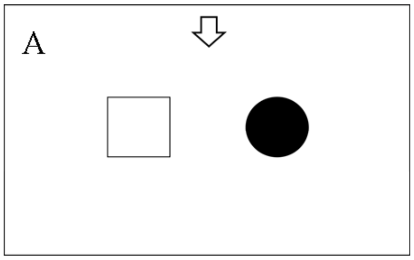


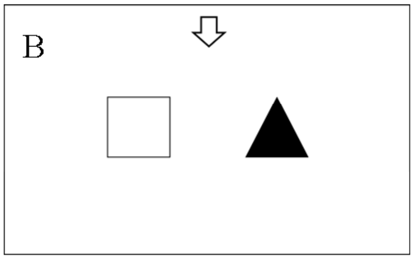
61. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:



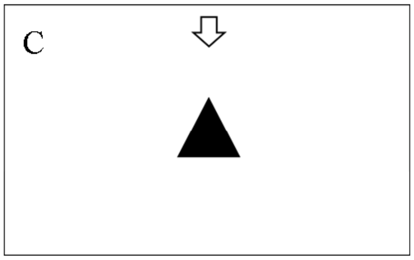


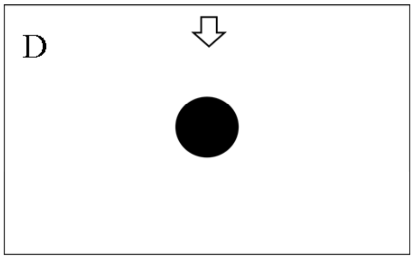
62. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:



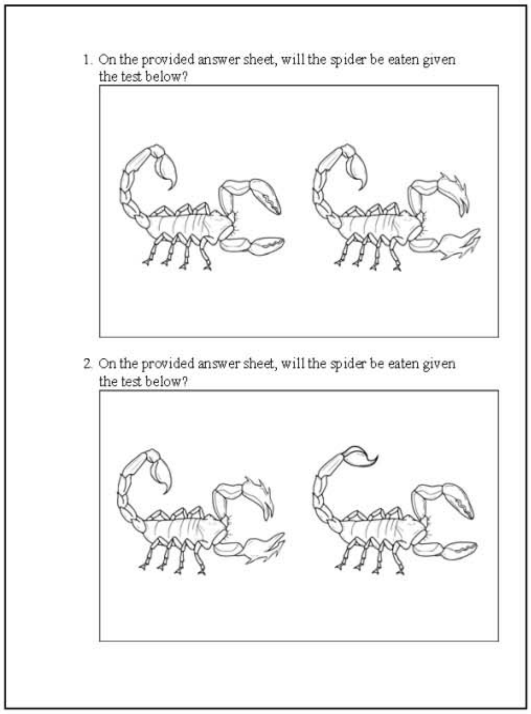


63. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:



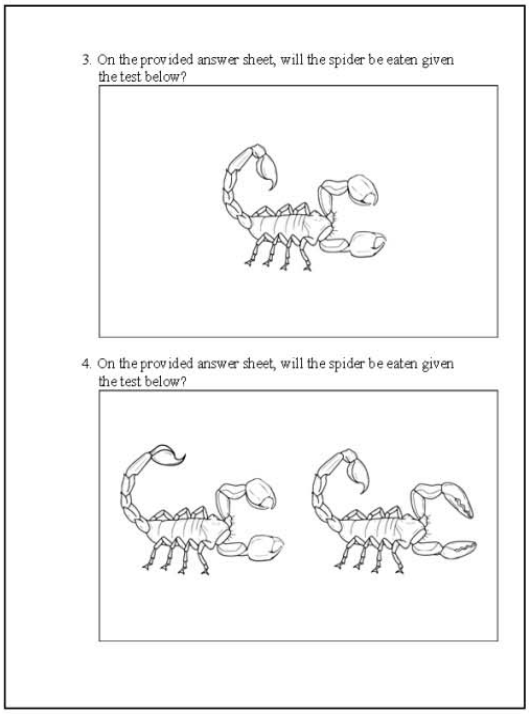


Scorpion – Spider Scenario



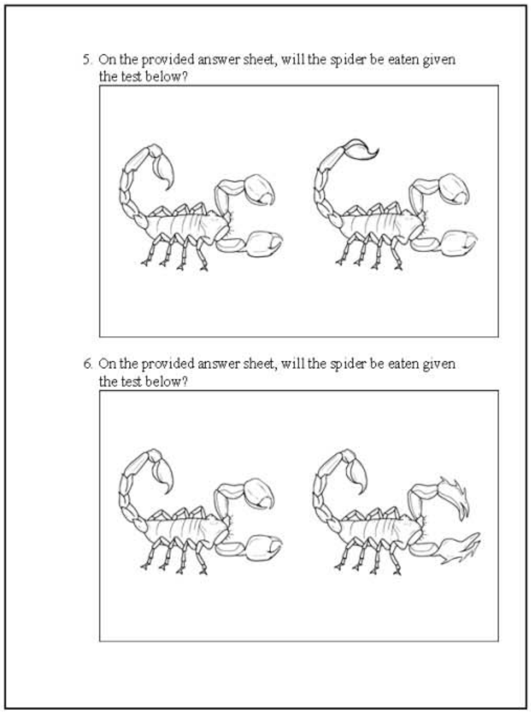
1. On the provided answer sheet, will the spider be eaten given the test below?

2. On the provided answer sheet, will the spider be eaten given the test below?



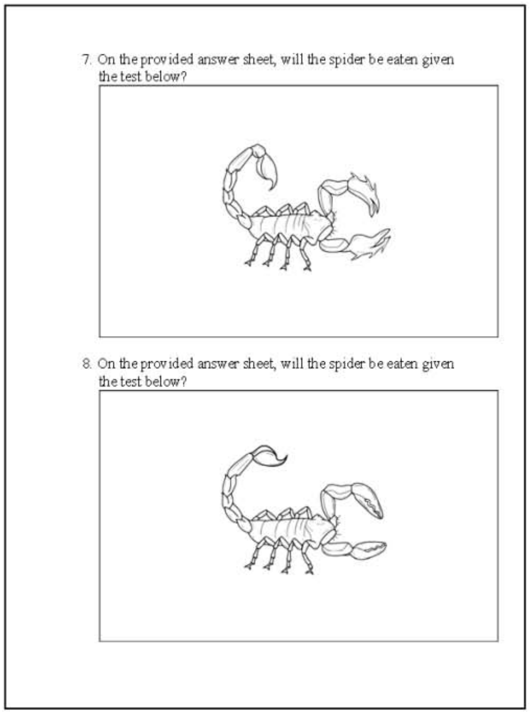
3. On the provided answer sheet, will the spider be eaten given the test below?

4. On the provided answer sheet, will the spider be eaten given the test below?



5. On the provided answer sheet, will the spider be eaten given the test below?

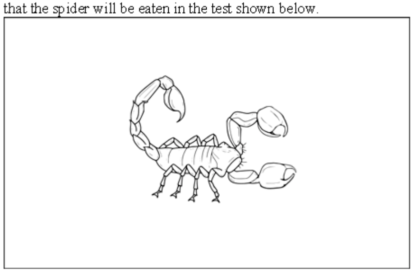
6. On the provided answer sheet, will the spider be eaten given the test below?



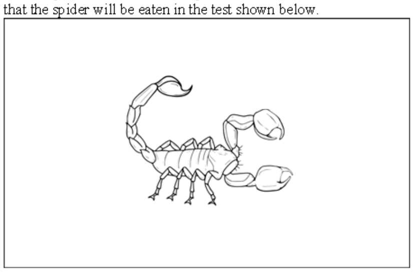
7. On the provided answer sheet, will the spider be eaten given the test below?

8. On the provided answer sheet, will the spider be eaten given the test below?

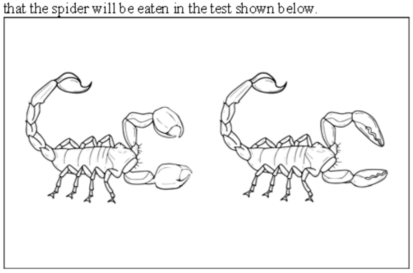
9. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.



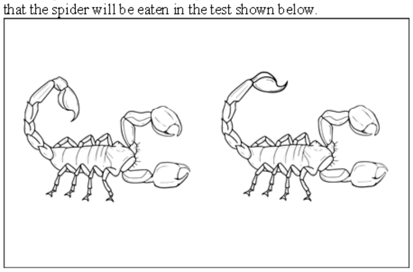
10. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.



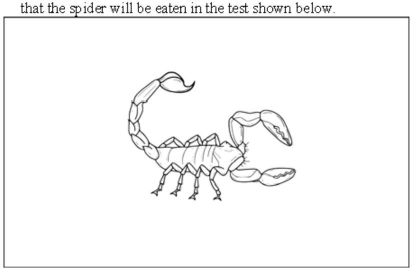
11. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.



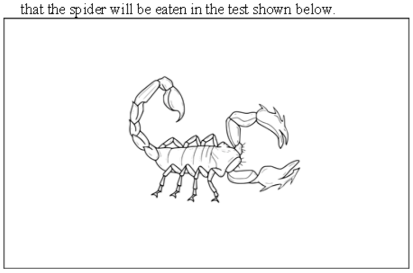
12. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.



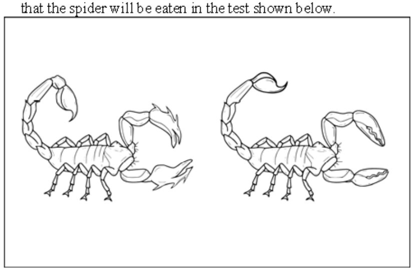
13. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.



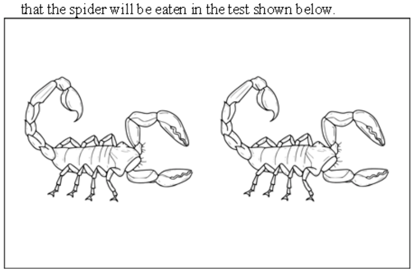
14. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.



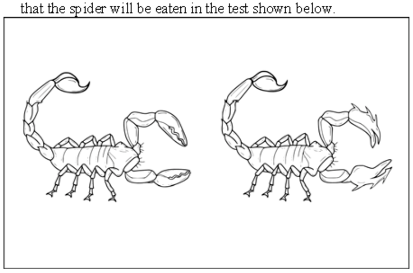
15. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.



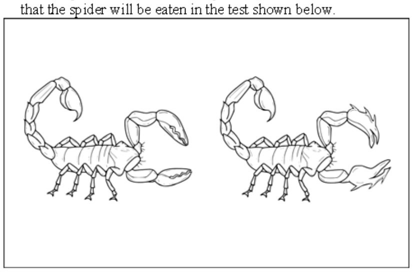
16. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.



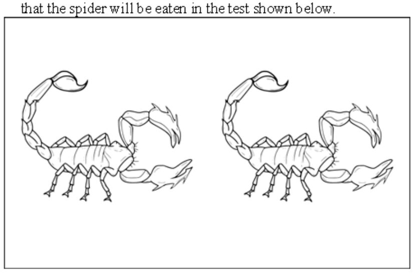
17. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.



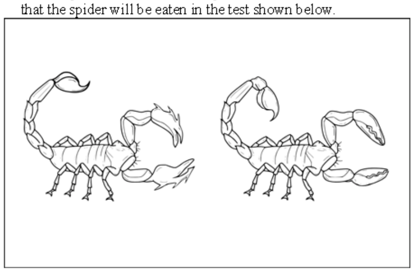
18. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.



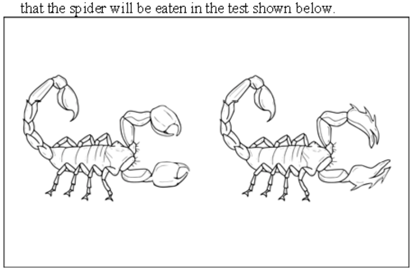
19. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.



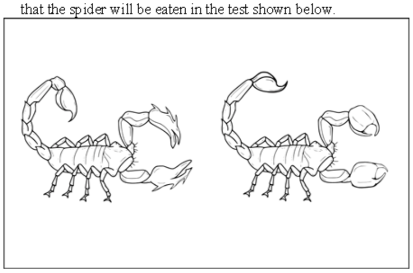
20. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.



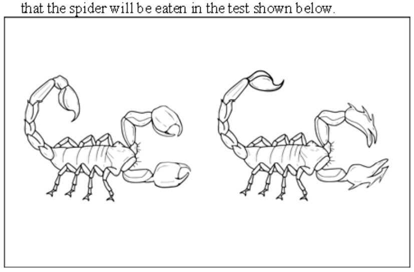
21. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.



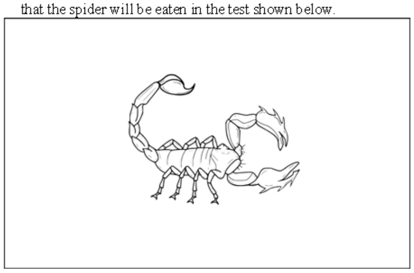
22. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.

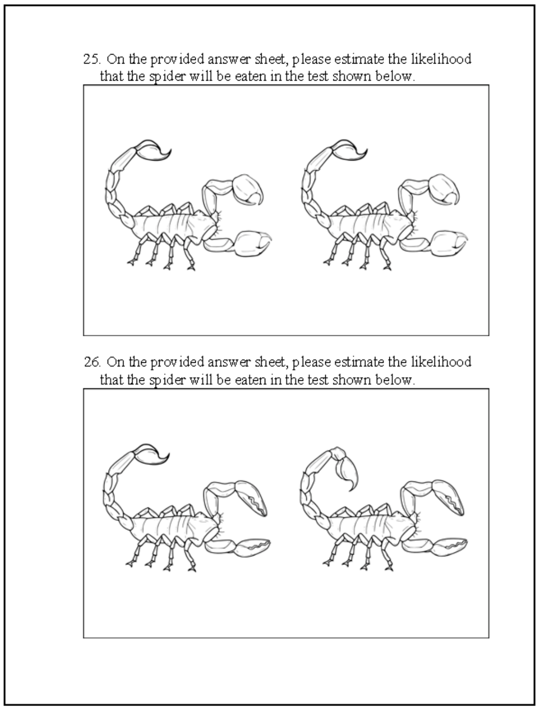


23. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.



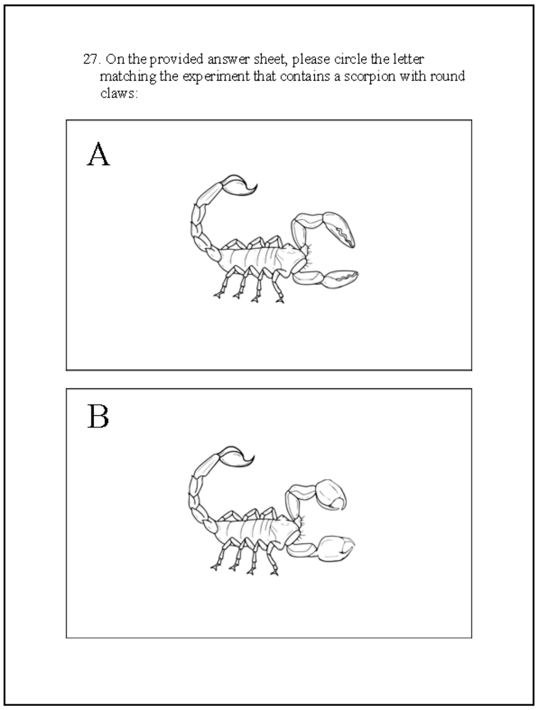
24. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.



