INT. J. SCI. EDUC., 19 NOVEMBER 2004,
VOL. 26, NO. 14, 1715–1731

Taylor & Francis
Taylor & Francis Group

**RESEARCH REPORT**

# Reasoning Maps: a generally applicable method for characterizing hypothesis-testing behaviour

*Brian White, University of Massachusetts Boston, Department of Biology, 100 Morrissey Boulevard, Boston, MA 02125, USA; e-mail: brian.white@umb.edu*

This paper presents a generally applicable method for characterizing subjects' hypothesis-testing behaviour based on a synthesis that extends on previous work. Beginning with a transcript of subjects' speech and videotape of their actions, a Reasoning Map is created that depicts the flow of their hypotheses, tests, predictions, results, and conclusions. The methods are described and then applied to a group of three undergraduate biology students testing hypothesis in an inquiry-based laboratory exercise, the Red and White Yeast Lab. Analysis of hypothesis-testing behaviour via Reasoning Maps reveals most of the features explored in previous studies in a unified context. In addition, Reasoning Maps allow analysis of higher-order patterns in hypothesis testing that are not possible using existing methods. We have designed these methods so that they will provide a common language for analysing and understanding hypothesis testing that will allow global comparisons of behaviour.

## Introduction

### *Hypothesis testing*

Hypothesis generation and testing are central to the scientific process. Because of this, many science education reform documents call for science students to learn this skill; for example, the National Research Council's (1996) National Science Education Standards call for students to know how to 'Design and conduct scientific investigations', and Bransford et al. (1999) strongly encourage science classroom activities where '… students design studies, collect information, analyze data, construct evidence, and then debate the conclusions that *they* derive from their evidence' (p. 171). In order to properly design curricula to teach this skill, it is necessary to have a detailed understanding of how students understand and perform hypothesis testing. This knowledge can then be used to design and evaluate appropriate educational interventions. A wide variety of techniques have been developed for this analysis. However, most of these techniques are best suited for the analysis of one particular task or the examination of a few particular features of hypothesis-testing behaviours. No methods exist that allow comparisons across all tasks and features. In order to facilitate the analysis of hypothesis testing in a variety of contexts, we have developed a technique for characterizing hypothesis-testing behaviour based on protocol analysis (Ericsson Simon 1984) called Reasoning Maps. In developing the reasoning maps, we drew upon work by philosophers of science as well as the extensive literature investigating hypothesis testing. Our goal was to create a system of analysis that incorporated the most widely-applicable features identified previously.

*International Journal of Science Education* ISSN 0950–0693 print/ISSN 1464–5289 online ©2004 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/0950069042000243718

The philosopher of science, Carl Hempel, uses Semmelweis' 1847 demonstration that childbed fever was caused by 'cadaveric material' transferred by medical students to mothers during childbirth as an example of the process of scientific inquiry (Hempel 1966). The process begins with a hypothesis; for example, 'childbed fever is caused by cadaveric material'. Tests of the hypothesis are then devised. These take the form of 'if … then' statements like 'If childbed fever is caused by cadaveric material, then requiring medical students to disinfect their hands before examining women in labor will reduce the incidence of childbed fever'. Hempel calls these statements 'test implications' (1966:7). The tests are then conducted and the results either support or refute the hypothesis. Studies of hypothesis testing in non-scientists have further divided this process into five components: hypothesis, test, prediction, result, the conclusion. In the section that follows we have selected studies that illustrate the range of tasks and types of analysis from the extensive literature on hypothesis testing. Although none of the studies address all of these in detail, this set emerges from the union of the previous work. In detail, these five components are:

- *Hypothesis*. Hempel calls hypotheses '… guesses at the connections that might obtain between the phenomena under study, at uniformities and patterns that might underlie their occurrence' (1966: 15). In studies of hypothesis testing, hypotheses take many forms and are subject to a wide range of constraints. In the most constrained situations, hypotheses are given in advance by the investigators, and the subjects' task is to evaluate them. In Tschirgi (1980), subjects were asked (for example) to choose the best experiment to determine whether adding honey to a cake recipe would improve its taste. Similarly, in Moshman (1979), students were asked how evidence reflected on the hypothesis: 'If a person uses fluoridated water, he will have healthy teeth' (p. 106). Less constrained studies required subjects to choose from a pre-defined set of hypotheses to investigate. Many of these involved determining which of a set of variables provided by the investigator has or does not have an effect on a particular outcome. For example, students were asked to determine which features of a computer-simulated car influenced its speed (Schauble 1990), which features of a boat influenced its speed (Schauble et al. 1991), or which environmental conditions in a computer-simulated environment influenced flooding levels (Kuhn et al. 2000). Kuhn and Phelps (1982) asked students to determine which of a set of chemicals produced a desired reaction. Similarly, Shute et al.(1989) asked students to determine the interactions between various economic factors in a computer-simulated town. Other tasks required subjects to solve a puzzle; Klahr (2000) asked subjects to determine the function of a particular program command on the behaviour of a robot, and in Newell and Simon (1972) a subject was asked to determine which numbers were represented by particular letters in an encoded arithmetic problem. Several studies (Kelly et al. 1998, Schauble et al. 1992) required students to identify the contents of disguised electrical components. The most unconstrained tasks approximated open-ended scientific inquiry. For example, in Lawson (2002), students generated hypotheses to explain why water rises in an enclosed space containing a burning candle. Similarly, in the studies of Mynatt, et al. (1978) and White and Frederiksen (1998), the hypotheses

were student-generated laws that attempted to explain the behaviour of a computer-simulated physical system.

- *Test*. Hempel describes these as part of the 'if … then' construction: 'If conditions of kind C are realized, then an event of kind E will occur' (1966: 19), and calls the conditions C an 'experimental test'. In studies of hypothesis testing, tests can be given in advance as part of the problem statement (Moshman 1979) or chosen from a pre-defined list (Tschirgi 1980), or may involve simple calculations (Newell and Simon 1972). In others, tests are conducted by altering parameters of a computer simulation (Klahr 2000, Mynatt et al. 1978, Schauble 1980, Shute et al. 1989, White and Frederiksen 1998); or by manipulations of physical objects: mixtures of chemicals (Kuhn and Phelps 1982), electrical circuits Kelly et al. 1998, Schauble et al. 1992, springs with weights and boats in canals (Schauble et al. 1991). or experiments involving candles in enclosed spaces (Lawson 2002).
- *Prediction*. This corresponds to the second part of Hempel's 'if … then' statement; the expected event E. In some studies (Mynatt et al. 1978, Newell and Simon 1972, Schauble et al. 1992, Shute et al. 1989), subjects could, and sometimes did, make predictions about the outcome of a particular test, while in other studies (Schauble et al. 1991), subjects were obliged to make specific predictions before conducting a test. Kuhn et al. (2000) went further by scoring predictions as correct or incorrect.
- *Result*. Hempel does not define these specifically, but instead refers to a variety of observations or experimental outcomes. In some studies, the results are given as part of the problem task (for example, Moshman 1979); in most other studies, they were generated in response to subject-generated tests. Results can take the form of output from a computer simulation: the speed of a car (Schauble 1990), responses of economic factors (Shute, et al. 1989), the movement of a robot (Klahr 2000), levels of flooding (Kuhn et al. 2000), or behaviour of objects in a computer simulation (Mynatt et al. 1978); or of the behaviour of materials manipulated by the subjects: the lighting of a light bulb in a simple circuit (Kelly et al. 1998, Schauble et al. 1992), or a chemical reaction (Kuhn and Phelps 1982), or water level in an enclosed space containing a candle (Lawson 2002).
- *Conclusions*. Hempel discusses how results inconsistent with a test implication can lead to the rejection of a hypothesis, while consistent results tend to support a hypothesis. In most cases examined, the correct hypothesis was known to investigators, so subjects' conclusions were scored as valid/correct or invalid/incorrect based on the investigators' knowledge of the correct answer. Several studies (Klayman and Ha 1987, Moshman 1979, Mynatt et al. 1978, Schauble 1990, Tschirgi 1980) also classified conclusions as positive (tending to confirm the hypothesis) or negative (tending to refute the hypothesis). In addition, some studies looked at the justification used for conclusions, requiring the subjects to cite a sufficient number of data points (Shute et al. 1989) or noting whether the conclusion was justified based on data or on prior beliefs (Schauble 1990).