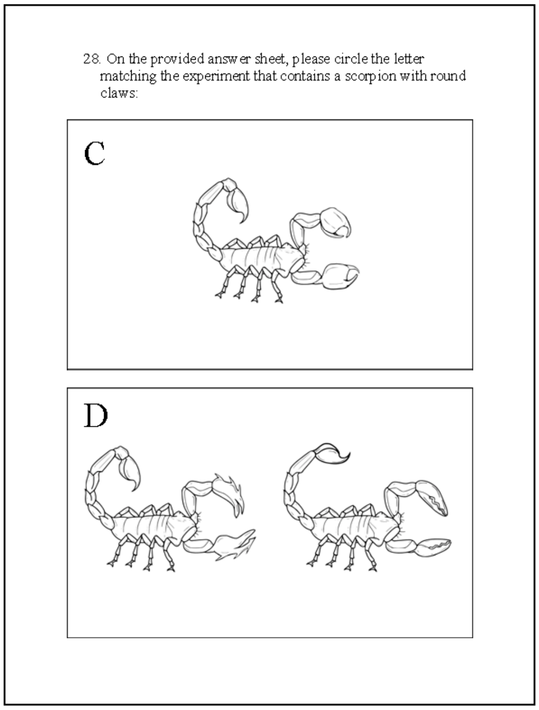


25. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.

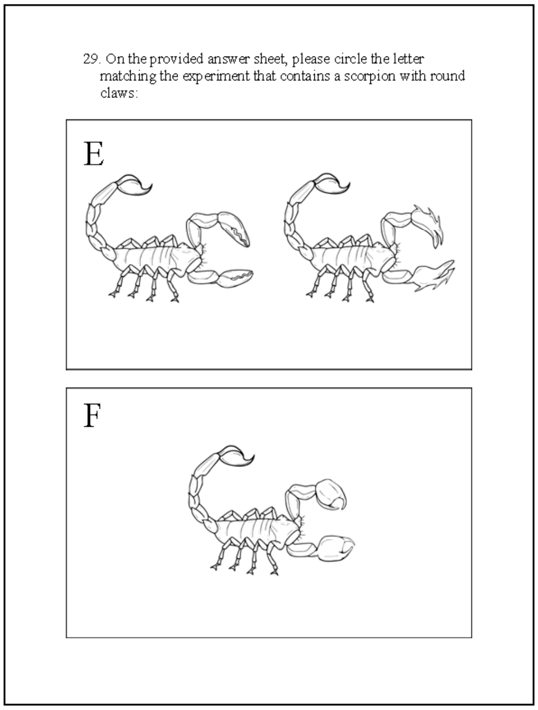
26. On the provided answer sheet, please estimate the likelihood that the spider will be eaten in the test shown below.



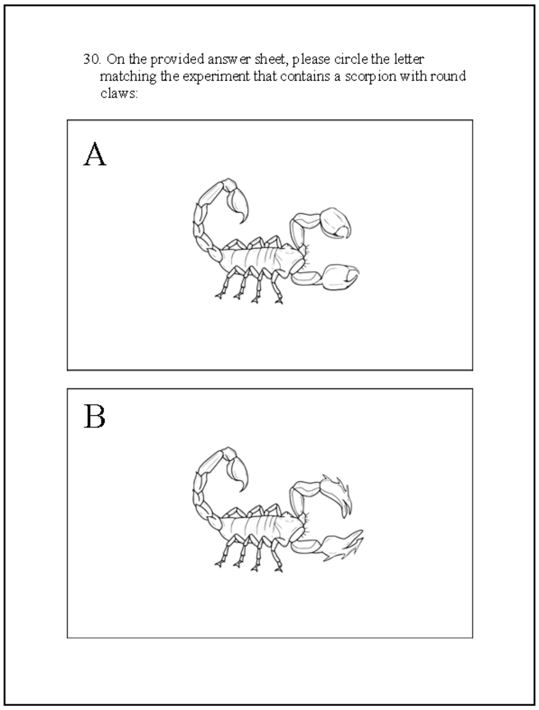
27. On the provided answer sheet, please circle the letter matching the experiment that contains a scorpion with round claws:



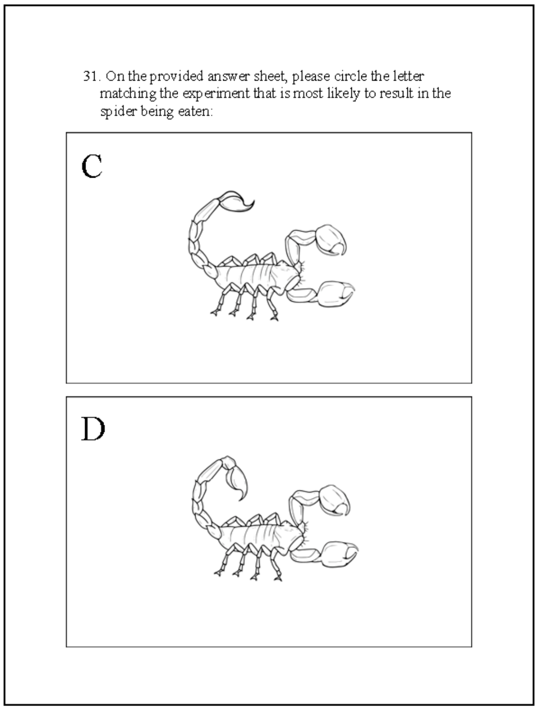
28. On the provided answer sheet, please circle the letter matching the experiment that contains a scorpion with round claws:



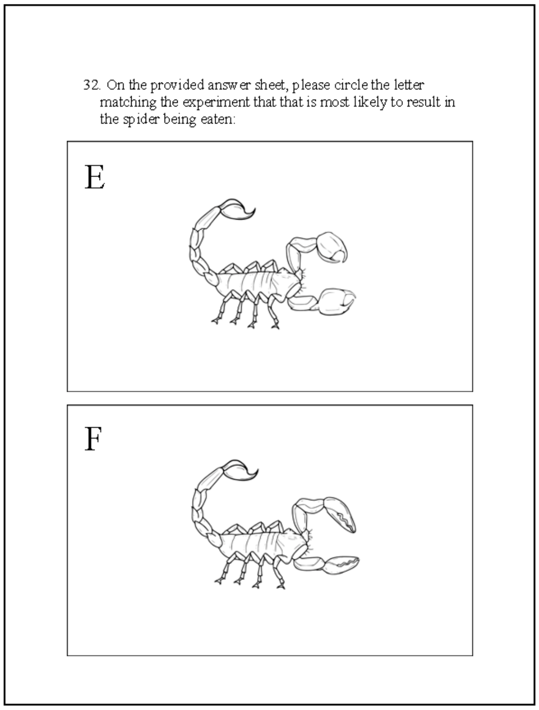
29. On the provided answer sheet, please circle the letter matching the experiment that contains a scorpion with round claws:



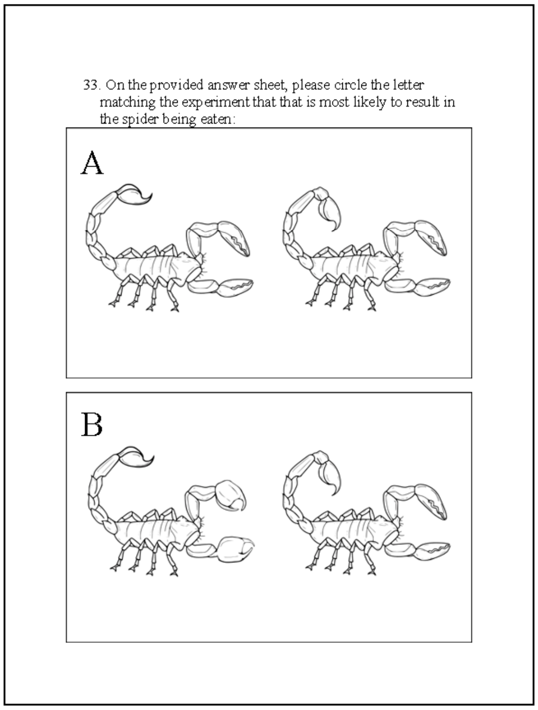
30. On the provided answer sheet, please circle the letter matching the experiment that contains a scorpion with round claws:



31. On the provided answer sheet, please circle the letter matching the experiment that is most likely to result in the spider being eaten:



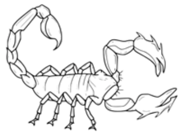
32. On the provided answer sheet, please circle the letter matching the experiment that that is most likely to result in the spider being eaten:



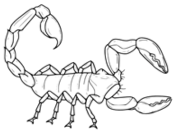
33. On the provided answer sheet, please circle the letter matching the experiment that that is most likely to result in the spider being eaten:

34. On the provided answer sheet, please circle the letter matching the experiment that that is most likely to result in the spider being eaten:

C

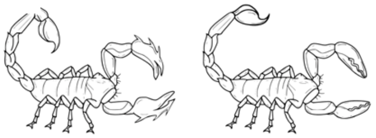


D

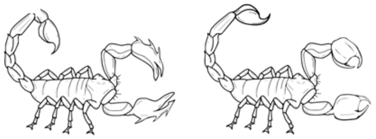


35. On the provided answer sheet, please circle the letter matching the experiment that that is most likely to result in the spider being eaten:

E

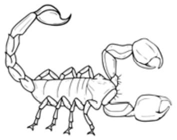


F

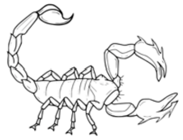


36. On the provided answer sheet, please circle the letter matching the experiment that that is most likely to result in the spider being eaten:

A

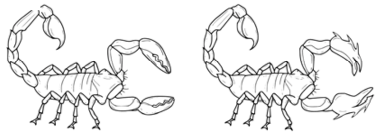


B

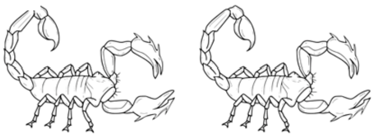


37. On the provided answer sheet, please circle the letter matching the experiment that that is most likely to result in the spider being eaten:

C

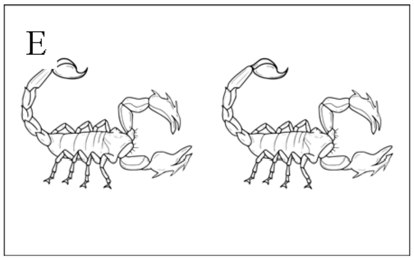


D

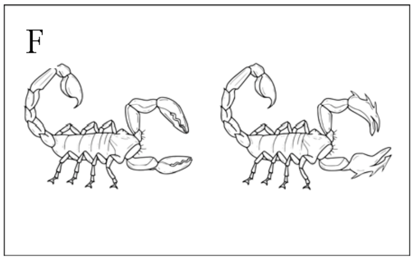


38. On the provided answer sheet, please circle the letter matching the experiment that that is most likely to result in the spider being eaten:

E

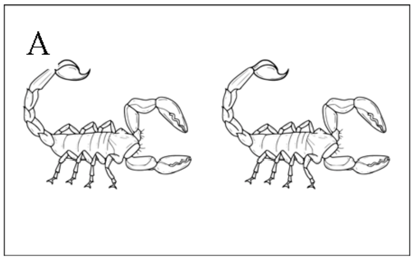


F

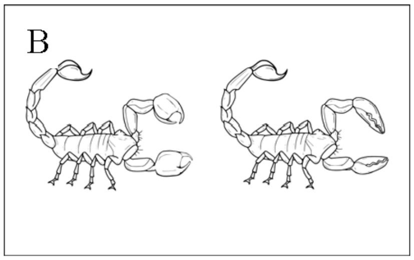


39. On the provided answer sheet, please circle the letter matching the experiment that that is most likely to result in the spider being eaten:

A

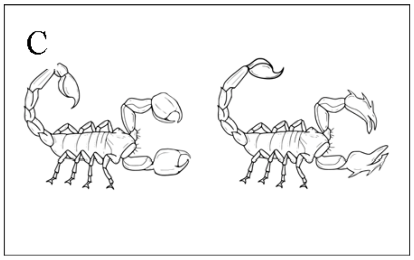


B

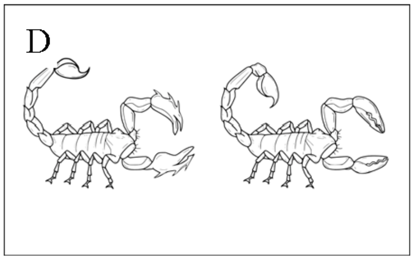


40. On the provided answer sheet, please circle the letter matching the experiment that that is most likely to result in the spider being eaten:

C

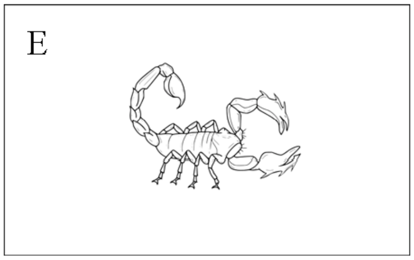


D

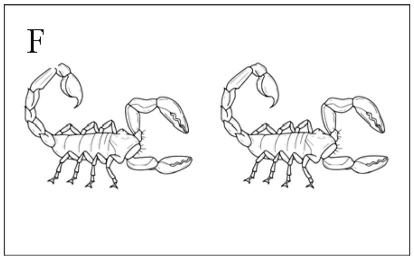


41. On the provided answer sheet, please circle the letter matching the experiment that that is most likely to result in the spider being eaten:

E

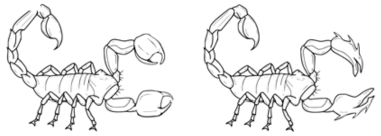


F

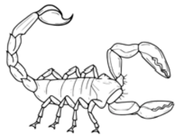


42. On the provided answer sheet, please circle the letter matching the experiment that that is most likely to result in the spider being eaten:

A

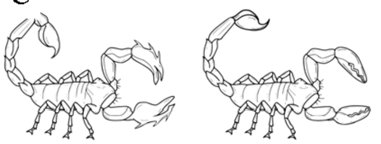


B

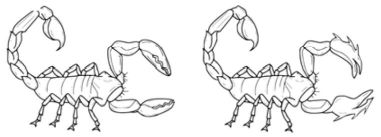


43. On the provided answer sheet, please circle the letter matching the experiment that that is most likely to result in the spider being eaten:

C

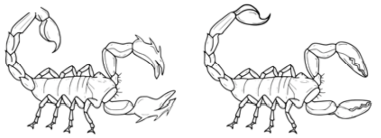


D

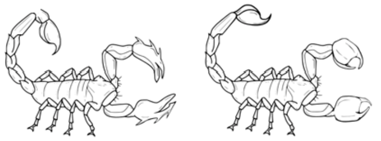


44. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:

A

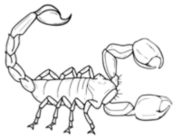


B

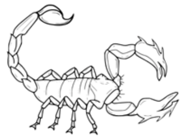


45. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:

C

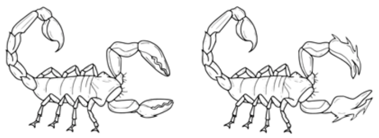


D

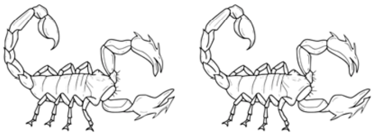


46. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:

E

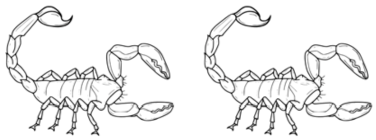


F

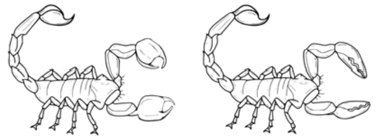


47. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:

A

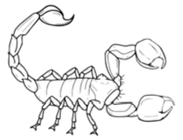


B

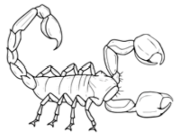


48. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:

C

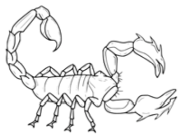


D

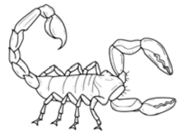


49. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:

E

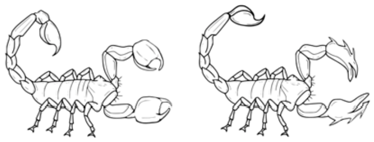


F

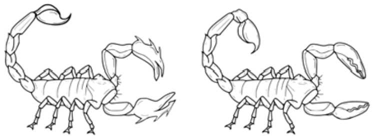


50. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:

A

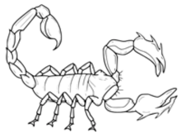


B

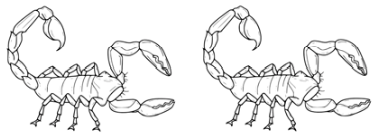


51. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:

C

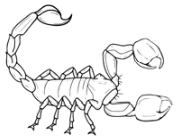


D

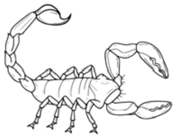


52. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:

E

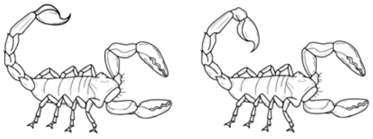


F

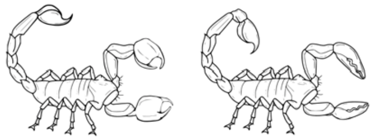


53. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:

A

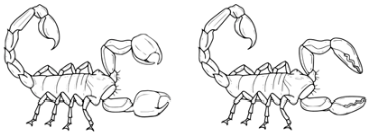


B

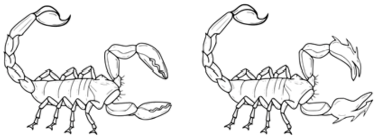


54. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:

C

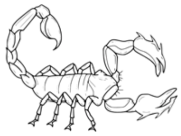


D

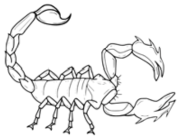


55. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:

E

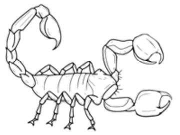


F

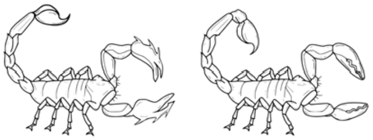


56. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:

A

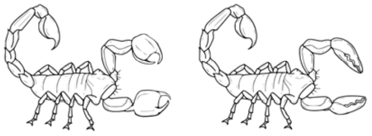


B

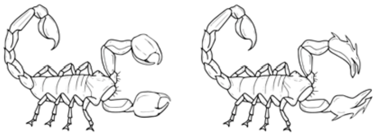


57. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:

C

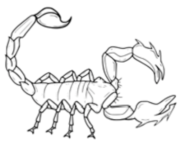


D

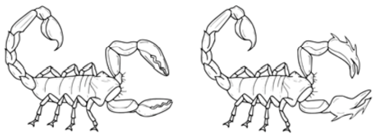


58. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:

E

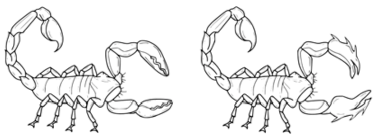


F

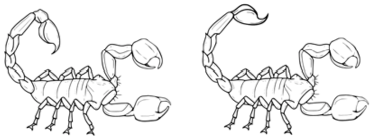


59. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:

A

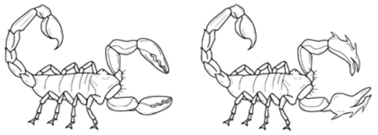


B

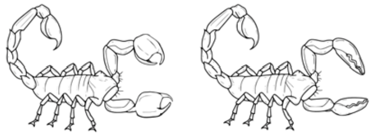


60. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:

C

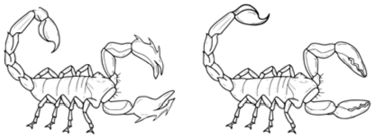


D

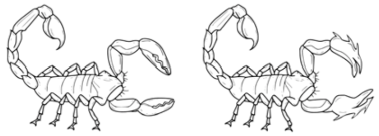


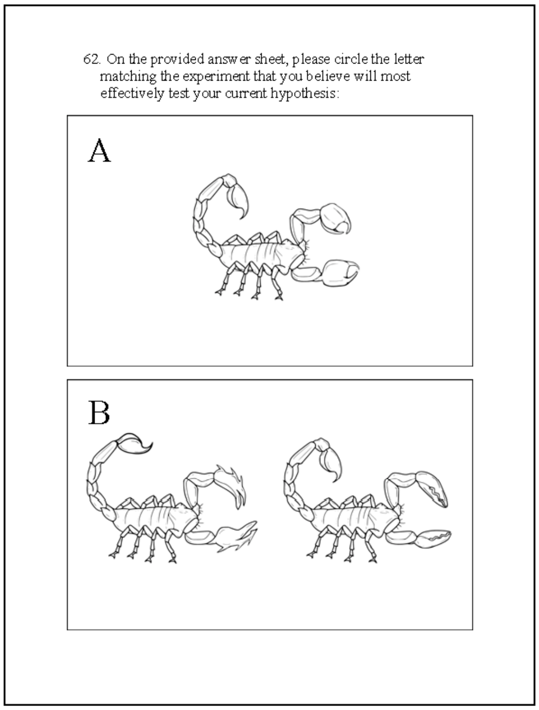
61. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:

E

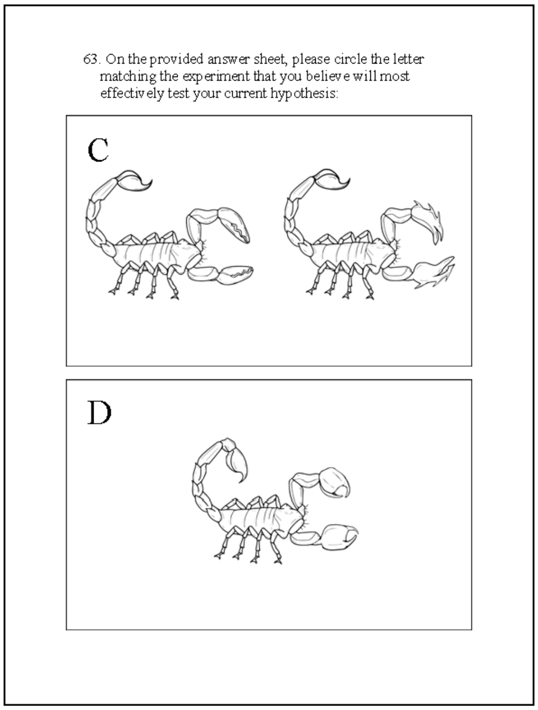


F





62. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:



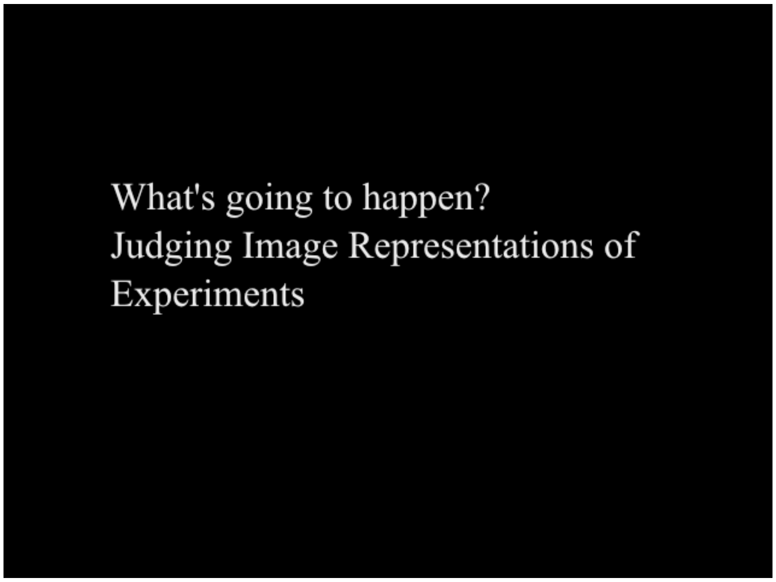
63. On the provided answer sheet, please circle the letter matching the experiment that you believe will most effectively test your current hypothesis:

APPENDIX D

TRANSCRIPTS AND SCREENSHOTS OF AUDIO/VIDEO INSTRUCTIONS

Transcript/Screenshots of General Instructions

Video: Display welcome screen (Figure D.1).



What's going to happen?
Judging Image Representations of
Experiments

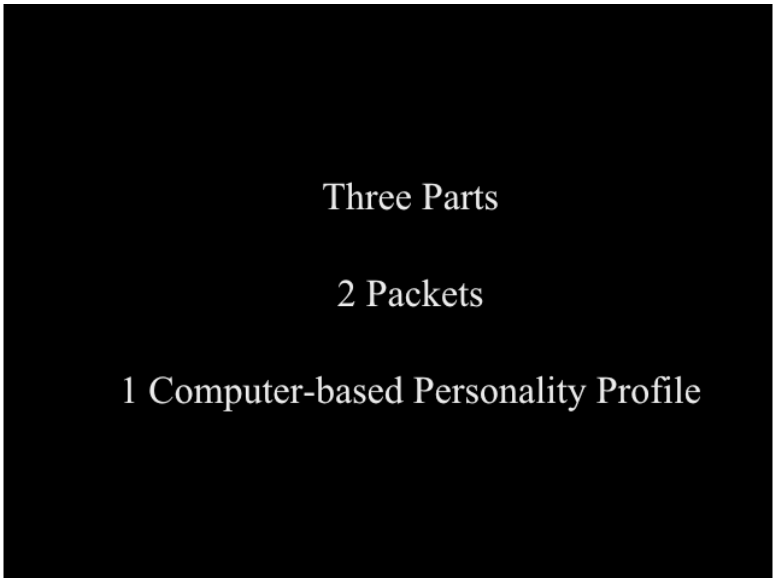
*Figure D.1.*Initial screen of the general instructions video.

Narrator: “Welcome to “What’s going to happen? Judging image representations of experiments.” This study should take less than 2 hours. You will receive 2 credits for your participation. In this study, we will ask you to read a description of a scenario and answer some questions about images that represent possible experiments related to the described scenario.”

Video: Display experiment structure screen (Figure D.2).

Narrator: “The experiment is done in three parts. There is a packet of questions, a personality profile, and a second packet of questions.”

Video: Display experiment materials screen (Figure D.3).



Three Parts

2 Packets

1 Computer-based Personality Profile

*Figure D.2.*The experiment structure screen describes the basic structure of the experiment for the participant.

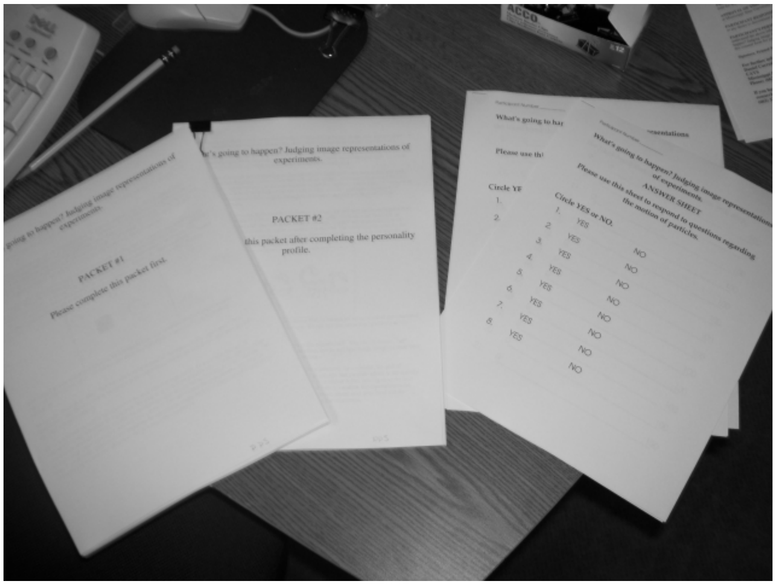
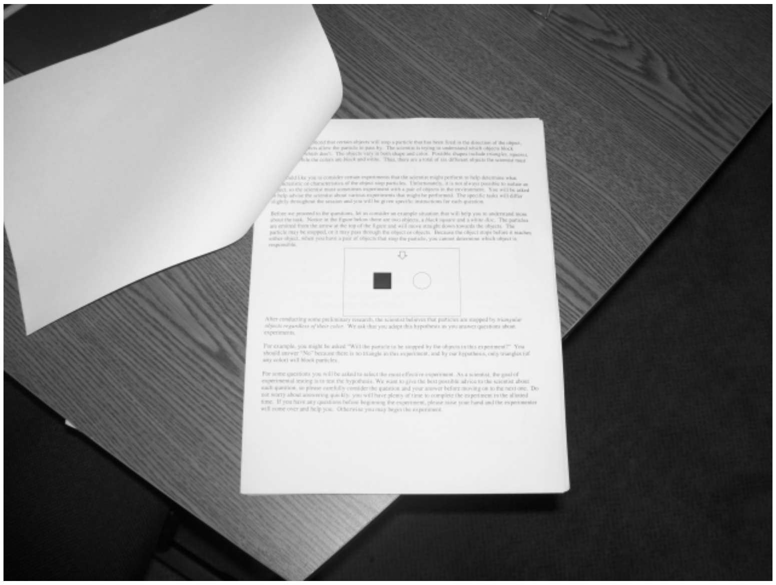


Figure D.3. The experiment materials screen displays a photo of two question packets and two answer packets.

Narrator: “The experimenter should have given you two packets of questions and two answer sheets. The packets should be labeled packet one and packet two. Please complete packet one first, then do the personality profile on this computer, and then complete the second packet.”

Video: Display questionnaire packet open to page one (Figure D.4).

Narrator: “Each of the packets begin with a description of a scientist facing a problem. The rest of the packet contains questions about possible experiments. Please carefully read the description before answering the questions about the experiments.”



lieved that certain objects will stop a particle that has been fired in the direction of the object,
jects allow the particle to pass by. The scientist is trying to understand which objects block
ticles don't. The objects vary in both shape and color. Possible shapes include triangles, squares,
hile the colors are black and white. Thus, there are a total of six different objects the scientist must

ould like you to consider certain experiments that the scientist might perform to help determine what
acteristic or characteristics of the object stop particles. Unfortunately, it is not always possible to isolate an
ject, so the scientist must sometimes experiment with a pair of objects in the environment. You will be asked
help advise the scientist about various experiments that might be performed. The specific tasks will differ
slightly throughout the session and you will be given specific instructions for each question.

Before we proceed to the questions, let us consider an example situation that will help you to understand more about the task. Notice in the figure below there are two objects, a *black square* and a *white disc*. The particles are emitted from the arrow at the top of the figure and will move straight down towards the objects. The particle may be stopped, or it may pass through the object or objects. Because the object stops before it reaches either object, when you have a pair of objects that stop the particle, you cannot determine which object is responsible.

After conducting some preliminary research, the scientist believes that particles are stopped by *triangular objects regardless of their color*. We ask that you adopt this hypothesis as you answer questions about experiments.

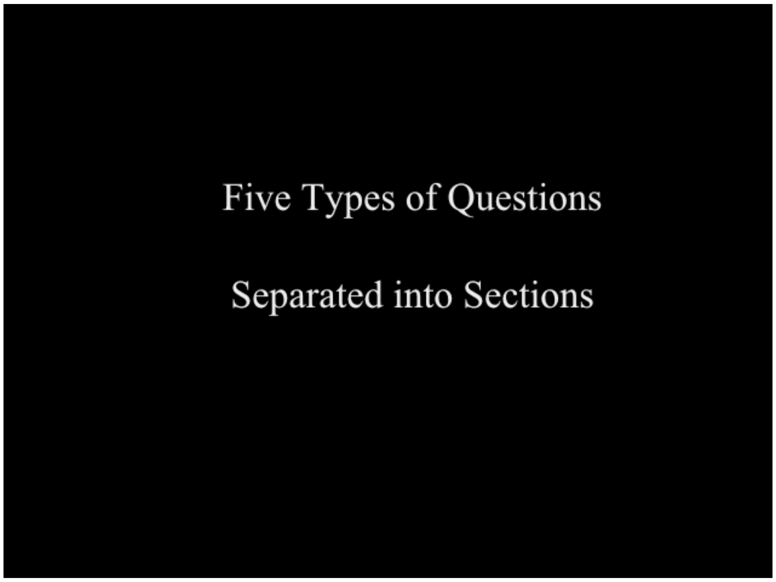
For example, you might be asked “Will the particle to be stopped by the objects in this experiment?” You should answer “No” because there is no triangle in this experiment, and by our hypothesis, only triangles (of any color) will block particles.

For some questions you will be asked to select the most effective experiment. As a scientist, the goal of experimental testing is to test the hypothesis. We want to give the best possible advice to the scientist about each question, so please carefully consider the question and your answer before moving on to the next one. Do not worry about answering quickly: you will have plenty of time to complete the experiment in the allotted time. If you have any questions before beginning the experiment, please raise your hand and the experimenter will come over and help you. Otherwise you may begin the experiment.

*Figure D.4.*The video shows a photo of the questionnaire packet open to the instruction page.

Video: Display five types of questions slide (Figure D.5).

Narrator: “There are five types of questions in e ach packet. The questions are grouped together and a page separates each section with a description of the upcoming questions.”



Five Types of Questions

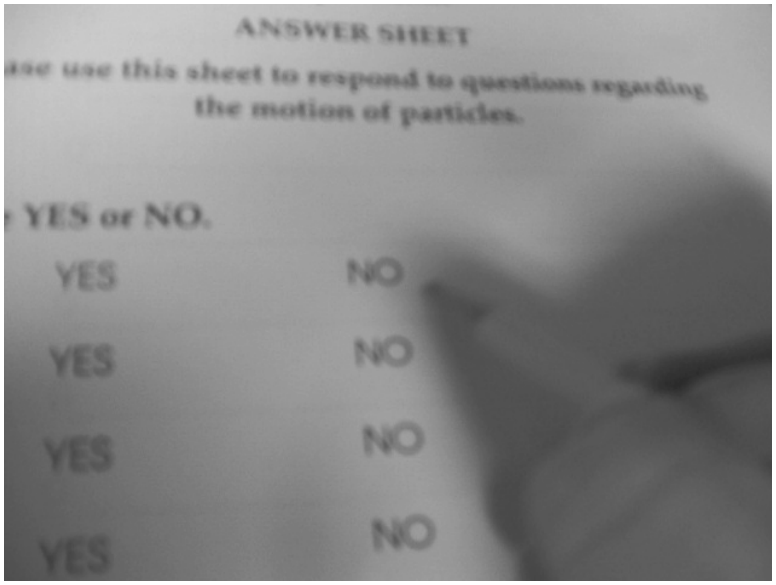
Separated into Sections

Figure D.5. Slide visually reinforces narration describing five sections each with its own type of question.

Video: Display video of experimenter responding to a question on the answer sheet. See Figure D.6 for a screenshot from the video.

Narrator: “On your answer sheet, simply circle yes if you believe the event will occur. Circle no if you believe the event is not going to occur.”

Video: Display photo of visual analog scales on answer sheet for positivity assessment questions (Figure D.7).



ANSWER SHEET

ase use this sheet to respond to questions regarding the motion of particles.

YES or NO.

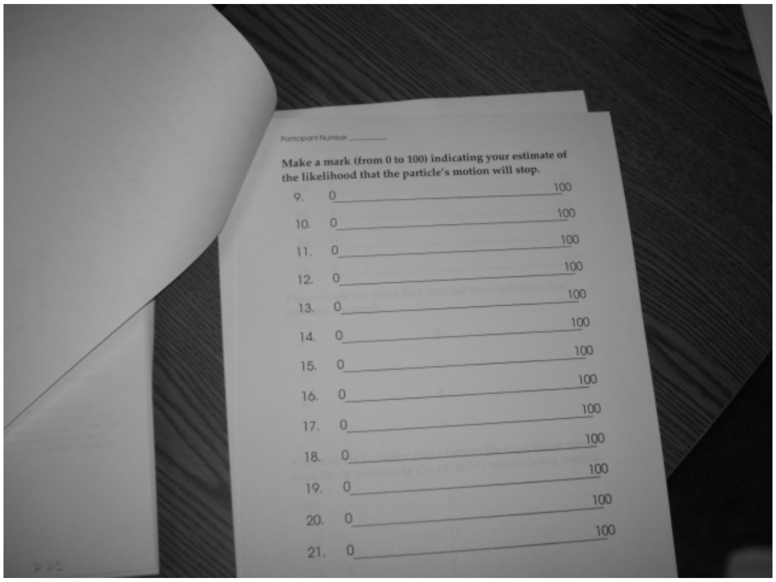
YES NO

YES NO

YES NO

YES NO

Figure D.6. Screenshot from a video of the experimenter circling “no” on the answer sheet.



Participant Number ________

Make a mark (from 0 to 100) indicating your estimate of the likelihood that the particle's motion will stop.

9. 0 ________________ 100

10. 0 ________________ 100

11. 0 ________________ 100

12. 0 ________________ 100

13. 0 ________________ 100

14. 0 ________________ 100

15. 0 ________________ 100

16. 0 ________________ 100

17. 0 ________________ 100

18. 0 ________________ 100

19. 0 ________________ 100

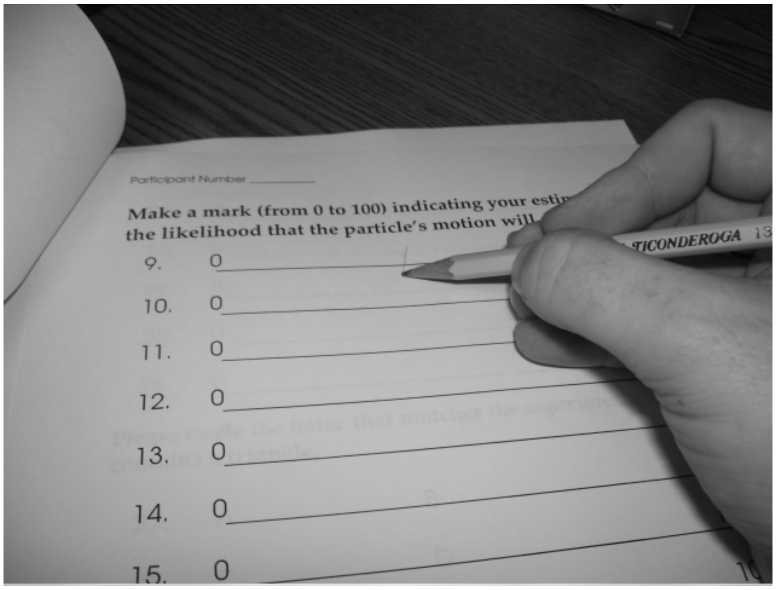
20. 0 ________________ 100

21. 0 ________________ 100

Figure D.7. Photo of the visual analog scales used to respond to the positivity assessment questions.

Narrator: “The second type of question will ask you to estimate the likelihood that an event will occur given the image that is representing the experiment.”

Video: Display video of experimenter responding to positivity assessment question using the visual analog scale. See Figure D.8 for a screenshot from the video.



Participant Number ________

Make a mark (from 0 to 100) indicating your esti
the likelihood that the particle's motion will

9. 0________

10. 0________

11. 0________

12. 0________

13. 0________

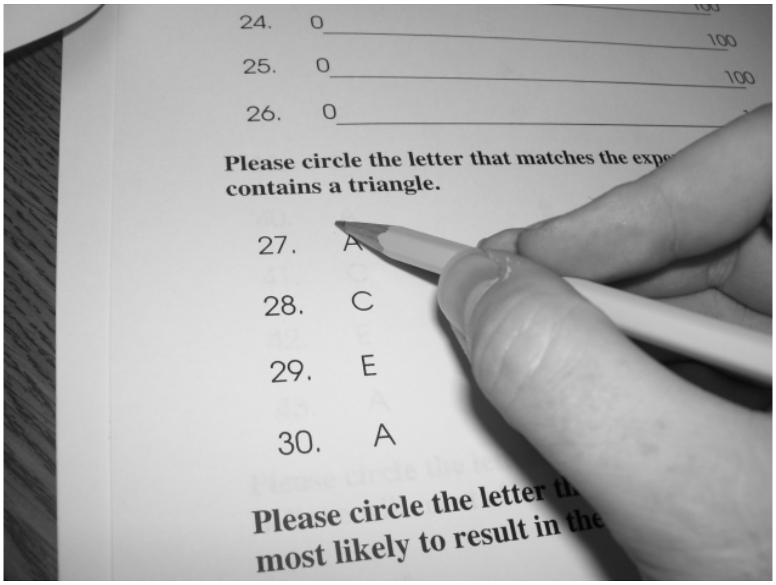
14. 0________

15. 0________ 10

Figure D.8. Screenshot taken from video of experimenter responding to positivity assessment question using visual analog scale.

Narrator: “On your answer sheet, simply make a vertical mark on the line that represents approximately what you think the likelihood is. For example, if you believe there is a 50/50 chance that the event will occur, you would make a mark approximately in the center between 0 and 100.”

Video: Display photo (Figure D.9) of responding to one of the test selection questions (catch trial, explicit positive selection, or hypothesis test selection).



24. 0____________________ 100

25. 0____________________ 100

26. 0____________________

Please circle the letter that matches the exp
contains a triangle.

27. A

28. C

29. E

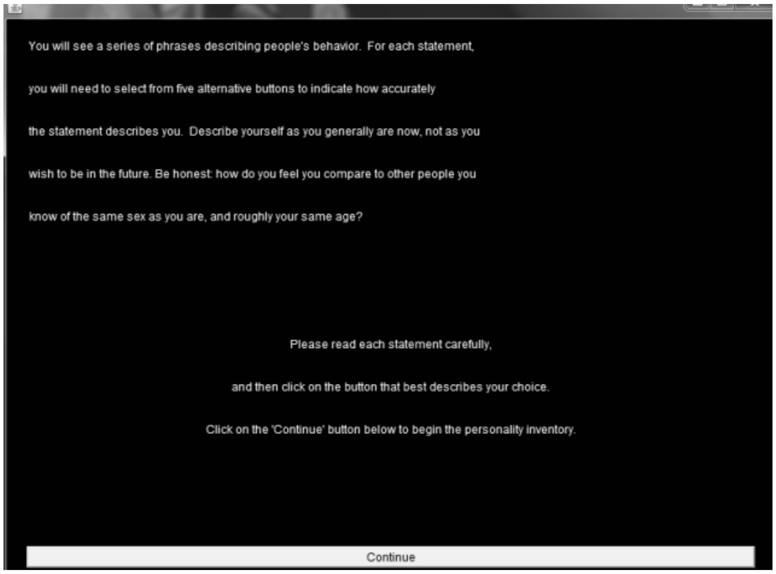
30. A

Please circle the letter th
most likely to result in the

*Figure D.9.*Photo of experimenter circling an answer to indicate selection of one of a pair of figures as in the catch trials, explicit positive selection and hypothesis test selection.

Narrative: “The third type of question asks you to find a specific feature in one of two experiment images. When you find the feature, circle the letter of the experiment that contains the feature on your answer sheet. The fourth type of questions asks you to select in which experiment of two is an event most likely to occur. Simply circle the letter that matches the experiment that you believe will most likely have the event occur on your answer sheet. The fifth type of question asks you to consider the description given at the beginning of the packet and select the test that will be most effective in reaching the scientist’s goal in solving the current problem.”

Video: Display screenshot of IPIP instructions screen (see Figure D.10).



You will see a series of phrases describing people's behavior. For each statement, you will need to select from five alternative buttons to indicate how accurately the statement describes you. Describe yourself as you generally are now, not as you wish to be in the future. Be honest: how do you feel you compare to other people you know of the same sex as you are, and roughly your same age?

Please read each statement carefully,

and then click on the button that best describes your choice.

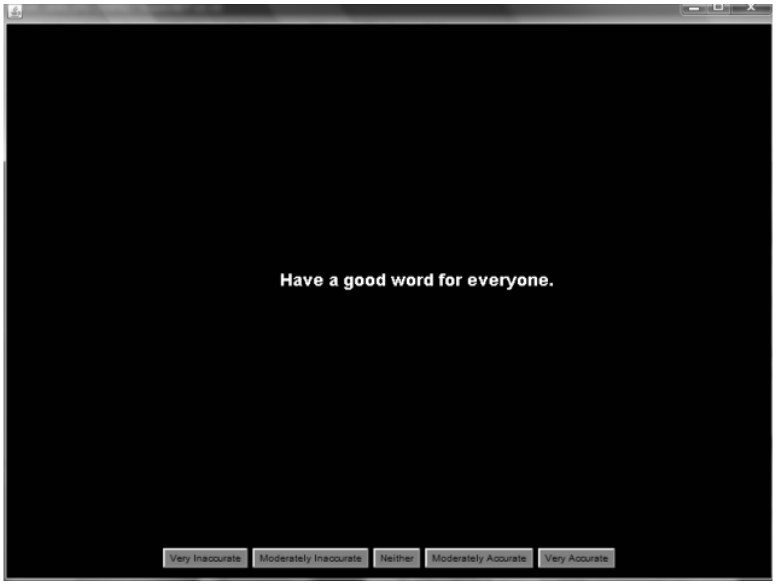
Click on the 'Continue' button below to begin the personality inventory.

*Figure D.10.*Screenshot of IPIP instruction screen.

Narrator: "Between the two packets, you will be asked to complete a personality profile on this computer. There will be a screen that will ask you to press a button that will start the personality profile. Please do not start the personality profile until you have completed the first packet of questions."

Video: Display screenshot of an IPIP item screen (see Figure D.11).

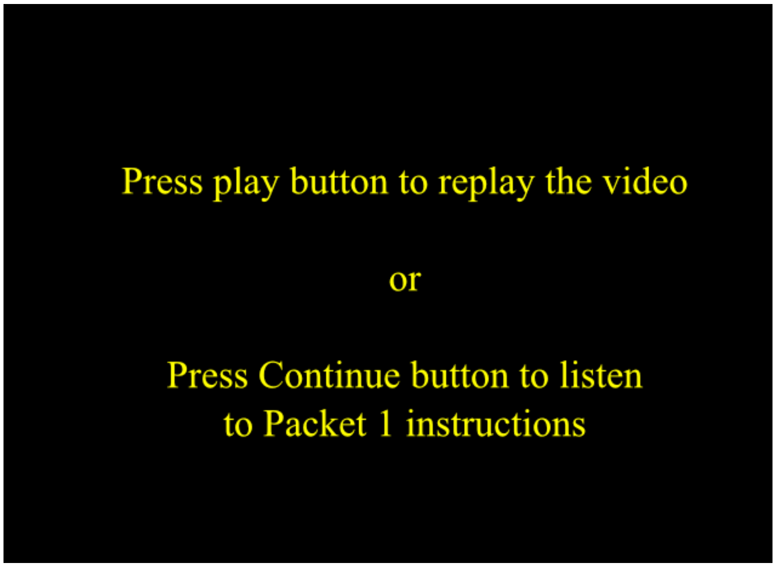
Narrator: "The personality profile is one hundred questions that are framed as a statement such as 'I am the life of the party.' Select the button that most matches your agreement with the statement on-screen. After you complete the personality profile, then you will be asked to continue on and complete the second packet of questions."



Have a good word for everyone.

Figure D.11. Screenshot of the IPIP item screen. The IPIP item is displayed centered in the display with the response buttons across the bottom of the screen.

Video: Display replay instructions slide (see Figure D.12).



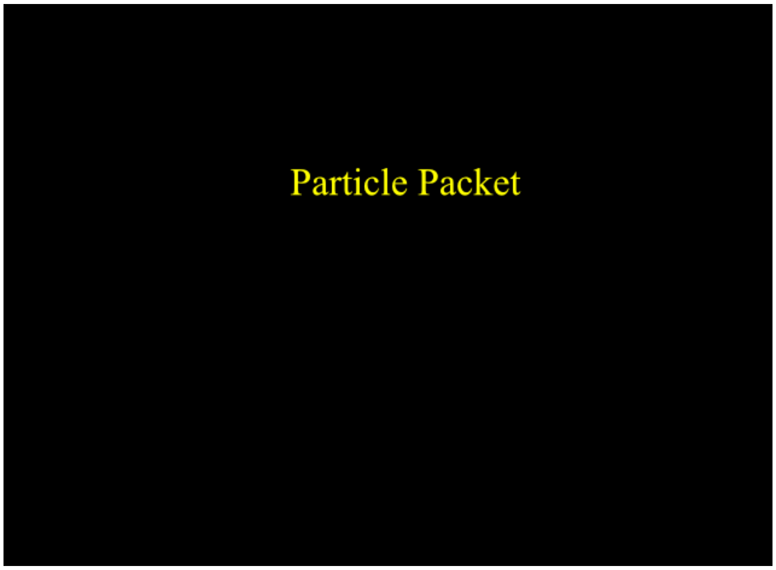
Press play button to replay the video

or

Press Continue button to listen to Packet 1 instructions

Figure D.12. Screenshot of final slide of the general instructions video. The slide provides instructions for replaying the instructions video.