*Connections between elements.* Some authors went on to characterize the connections between these elements and thus examine the flow of the process followed by the subjects. Schauble et al. (1991) described two different approaches to hypothesis

testing: 'an engineering model characterized by the more familiar goal of manipulating variables to produce a desired outcome', and a science model '[which] was associated with broader exploration, more selectiveness about evidence interpreted, and greater attention to establishing that some variables are not causal' (p. 859). Several studies examined the strategy of controlling variables (Kuhn and Phelps 1982, Kuhn et al. 2000, Shute et al. 1989, Tschirgi 1980) and the HOTAT ('hold one thing at a time') and VOTAT ('vary one thing at a time') approaches to experimental design. Mynatt et al. (1978) also looked at whether hypotheses were modified, temporarily abandoned, permanently abandoned, or retested following either confirmation or disconfirmation. Other studies examined this process through diagrams indicating each of the subject's moves. One type of diagram, the problem behaviour graph (Newell and Simon 1972) or student procedure graphs (Shute, et al. 1989) showed the particular tests performed and step-by-step changes in the state of the subject's knowledge. While each of these were explicitly temporal, others used diagrams that were not. Schauble et al. (1992) diagrammed students' processes as a flowchart of tests and decisions based on their results; a subject could pass through the same parts of a flowchart more than once. Finally, Kelly et al. (1998) used Toulmin's (1958) argument frame to organize individual hypotheses, test, prediction, result, and conclusion sequences.

All of these studies highlighted important issues in and relevant measures of hypothesis testing. In developing Reasoning Maps, we set out to create a scheme that incorporated all of the most common features and measures that had been used previously. Because Reasoning Maps are based on a synthesis of previous studies, virtually all of the previous analyses can be performed on a Reasoning Map of students' behaviour. In this way, our maps allow analyses like those performed previously as well as comparisons across different hypothesis-testing tasks that had not been possible with more locally applicable measures used previously.

### *The Red and White, Yeast Lab*

Analysis of hypothesis testing requires a context. Because our goal was to treat this in the most general context, we sought a hypothesis-testing task with as few constraints as possible. In addition, because wanted to produce a step-by-step characterization of the moves our subjects made, we required a task where the students leave an extensive verbal record of their thought processes. For these reasons, we chose to examine introductory-level undergraduate students' reasoning in the context of a particular inquiry-based biology laboratory exercise (White 1999). In this laboratory exercise, the Red and White Yeast Lab (RWYL), students invent their own hypotheses, design their own experiments, and evaluate their own data as they explore a biological phenomenon.

The RWYL is based on the biological phenomenon shown in figure 1. It is a patch of an engineered strain of bakers' yeast that has been grown for one week on solid medium on a petri dish. The centre of the patch is red and the outside edge is white. The students' task is to find out as much as they can about why the centre is red and the edge is white. The tools they are given are sterile toothpicks for taking samples of the yeast and fresh plates of nutrient medium on which to grow these samples. These tests are therefore re-creations and permutations of the original patch rather than examinations or dissections. When grown for a week, like the original patch, their samples produce results that can then be evaluated.

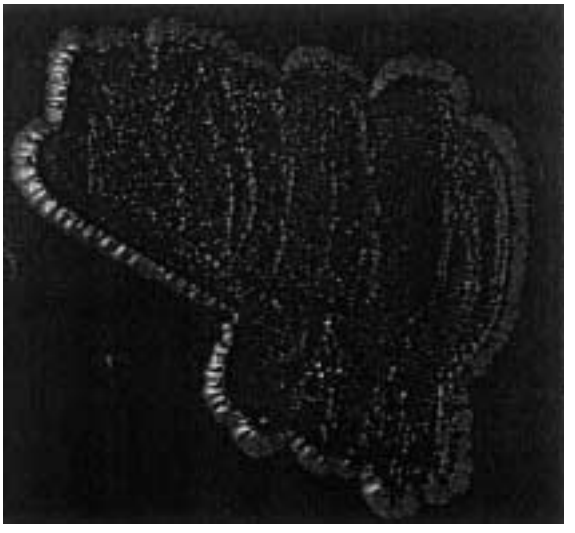


**Figure 1. The biological phenomenon: a patch of an engineered strain of Bakers' Yeast grown for one week on solid medium; the patch has a red centre and a white edge.**

Students work in groups of three over the course of three weekly laboratory sessions as they explore this problem. The week 1 session begins with a brief introduction to the biology of yeast and the tools they have available. Students first make hypotheses to explain why the centre is red and the edge white. They then design and carry out a first round of tests to address these hypotheses. During the week 2 session, they discuss the results from the round 1 tests, relate them to their hypotheses from week 1, develop new hypotheses, and design and set up a second round of tests. During the week 3 session, they discuss the results from their round 1 and round 2 tests in the light of all of their hypotheses and try to reach a consensus as a class. Throughout this process, the laboratory teaching assistants (TAs) are instructed to let the students do the thinking; students should generate hypotheses and tests and evaluate data themselves.