Transcript/Screenshots of Instructions for MDT Microworld Scenario



Particle Packet

Figure D.13. Welcome screen for MDT Microworld packet audio instructions.

Video: Display packet welcome screen (see Figure D.13).

Narrator: “Please open your packet to the first page containing the description of the scenario. This scenario involves the investigation of the motion of particles past objects.”

Video: Display MDT Microworld example screen (see Figure D.14).

Narrator: “A scientist has noticed that certain objects will stop a particle that has been fired in the direction of an object while other objects allow a particle to pass by. The scientist is trying to understand which objects block particle motion and which do not.

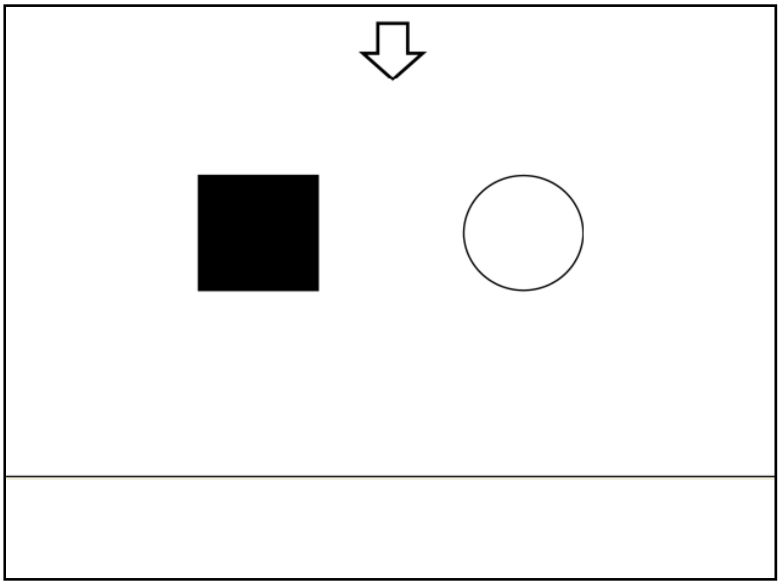


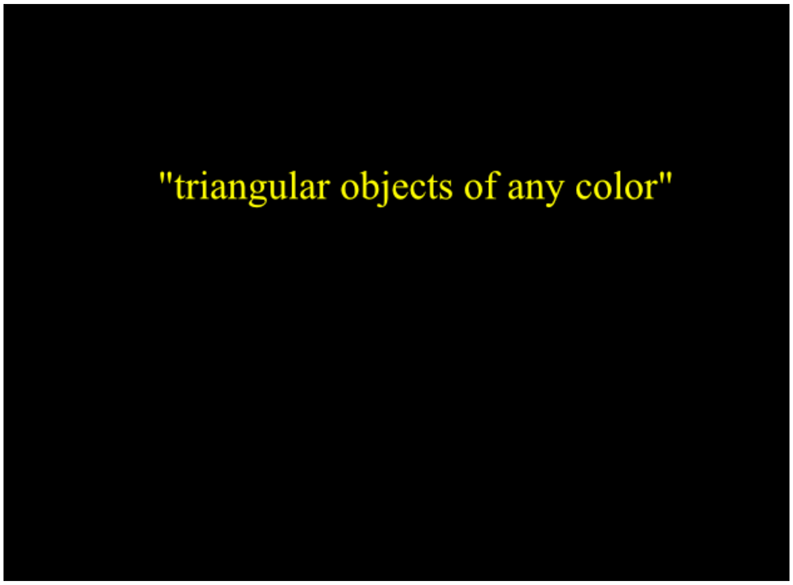
*Figure D.14.*MDT Microworld example figure.

Narrator (High Miss Cost only): “If the scientist can determine what is stopping particle motion, he can use this knowledge to develop a new treatment for cancer.”

Narrator: “The objects can vary in shape and color. Some possible shapes include triangles, squares and discs and black and white colors. There are a total of six different objects that the scientist must consider. We are going to ask you to consider possible experiments that the scientist might perform to help determine what characteristic of the objects stop particle motion. Sometimes it is impossible to isolate a single object so the scientist will sometimes experiment with a pair of objects in an environment. You will be asked to help advise the scientist about the various experiments that might be performed. The image that is on this page is an example of how the experiments are represented in the questions section of this packet. The image shown here has two objects: a black square and a white disc. Particles are emitted from the arrow at the top of the figure and move straight down towards the objects. The particle may stop moving or it may pass between the objects. If an object stops the particles motion and you have two objects, you cannot determine which of the two objects are responsible."

Video (All but Alternative Hypothesis): Display given hypothesis (see Figure D.15).

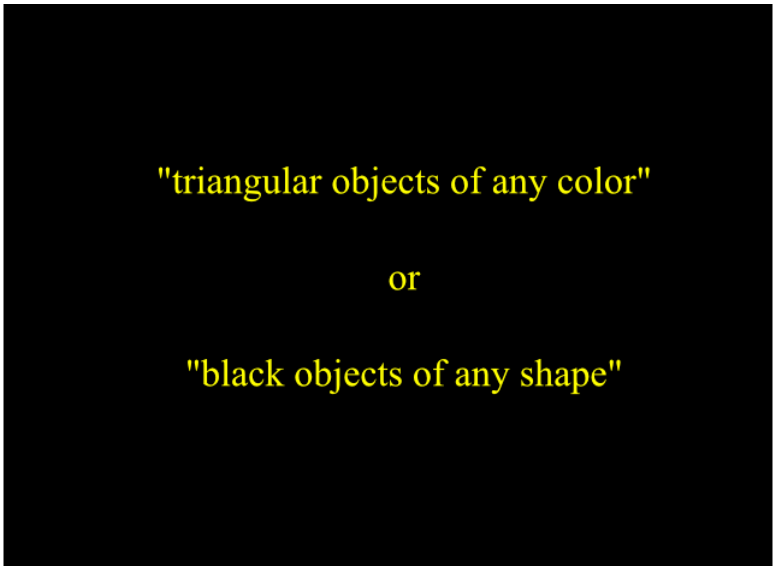
Video (Alternative Hypothesis): Display primary and alternative hypothesis (see Figure D.16).



"triangular objects of any color"

Figure D.15. Screenshot of video instructions giving the hypothesis.

Narrator: "For this experimental scenario, the scientist believes that particles are stopped by triangular objects regardless of their color."



"triangular objects of any color"

or

"black objects of any shape"

Figure D.16. Screenshot of video instructions giving primary and alternative hypothesis.

Narrator (Common Event only): “During preliminary research, the particles were stopped regularly by the objects.”

Narrator (Alternative Hypothesis only): “Another researcher working with the same particles believes that the particles are stopped by black particles regardless of their shape.”

Narrator: “We ask that you adopt this hypothesis as you answer questions about the experiments. An example question might be: “Will the particle be stopped by the objects in this experiment?” and you would be given an image of the experiment. So given the image on this page, you should answer “no” because there is no triangle in this experiment and by the given hypothesis only triangles of any color will block the particles.”

Narrator (Alternative Hypothesis only): “However, you may answer “yes” under the alternative hypothesis provided by the other researcher because there is a black object in the experiment.”

Narrator (Test only): “As a scientist, the goal of experimental testing is to test the current hypothesis.”

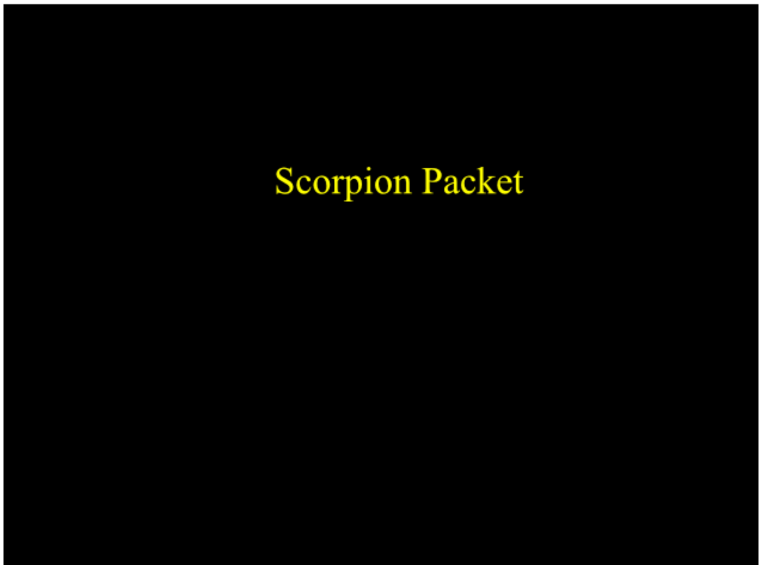
Narrator (Confirm only): “As a scientist, the goal of experimental testing is to prove the current hypothesis is the correct one.”

Narrator (Disconfirm only): “As a scientist, the goal of experimental testing is to falsify the current hypothesis.”

Narrator: “We want to give the best possible advice to the scientist about each question so please carefully consider the question and your answer before moving on to the next one. If you have any questions before beginning the experiment, please raise your hand and the experimenter will come over and help you. Otherwise you may begin your experiment. If you would like to replay this video, press the play button on the bottom left. If you are ready to continue press the continue button at the bottom of this window.”

Transcript/Screenshots of Instructions for Scorpion-Spider Scenario

Video: Display packet welcome screen (see Figure D.17).



Scorpion Packet

Figure D.17. Welcome screen for Scorpion-Spider packet audio instructions.

Narrator: Please open your packet to the first page that contains the description of the scenario. This scenario involves scorpions attacking and consuming a particular species of spider. A scientist is working with farmers to try to determine which types of scorpions eat the spiders and which do not."

Narrator (High Miss Cost only): "The spiders are beneficial to pest control and the scientist must discover which scorpions are eating the spiders to avoid significant loss for the farmers."

Video: Display scorpion tail and claw features (see Figure D.18).

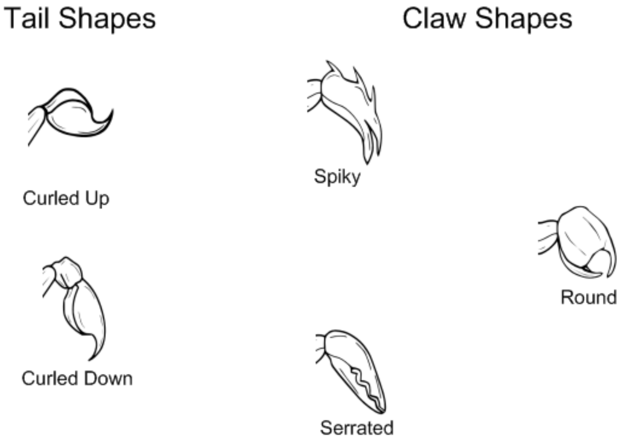


Figure D.18. Scorpion tail and claw features and their labels.

Narrator: “The scorpion species can be identified by their tail shape: up or down; and their claw shape: round, spiky, or serrated. There are a total of six types of scorpion. We are going to ask you to consider some possible experiments that will allow the scientist to understand what features determine which scorpions will eat spiders.”

Video: Display Scorpion-Spider example figure (see Figure D.19).

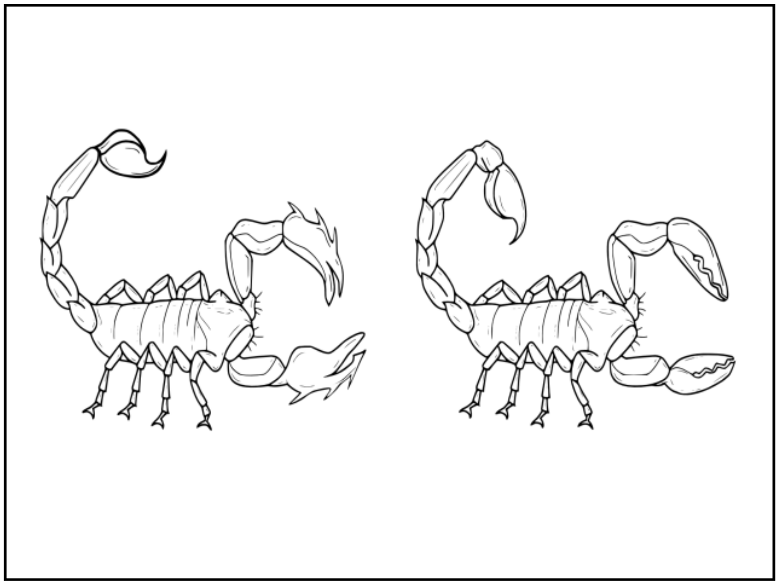
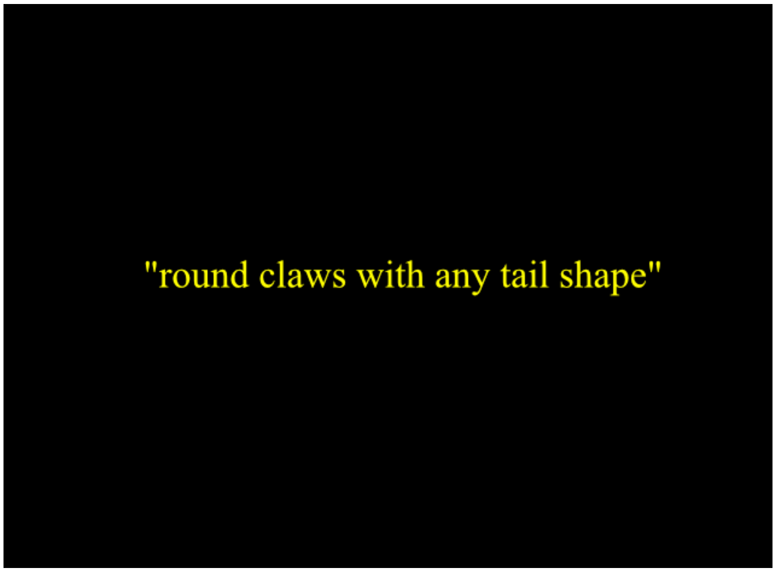


*Figure D.19.*Scorpion-Spider example figure.

Narrator: “You will be asked to help advise the scientist in selecting the experiments that might be performed. The image on this page represents two scorpion types: the left one with spiky claws and an up-turned tail; the right one with serrated claws and a down-turned tail.”

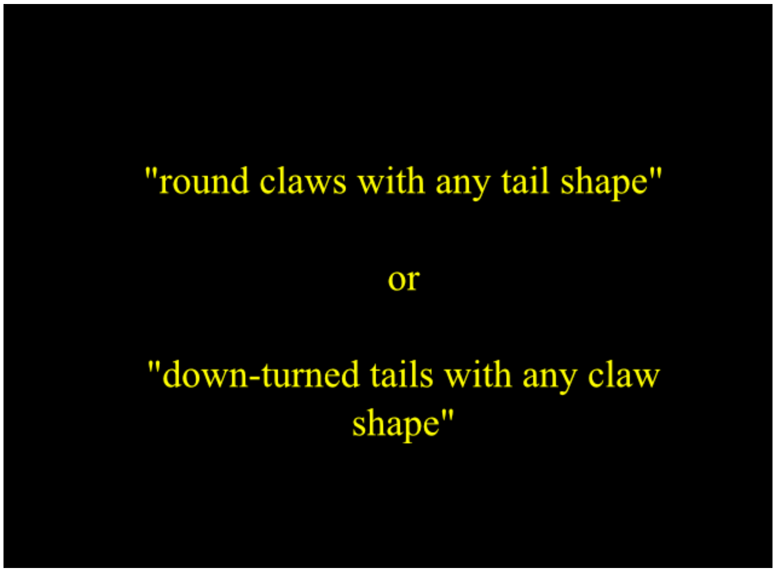
Video (All But Alternative Hypothesis): Display given hypothesis (see Figure D.20).

Video (Alternative Hypothesis only): Display primary and alternative hypotheses (see Figure D.21).



"round claws with any tail shape"

Figure D.20. Screenshot of video instructions giving hypothesis.



"round claws with any tail shape"

or

"down-turned tails with any claw shape"

Figure D.21. Screenshot of video instruction giving primary and alternative hypotheses.

Narrator (All but Alternative Hypothesis): “Based on preliminary field studies, the scientist believes that the spiders are being attacked and consumed by scorpions with the round claws regardless of their tail shape.”

Narrator (Alternative Hypothesis only): “Other scientists have suggested instead that scorpions with tails that turn down regardless of their claw shape are responsible for the deaths of the spiders.”

Narrator (Common Event only): “The farmers have noted that spiders rarely survive long once the scorpions are introduced to their fields.”

Narrator: “We are going to ask that you adopt this hypothesis as you answer questions about the experiments. For example, you might be asked: will the spider be eaten in this experiment? Given the image above, on the paper, you should answer “no” because none of the scorpions in this experiment have round claws and by our hypothesis scorpions must have round claws and any tail shape to attack and consume the spider.”

Narrator (Alternative Hypothesis only): “However, given the alternative hypothesis suggested by other scientists, you may answer “yes” because one of the scorpions in the environment has a tail that is curled down.”

Narrator (Test only): “As a scientist, the goal of experimental testing is to test your current hypothesis.”

Narrator (Confirm only): “As a scientist, the goal of experimental testing is to prove that the current hypothesis is the correct one.”

Narrator (Disconfirm only): “As a scientist, the goal of experimental testing is to falsify your current hypothesis.”

Narrator: “We want you to give the best possible advice to the scientist about each question so please carefully consider the question and your answer before moving on to the next one. If you have any questions before beginning the experiment, please raise your hand and the experimenter will come over and help you. Otherwise you may begin your experiment. If you would like to replay this video, press the play button on the bottom left. If you are ready to continue press the continue button at the bottom of this window.”

APPENDIX E

GENERALIZED ETA-SQUARED

Effect size measures provide a quantified and standardized measure of the difference between the null hypothesis and the alternative hypothesis that is independent of sample size (Olejnik & Algina, 2003). There are several different measures of effect size that largely depend on the statistical test or research design used for the analysis. Olejnik and Algina (2003) discuss implications of research design on the eta squared family of measures of effect size and present a new generalized version of eta squared that attempts to define an effect size measure that can be compared across multiple research designs. Bakefield (2005) recommends the use of generalized eta squared (η^2_G*)* rather than partial eta squared (η^2_P) or omega squared (ω^2) for repeated-measures designs.

Eta squared *(η^2)* is a measure of the proportion of variance explained by an effect. Eta squared is calculated for an effect by dividing the sum of squares for the factor by the total sum of squares. Eta squared has problems with certain research designs due to the use of the total sum of squares as the denominator that limits the comparability of effect size across different research designs. For example, given a within-subjects analysis of a factor and a between-subjects analysis of a factor, eta squared may not be comparable for the two experiments.

Partial eta squared (η^2_P) is an improvement over eta squared that is calculated as the effect sum of squares divided by sum of the effect sum of squares and the subjects-within-cells sum of squares. Conveniently, partial eta squared is the effect size estimate provided by SPSS. However, Olejnik and Algina (2003) indicate that partial eta squared loses its ability to be compared across designs if the design includes a blocking factor.

Generalized eta squared (η^2_G*)* was developed by Olejnik and Algina (2003) to respond to the issue of comparability across research designs. Generalized eta squared's computation is based on the assumption that there are two sources of variance in data: First, the factors manipulated in the study introduce variance. Second, the individual differences of the subjects introduce variance. Generalized eta squared is calculated as the effect sum of squares divided by the sum of the effect sum of squares and the sum of the measured sums of squares for the study. Bakeman (2005) provides tables outlining computation of partial eta squared and generalized eta squared for various designs. Table E.1 demonstrates the calculation of partial eta squared and generalized eta squared using the repeated-measures analysis of catch trial questions including inaccurate participants from Table 4.2 as an example.

Table E.1

Calculation of Partial Eta Squared and Generalized Eta Squared

	SS	η^2_P Calculation	η^2_P	η^2_G Calculation	η^2_G
Order (O)	.005	$SS_O/(SS_O + SS_{s/OI})$	.001	$SS_O/(SS_O + SS_{s/OI} + SS_{Ss/OI})$	.000
Instruction (I)	.081	$SS_I/(SS_I + SS_{s/OI})$	.010	$SS_I/(SS_I + SS_{s/OI} + SS_{Ss/OI})$	.006
O × I	.143	$SS_{OI}/(SS_{OI} + SS_{s/OI})$	.018	$SS_{OI}/(SS_{OI} + SS_{s/OI} + SS_{Ss/OI})$	.010
Error (s/OI)	7.952				
Scenario (S)	.336	$SS_S/(SS_S + SS_{Ss/OI})$	.049	$SS_S/(SS_S + SS_{s/OI} + SS_{Ss/OI})$	.023
S × O	.001	$SS_{SO}/(SS_{SO} + SS_{Ss/OI})$	.000	$SS_{SO}/(SS_{SO} + SS_{s/OI} + SS_{Ss/OI})$	.000
S × I	.033	$SS_{SI}/(SS_{SI} + SS_{Ss/OI})$	.005	$SS_{SI}/(SS_{SI} + SS_{s/OI} + SS_{Ss/OI})$	.002
S × O × I	.166	$SS_{SOI}/(SS_{SOI} + SS_{Ss/OI})$	.024	$SS_{SOI}/(SS_{SOI} + SS_{s/OI} + SS_{Ss/OI})$	.011
Error (Ss/OI)	6.592				

APPENDIX F

ADDITIONAL STATISTICS

Basic Event Prediction

Inaccurate Participants

ANOVA including inaccurate participants. A repeated-measures ANOVA with one within-subjects factor (scenario) and two between-subjects factors (presentation order and instruction condition) was performed to investigate the effect of presentation order, scenario, and instruction condition on basic event prediction response accuracy. All participants were included in the analysis. Table F.1 lists the results of the repeated-measures ANOVA.

Table F.1

Repeated-Measures Analysis of Variance of Mean Response Accuracy on Basic Event Prediction, Including Inaccurate Participants (N = 183)

Source	*df*	*F*	η^2_P	η^2_G	*p*
			Between Subjects		
Order (O)	1	.514	.003	.002	.474
Instruction (I)	5	11.612	.253	.197	.000*
O × I	5	.657	.019	.014	.656
S within-group error	171	(.047)			
			Within Subjects		
Scenario (S)	1	11.20	.061	.025	.001*
S × O	1	4.243	.024	.010	.041*
S × I	5	.655	.019	.007	.658
S × O × I	5	.477	.014	.005	.793
S × *S* within-group error	171	(.030)			

Note. Values in parentheses indicate mean square errors. *S* = subjects.
*$p < .05$

The main effect of presentation order was not significant. The main effect of instruction condition, *F* (5, 171) = .378, *p* = .863, η^2_G = .197*,* was significant because the scoring in this analysis did not account for the alternative hypothesis instruction condition's additional hypothesis. The interaction between presentation order and instruction condition was not significant. Figure F.1 depicts the estimated mean response accuracy by instruction condition.

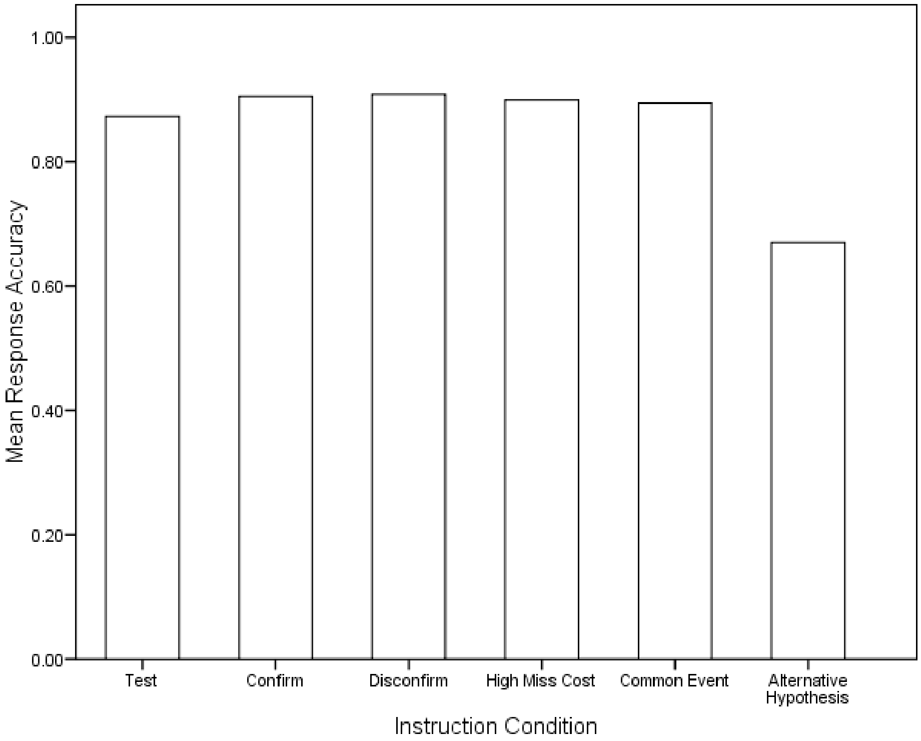


Figure F.1. Estimated overall mean response accuracy on basic event prediction questions by instruction condition, including inaccurate participants.

A Tukey's post-hoc analysis of the main effect of the instruction condition revealed that participants in the alternative hypothesis instruction condition had a lower estimated mean response accuracy (*M* = 67.0%, SE = .02) on the basic event predictions than the other instruction conditions (*p* < .001 for all comparisons). The lower response accuracy was due to the response accuracy score failing to account for the additional hypothesis given in the alternative hypothesis instructions.

The main effect of scenario was significant, *F* (1, 171) = 14.933, *p* < .001, η^2_G = .025. There was also an interaction effect between scenario and presentation order, *F* (1, 171) = 9.99, *p* = .002, η^2_G = .01. The interaction between scenario and instruction condition and the three-way interaction between scenario, instruction condition, and presentation order were not significant. Figure F.2 graphically depicts the Scenario × Presentation Order effect.

The Scenario × Presentation Order interaction indicates that basic event prediction for the MDT Microworld scenario and the Scorpion-Spider scenario depends on the order of presentation for the scenarios. Examining Figure F.2, there appears to be a possible learning effect (higher accuracy on the second scenario) for participants that received the Scorpion-Spider scenario first. Multiple pairwise comparisons using the Sidak adjustment were performed to investigate the simple effects of the Scenario × Presentation Order interaction. The analysis of the simple effect of scenario for the participants that received the Scorpion-Spider scenario first verified that response accuracy did increase significantly from the first scenario (Scorpion-Spider, *M* = 81.5%, *SD* = .23) to the second scenario (MDT Microworld, *M* = 91.4%, *SD* = .16).

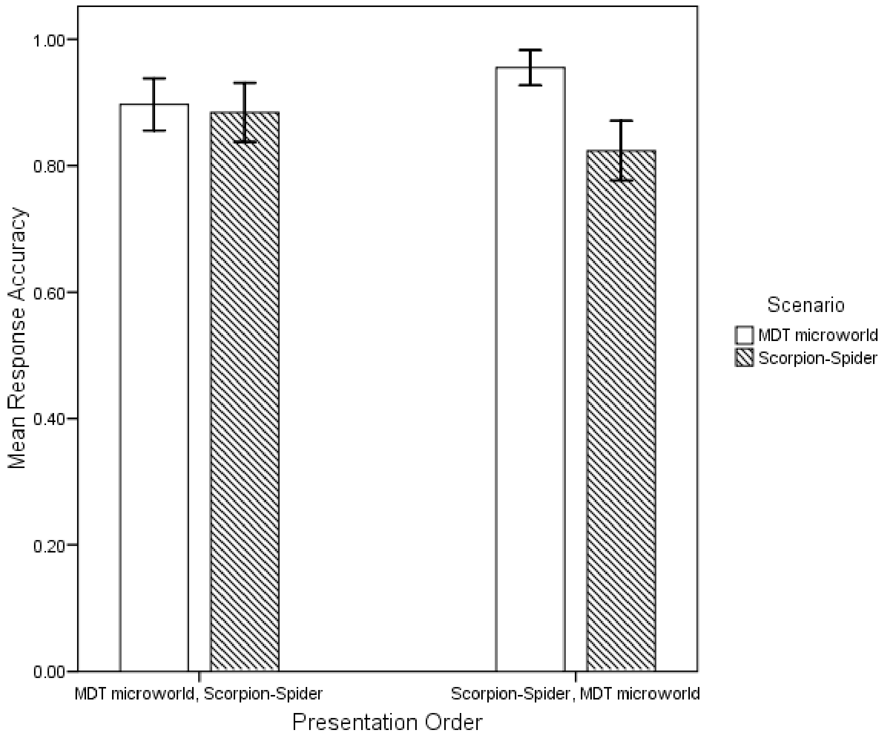


Figure F.2. Mean response accuracy on basic event prediction questions for presentation order by scenario interaction including inaccurate participants. Error bars indicate 95% confidence interval.

The analysis of the simple effect of presentation order for the MDT Microworld scenario indicates that response accuracy on the basic event prediction questions for the MDT Microworld scenario is lower for participants that receive that scenario first (*M* = 89.7%, *SD* = .19) than for participants that receive the Scorpion-Spider scenario first (*M* = 95.5%*, SD* = .13, *p* = .03). The presentation order effect for the Scorpion-Spider scenario was not significant (*p* = .523). The scenario main effect is a result of the significantly higher performance on the MDT Microworld scenario by the Scorpion-Spider, MDT Microworld presentation order.

In the analysis of catch trials, it was noted that most of the subjects dropped from the analysis were inaccurate in the Scorpion-Spider scenario. If the same participants were also inaccurate on the basic event prediction questions (possibly due to a misunderstanding similar to the confusion over scorpion claws seen in the pilot study), the inaccuracy on the Scorpion-Spider scenario could explain the Scenario × Presentation Order interaction and scenario main effect.

ANOVA of mean response accuracy, excluding inaccurate participants. A second repeated-measures ANOVA with one within-subjects factor (scenario) and two between-subjects factors (presentation order and instruction condition) was performed excluding the inaccurate participants from the analysis. The results of the ANOVA are listed in Table F.2.

Excluding the inaccurate participants did not significantly change the results of the between-subjects effects. The main effect of presentation order and the interaction between presentation order and instruction condition were not significant. The main effect of instruction condition was significant, F (5, 147) = 20.094, $p < .001$, $\eta^2_G = .302$. In Figure F.3, as in the previous analysis, the mean response accuracy for participants in the alternative hypothesis instruction condition is lower than the mean response accuracy for the other instruction conditions. This instruction condition effect is again due to the additional hypothesis given in the alternative hypothesis instruction condition (see Chapter IV for analysis of hypothesis adjusted mean response accuracy).

Table F.2

Repeated-Measures Analysis of Variance of Mean Response Accuracy on Basic Event Prediction, Excluding Inaccurate Participants (n = 159)

Source	*df*	*F*	η^2_P	η^2_G	*p*
			Between Subjects		
Order (O)	1	.216	.001	.001	.643
Instruction (I)	5	20.094	.406	.302	$.000^*$
O × I	5	.467	.016	.010	.800
S within-group error	147	(.029)			
			Within Subjects		
Scenario (S)	1	.512	.003	.001	.476
S × O	1	5.192	.034	.013	$.024^*$
S × I	5	1.235	.040	.015	.296
S × O × I	5	1.830	.059	.022	.110
S × *S* within-group error	147	(.017)			

Note. Values in parentheses indicate mean square errors. *S* = subjects.
$^*p < .05$

In the new analysis excluding the inaccurate participants, the main effect of scenario was not significant. The interaction between scenario and instruction condition and the three-way interaction between scenario, presentation order, and instruction condition were not significant. The interaction between scenario and presentation order was still significant, *F* (1, 147) = 5.192, *p* = .024, η^2_G = .013.

The possibility of a small learning effect where response accuracy improves on the second scenario remains (see Figure F.4). Multiple pairwise comparisons with the Sidak adjustment were performed to investigate the simple effects of the Scenario × Presentation Order interaction. For the participants in the Scorpion-Spider, MDT Microworld presentation order, mean response accuracy did increase from the first scenario, Scorpion-Spider (*M* = 87%, *SD* = .19), to the second scenario, the MDT

Microworld (*M* = 91.4%, *SD* = .16, *p* = .035). However, while participants that received the MDT Microworld first did have a slight increase in mean response accuracy from the first scenario, MDT Microworld (*M* = 87.3%, *SD* = .18), to the second scenario, Scorpion-Spider (*M* = 89.6%, *SD* = .16), the difference was not significant (*p* = .272).

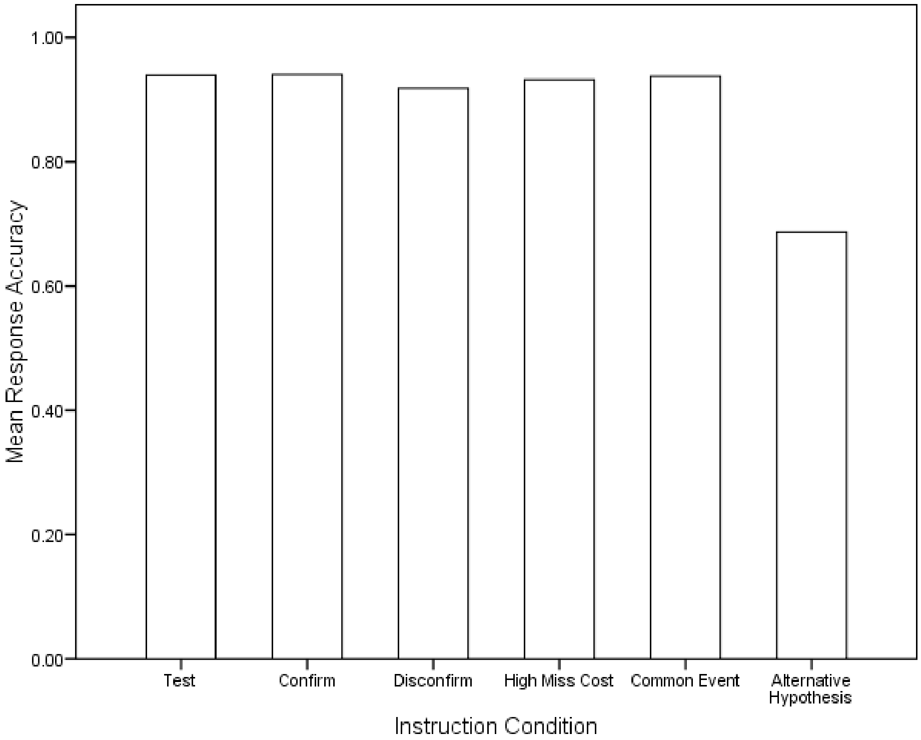


Figure F.3. Estimated overall mean response accuracy on basic event prediction questions for instruction condition excluding inaccurate participants.

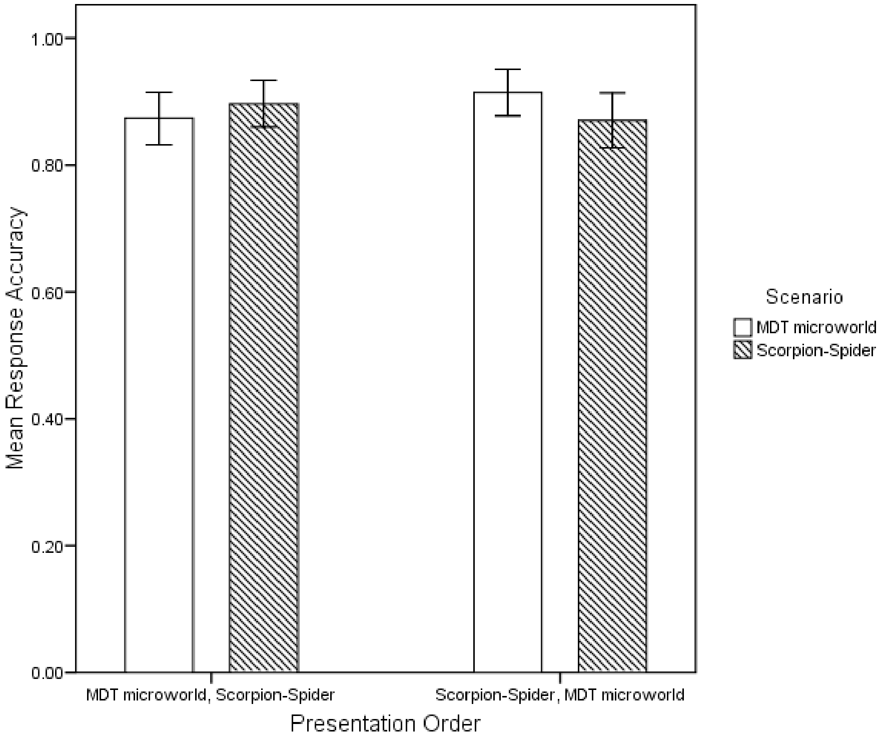


Figure F.4. Mean response accuracy on basic event prediction questions for presentation order by scenario interaction excluding inaccurate participants. Error bars indicate 95% confidence interval.

In summary, removing the inaccurate participants from the analysis of response accuracy for the basic event prediction questions removed a spurious main effect of scenario. The increased response accuracy in the second scenario revealed by the interaction between scenario and presentation order suggests a possible learning effect. The absence of a clear learning effect for both presentation orders indicates that starting with the Scorpion-Spider scenario was worse for performance than starting with the MDT Microworld scenario for the basic event prediction questions. The possibility of a learning effect is not particularly relevant to the primary purpose of this research.

Although the significantly lower mean response accuracy of participants in the alternative hypothesis instruction condition is relevant, the alternative hypothesis condition's additional hypothesis (HA) was not accounted for in this analysis. The participants in the alternative hypothesis instruction condition were re-scored to account for the alternative hypothesis (See Chapter IV for the analysis of the hypothesis adjusted mean response accuracy).

Positivity Assessment

Inaccurate Participants

ANOVA including inaccurate participants. A repeated-measures ANOVA with one within-subjects factor (scenario) and two between-subjects factors (presentation order and instruction condition) was performed to investigate the effect of presentation order, scenario, and instruction condition on the positivity assessment deviation from the expected values. Table 4.10 lists the results of the ANOVA.

The main effect of instruction was significant, F (5,171) = 9.675, $p < .001$, $\eta^2_G =$ .149. The main effect of order and the interaction between order and instruction were not significant. A Tukey's HSD post-hoc analysis of the main effect of instruction revealed that the mean deviation from the expected values for the participants in the alternative hypothesis instruction condition was significantly larger ($p < .001$) than the mean deviation of the other instruction conditions. The increased error for the alternative hypothesis group was due to the additional hypothesis under consideration during positivity assessment.

Table F.3

Repeated-Measures Analysis of Variance of Mean Deviation of Positivity Assessment from Expected Values Given H1, Including Inaccurate Participants (N = 183)

Source	*df*	*F*	η^2_P	η^2_G	*p*
			Between Subjects		
Order (O)	1	.537	.003	.002	.465
Instruction (I)	5	9.675	.221	.149	.000*
O × I	5	.418	.012	.007	.836
S within-group error	171	(.031)			
			Within Subjects		
Scenario (S)	1	2.607	.015	.006	.108
S × O	1	8.291	.046	.018	.004*
S × I	5	1.117	.032	.012	.353
S × O × I	5	.501	.014	.006	.775
S × *S* within-group error	171	(.019)			

Note. Values in parentheses indicate mean square errors. *S* = subjects.
**p* < .05

The main effect of scenario was not significant. There was a significant interaction between scenario and presentation order, *F* (1, 171) = 8.291, *p* = .004, η^2_G = .018. The interaction between scenario and instruction and the three-way interaction between scenario, instruction, and presentation order were not significant. Figure F.5 graphically depicts the Scenario × Presentation Order effect.

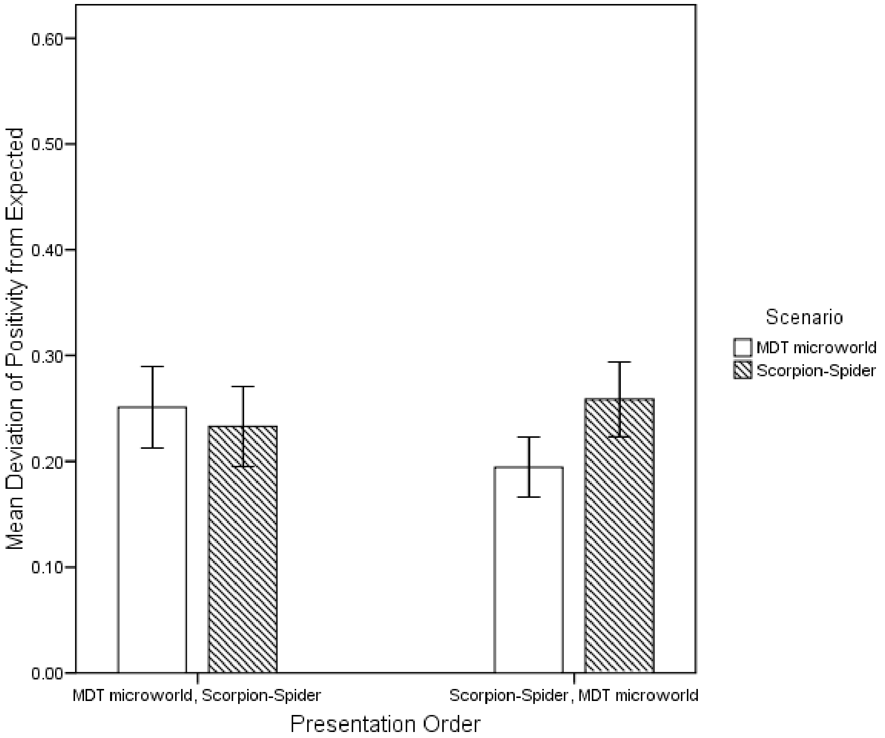


Figure F.5. Mean deviation of positivity assessment from expected values given the primary hypothesis (H1) for scenario by presentation order, including inaccurate participants. Error bars indicate 95% confidence interval.

Participants' positivity assessments were closest to the expected values given H1 on the second scenario (MDT Microworld; *M* = .1944, *SD* = .1375) performed by the Scorpion-Spider, MDT Microworld presentation order group. Again, this suggests a learning effect when shifting from the Scorpion-Spider scenario to the MDT Microworld scenario. The overall greater deviation from the expected values for the Scorpion-Spider scenario may have been due to the inclusion of the inaccurate subjects.

ANOVA excluding inaccurate participants. Given the results including the inaccurate participants, a new analysis was performed excluding the inaccurate participants from the analysis. Table F.4 lists the results of the ANOVA.

Table F.4

Repeated-Measures Analysis of Variance of Deviation of Positivity Assessment from Expected Values Given H1, Excluding Inaccurate Participants (n = 159)

Source	*df*	*F*	η^2_P	η^2_G	*p*
			Between Subjects		
Order (O)	1	.282	.002	.001	.596
Instruction (I)	5	16.133	.354	.248	$.000^*$
O × I	5	.836	.028	.017	.526
S within-group error	147	(.021)			
			Within Subjects		
Scenario (S)	1	.464	.003	.001	.497
S × O	1	13.107	.082	.034	$.000^*$
S × I	5	.805	.027	.011	.548
S × O × I	5	.876	.029	.012	.499
S × *S* within-group error	147	(.014)			

Note. Values in parentheses indicate mean square errors. *S* = subjects.
$^*p < .05$

The main effect of instruction condition was significant, *F* (5,147) = 16.133, $p < .001$, $\eta^2_G = .248$. The main effect of presentation order and the interaction between presentation order and instruction condition were not significant. As in the previous analysis of positivity assessment, Tukey's HSD post-hoc analysis of the main effect of instruction revealed the alternative hypothesis group positivity assessment deviated significantly more ($p < .001$) from the expected positivity assessments when the expected values were based on the primary hypothesis. The increased deviation for the alternative hypothesis group was likely due to the additional hypothesis under consideration during assessment that was not accounted for in this analysis.

The only significant within-subjects effect was the interaction between scenario and presentation order was significant, F (1, 147) = 13.107, $p < .001$, $\eta^2_G = .034$. Figure F.6 graphically depicts the Scenario × Presentation Order effect.

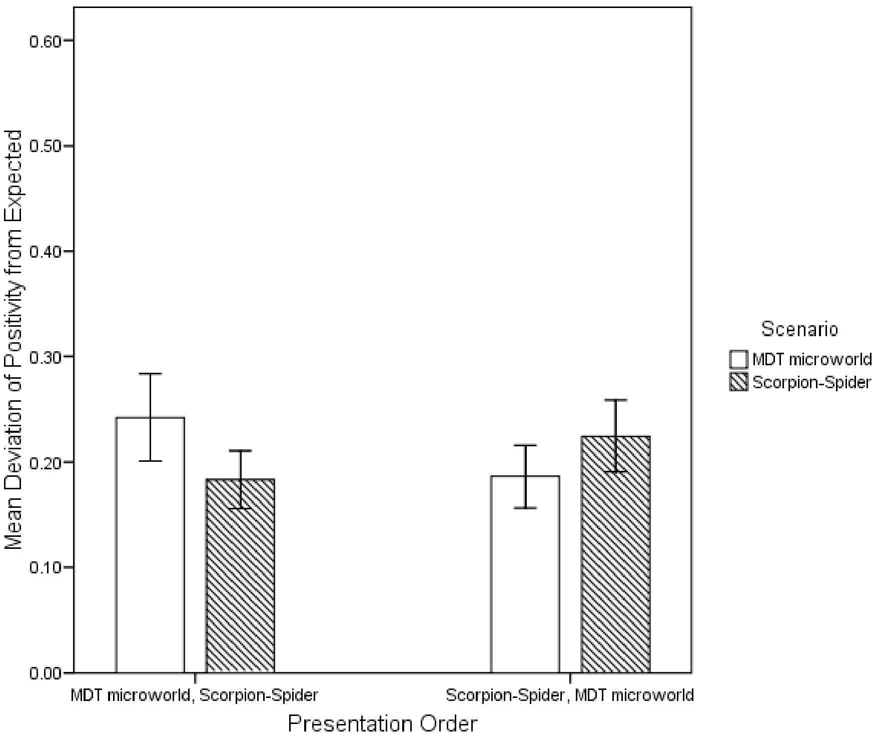


Figure F.6. Mean deviation of positivity assessment from expected values given the primary hypothesis (H1) for scenario by presentation order, excluding inaccurate participants. Error bars indicate 95% confidence interval.

Participants' positivity assessments were closest to the expected values for H1 for the second scenario regardless of the content of the scenario. For the MDT Microworld, Scorpion-Spider presentation order, the mean positivity assessments for the second scenario (Scorpion-Spider scenario; *M* = .1833, *SD* = .1215) were closer to the expected values than for the first scenario (MDT Microworld scenario; *M* = .2421, *SD* = .1844). Likewise, for the Scorpion-Spider, MDT Microworld presentation order, the mean positivity assessments for the second scenario (MDT Microworld scenario; *M* = .1863, *SD* = .1336) were closer than for the first scenario (Scorpion-Spider scenario; *M* = .2245, *SD* = .1534).

Removing the inaccurate participants resulted in a more consistent pattern of results where mean deviation from the expected values for positivity assessment was reduced in the second presented scenario. Because the expected positivity assessment values were set at the extreme levels (0% or 100%), these results suggest that participants were more extreme in their positivity assessments in the second scenario. This may have reflected increasing confidence in their assessments from the first scenario to the second scenario.

Explicit Positive Test Selection

Inaccurate Participants

ANOVA including inaccurate participants. A repeated-measures ANOVA was performed to investigate the effect of presentation order, scenario, and instruction condition on the deviation of explicit positive test selection from the expected values. Table F.5 lists the results of the ANOVA.

Table F.5

Repeated-Measures Analysis of Variance of Mean Deviation from Expected Values for Explicit Positive Test Selection Questions, Including Inaccurate Participants (N = 183)

Source	*df*	*F*	η^2_P	η^2_G	*p*
			Between Subjects		
Order (O)	1	10.267	.057	.027	.002*
Instruction (I)	5	.714	.020	.010	.614
O × I	5	1.971	.054	.026	.085
S within-group error	171	(.007)			
			Within Subjects		
Scenario (S)	1	154.513	.475	.327	.000*
S × O	1	.330	.002	.001	.567
S × I	5	2.657	.072	.04	.024*
S × O × I	5	1.266	.036	.02	.281
S × *S* within-group error	171	(.008)			

Note. Values in parentheses indicate mean square errors. *S* = subjects.
$*p < .05$

The main effect of presentation order was significant, *F* (1, 171) = 10.267, *p* = .002. The main effect of instruction condition was not significant and the interaction between presentation order and instruction condition was also not significant. An examination of the means for the two presentation order groups reveals that participants that received the MDT Microworld first had a significant preference for top figures (*M* = -.02) compared to participants that received the Scorpion-Spider scenario first, who had no preference for top or bottom figures (*M* = .008).

The main effect of scenario was highly significant, *F* (1, 171) = 154.513, *p* < .001 with a large estimated effect size (η^2_G = .327). The interaction between scenario and presentation order was not significant. There was a significant effect for the interaction between scenario and instruction condition, *F* (5, 171) = 2.657, *p* = .024. The three-way interaction between scenario, presentation order, and instruction condition was not significant. Figure F.7 depicts the interaction between scenario and instruction condition.

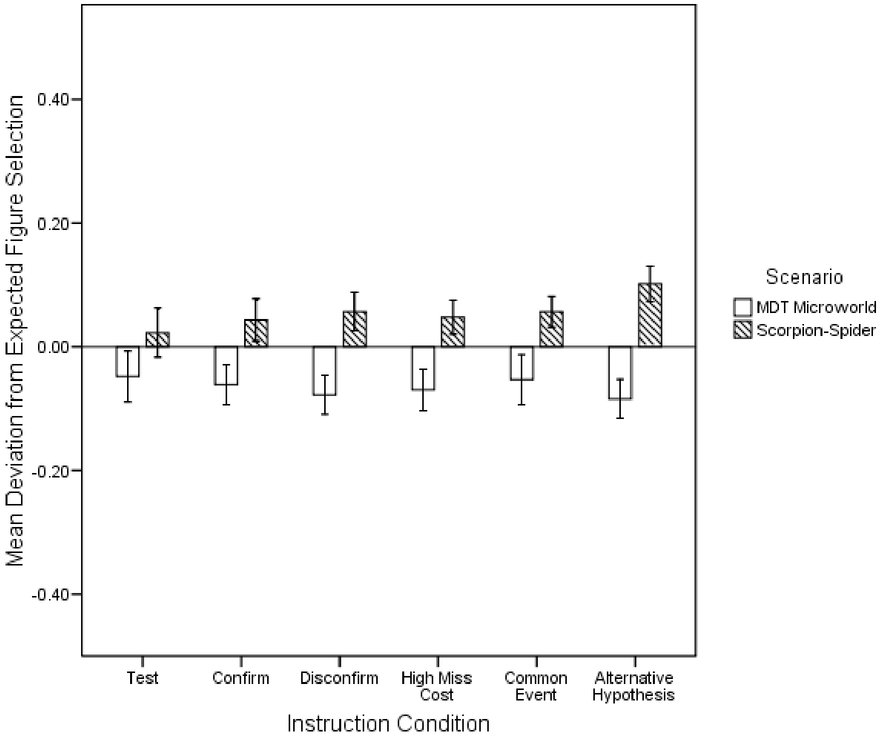


Figure F.7. Mean deviation from expected figure selection given the primary hypothesis (H1) for scenario by instruction condition. Error bars indicate 95% confidence interval.

A multiple comparisons analysis using the Sidak adjustment was performed to determine the nature of the interaction. A single instruction condition, the alternative hypothesis group, was significantly different from only one other instruction condition, the test condition. The difference was only significant in the Scorpion-Spider scenario.

However, all of the instruction conditions differed in the *sign* (positive or negative) of the mean deviation. In the MDT Microworld, the mean deviation is negative

(*M* = -.066, *SD* = .096). In the Scorpion-Spider scenario, the mean deviation is positive (*M* = .055, *SD* = .087). The direction of deviation indicates an effect of figure position on test selection (positive = bottom figure, negative = top figure). There was no difference in the number of predicted positive and negative figures between the scenarios so the source of this effect is difficult to discern. As in the positive assessment questions, some additional factor was affecting participant selections for the explicit positive test selection questions.

ANOVA excluding inaccurate participants. A new analysis was performed excluding the inaccurate participants from the analysis. The results of the repeated-measures ANOVA are listed in Table F.6.

Dropping the inaccurate participants made no significant difference in the analysis of explicit positive test selection. As in the analysis including the inaccurate participants, there was a main effect of presentation order, F (1, 147) = 14.012, p < .001, η^2_G = .042. The main effect of instruction condition and the interaction between presentation order and instruction condition were not significant. An examination of the means again reveals that the participants that received the MDT Microworld first selected significantly more top figures than predicted (M = -.025) and the participants that received the Scorpion-Spider scenario first did not appear to have an overall preference for top or bottom figures (M = .008).

Table F.6

Repeated-Measures Analysis of Variance of Mean Deviation from Expected Values for Explicit Positive Test Selection Questions, Excluding Inaccurate Participants (n = 159)

Source	*df*	*F*	η^2_P	η^2_G	*p*
			Between Subjects		
Order (O)	1	14.012	.087	.042	.000*
Instruction (I)	5	1.814	.058	.027	.113
O × I	5	1.898	.061	.029	.098
S within-group error	147	(.006)			
			Within Subjects		
Scenario (S)	1	178.28	.548	.398	.000*
S × O	1	.522	.004	.002	.471
S × I	5	3.118	.096	.055	.011*
S × O × I	5	.690	.023	.013	.632
S × *S* within-group error	147	(.007)			

Note. Values in parentheses indicate mean square errors. *S* = subjects.
**p* < .05

There was a main effect of scenario, *F* (1, 147) = 178.28, *p* < .001, η^2_G = .398. There was again an interaction between scenario and instruction condition, *F* (5, 147) = 3.118, *p* = .011. The interaction between scenario and presentation order was not significant and the three-way interaction between scenario, presentation order, and instruction condition was also not significant. Figure F.8 depicts the interaction between scenario and instruction condition.

Multiple pairwise comparisons using the Sidak adjustment were performed to determine the nature of the interaction. With the inaccurate participants dropped from the analysis, the alternative hypothesis instruction condition (*M* = -.093, *SD* = .0698) was significantly different from two other instruction condition, the test condition (*M* = .009, *SD* = .1128, *p* < .001) and the confirm condition (*M* = .043, *SD* = 0803, *p* = .036) for only the Scorpion-Spider scenario.

The main effect of scenario is also clear in Figure F.8. Dropping the inaccurate participants had no significant effect on the difference between the figure selections in the two scenarios. In the MDT Microworld, participants tend to select more top screens than predicted (*M* = -.073, *SD* = .086). In the Scorpion-Spider scenario, participants tend to select more bottom screens than predicted (*M* = .057, *SD* = .085).

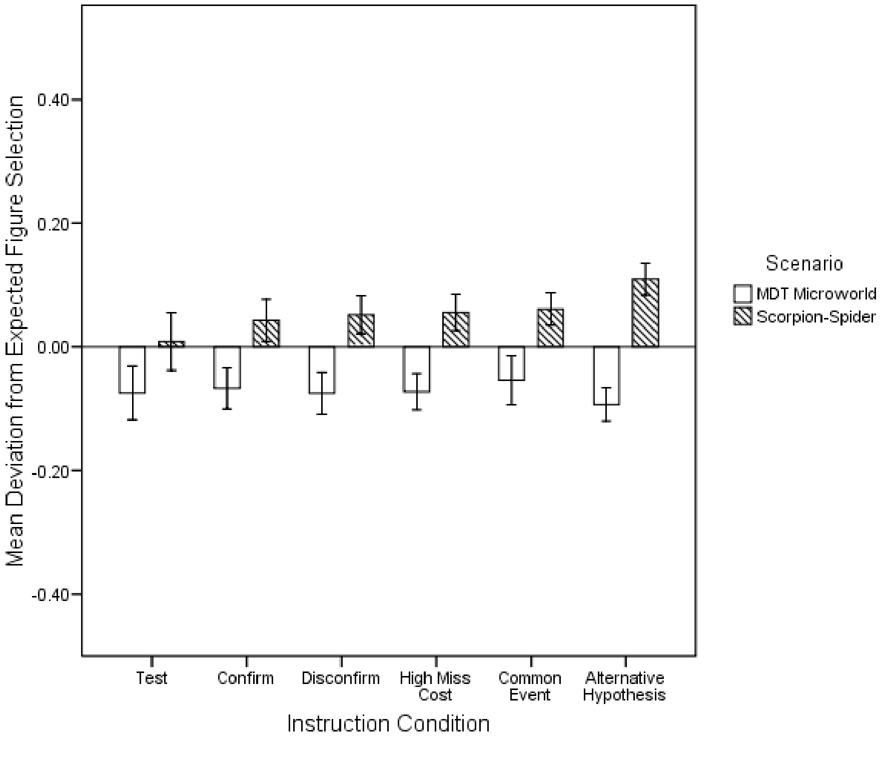


Figure F.8. Mean deviation from expected figure selection given the primary hypothesis (H1) for scenario by instruction condition. Error bars indicate 95% confidence interval.

Hypothesis Test Selection

Inaccurate Participants

ANOVA including inaccurate participants. A repeated-measures ANOVA was performed to investigate the effect of presentation order, scenario, and instruction condition on the proportion of positive tests selected in the hypothesis test selection questions. Table F.7 lists the results of the ANOVA.

Table F.7

Repeated-Measures Analysis of Proportion of Positive Tests Selected for Hypothesis Test Selection Questions, Including Inaccurate Participants (N = 183)

Source	*df*	*F*	η^2_P	$\eta^2_G{}^a$	*p*
			Between Subjects		
Order (O)	1	.015	.000	.000	.901
Instruction (I)	5	2.120	.058	.042	.065
O × I	5	.729	.021	.015	.602
S within-group error	171	(.083)			
			Within Subjects		
Scenario (S)	1	4.485	.026	.008	.036*
S × O	1	1.848	.011	.003	.176
S × I	5	1.036	.029	.009	.398
S × O × I	5	.298	.009	.003	.914
S × *S* within-group error	171	(.008)			

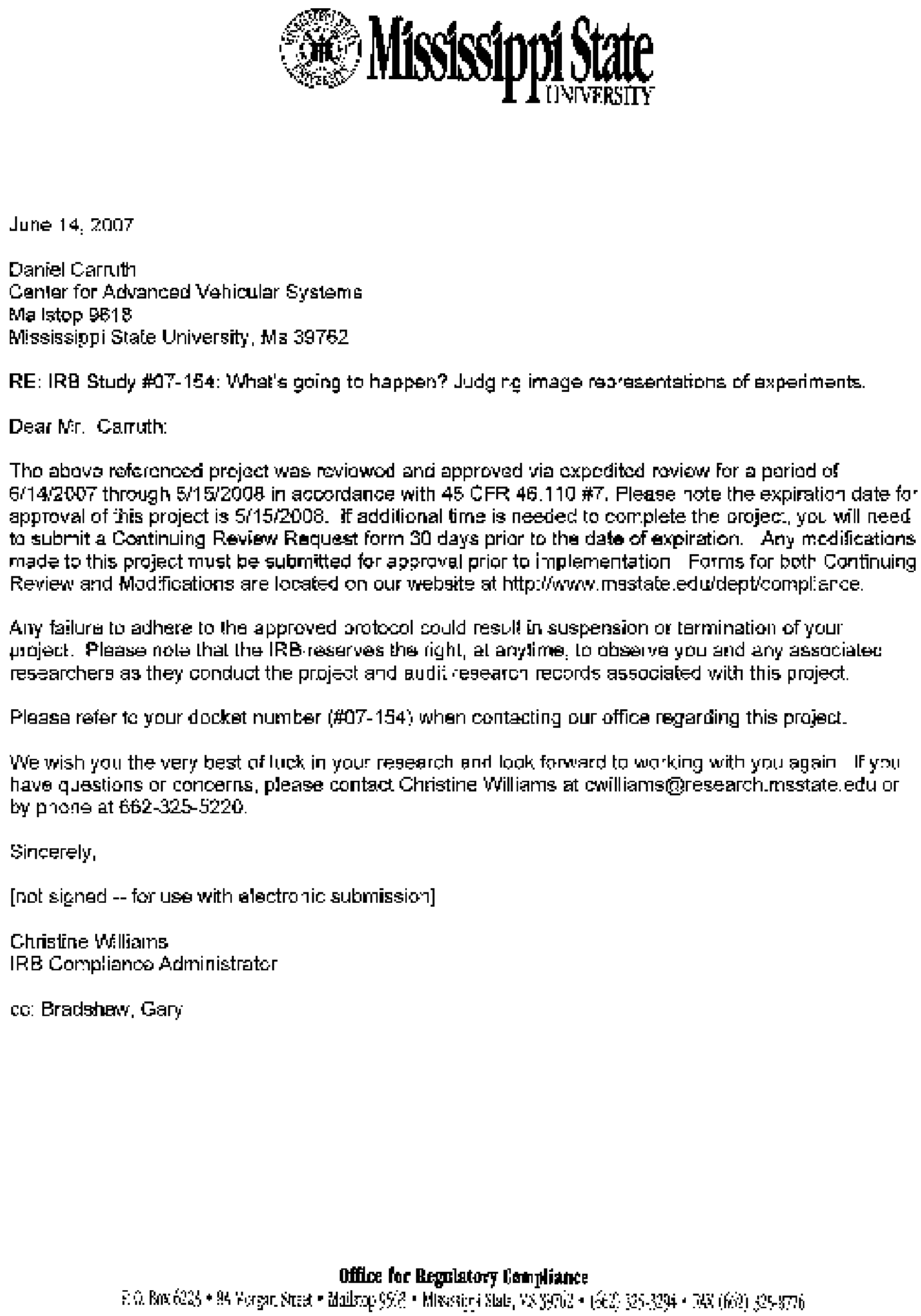
Note. Values in parentheses indicate mean square errors. *S* = subjects.
*p < .05

The main effect of presentation order, the main effect of instruction condition, and the interaction between presentation order and instruction condition, were all not significant.

The main effect of scenario was significant, *F* (1, 171) = 4.485, *p* = .036, η^2_G = .008, with a very small estimated effect size. The interaction between scenario and order, between scenario and instruction condition, and between scenario, presentation order, and instruction condition were all not significant. An examination of the means for the two scenarios reveals that participants in the Scorpion-Spider scenario (*M* = 81%, *SD* = .26) had a significantly lower proportion of positive tests selected than the participants in the MDT Microworld (*M* = 85%, *SD* = .22). Although this suggests a possible effect of thematic content, previous analyses have demonstrated that the inaccurate participants may have an impact on scenario effects. See Chapter IV for analyses of hypothesis test selection questions excluding inaccurate participants.

APPENDIX G

IRB APPROVAL FORM



Mississippi State University

June 14, 2007

Daniel Carruth
Center for Advanced Vehicular Systems
Mailstop 9618
Mississippi State University, Ms 39762

RE: IRB Study #07-154: What's going to happen? Judging image representations of experiments.

Dear Mr. Carruth:

The above referenced project was reviewed and approved via expedited review for a period of 6/14/2007 through 5/15/2008 in accordance with 45 CFR 46.110 #7. Please note the expiration date for approval of this project is 5/15/2008. If additional time is needed to complete the project, you will need to submit a Continuing Review Request form 30 days prior to the date of expiration. Any modifications made to this project must be submitted for approval prior to implementation. Forms for both Continuing Review and Modifications are located on our website at http://www.msstate.edu/dept/compliance.

Any failure to adhere to the approved protocol could result in suspension or termination of your project. Please note that the IRB reserves the right, at anytime, to observe you and any associated researchers as they conduct the project and audit research records associated with this project.

Please refer to your docket number (#07-154) when contacting our office regarding this project.

We wish you the very best of luck in your research and look forward to working with you again. If you have questions or concerns, please contact Christine Williams at cwilliams@research.msstate.edu or by phone at 662-325-5220.

Sincerely,

[not signed -- for use with electronic submission]

Christine Williams
IRB Compliance Administrator

cc: Bradshaw, Gary

Office for Regulatory Compliance
P.O. Box 6223 • 84 Morgan Street • Mailstop 9563 • Mississippi State, MS 39762 • (662) 325-3294 • FAX (662) 325-8776