The phenomenon underlying the red and white colour pattern in the RWYL is complex. Rather than having one single correct answer, there are several correct findings that students can arrive at with the tools they have. The molecular details of the process leading to the red centre and white edge are described in detail by White (1999). Briefly, the red or white colour is determined genetically: cells containing an unstable genetic element (a plasmid) are red; and cells lacking this plasmid are white. Normally, a red cell divides to give two red daughters. However, since the plasmid is unstable, in roughly 1% of the cell divisions, the plasmid is not transferred to one daughter cell and thus red cell gives rise to a red daughter and a white daughter. Once a cell has lost the plasmid and become white, the plasmid can never be re-gained and thus all of that cell's descendants will be white. White cells also grow faster than red cells. Thus a red sample will always grow up to produce a mixture of red and white, while a white sample will give only white. The centre of the patch is red because the cells there are more crowded, have less access to nutrients, and therefore grow more slowly than those at the edge. As a result, the cells at the centre have had fewer

chances to lose their plasmids and remain red. The cells at the edge of the patch are exposed to more nutrients and therefore grow faster. As a result, they have had more chances to lose the plasmid and therefore become white.

With the simple tools available, students can arrive at the following correct findings:

- both red and white are alive and capable of reproduction;
- red always grows to give red and white;
- white always grows up to give only white;
- colour does not depend solely on nutrient availability or waste accumulation; and
- white grows faster than red.

This laboratory exercise was designed to expose students to some of the subtleties and complexities of scientific research within the restricted environment of an undergraduate teaching laboratory. Furthermore, analysis of students' thought processes is facilitated by features specific to the RWYL. First, because samples take one week to produce results, the RWYL moves at a relatively slow pace, which allows the students plenty of time to discuss their reasoning. Second, the students work in groups of three, which forces them to explain their reasoning out loud to each other. Finally, the material of the laboratory is very unfamiliar to the students, so that very little 'goes without saying'. As a result, transcripts of the students' speech during the RWYL provide information that is sufficient to examine the primary features of their reasoning.

Starting from this rich data set, we have developed a scheme for diagramming the course of their work that simplifies the transcript while preserving the important features of their reasoning process. Using this scheme, we can identify and characterize the students' moves as they explore the phenomenon. This scheme also allows us to evaluate the overall quality of their process in terms of its products, their conclusions. Although our analysis simplifies the students' process considerably, we are able to describe the major features of the students' behaviour at a revealing level of detail.

This study presents the application of methods for analysis, characterization and evaluation of students' hypothesis-testing behaviour to a group of students in the RWYL. We discuss what this analysis shows about the RWYL and then apply our methods to another hypothesis-testing task from the literature. Our goal is to produce a set of methods that will allow analyses like those performed previously in addition to higher-order comparisons between different hypothesis-testing tasks.

## Methods

### *Subjects*

The subjects were undergraduate students enrolled in General Biology I (Biol 111) at the University of Massachusetts Boston. Biol 111 is the first semester introductory biology course for Biology majors; it is also a requirement for students in Psychology, Nursing, and Human Performance and Fitness. The average age of Bio 111 students is 22.3, which reflects a large population of returning students; they are 76% female and 42% non-white. The course consists of three 50-minute lectures and one three-hour laboratory session per week. Lectures (one section of 250

students) were given by the principal investigator; laboratory sections (12 sections of roughly 20 students each) are taught by graduate TAs supervised by the principal investigator.

Throughout the semester, students work in groups of three during the laboratory sessions. This study examines one such group. The group consisted of two females and one male aged 19, 20, and 22 years; two were Biology majors, one was a Psychology major; their final course grades were C–, B–, and A–, respectively.

### *Procedure*

Students were video-taped and audio-taped as they completed the RWYL. Transcripts were prepared from the sections of the tapes where the students and the TA were discussing the RWYL or where the TA was talking to the laboratory section as a whole. Portions of class-wide discussions that did not involve the students in the study were not transcribed.

We have devised a method for diagramming the students' reasoning process; the general format of these Reasoning Maps is shown in figure 2. The five elements of the diagram are based on a synthesis of previous studies of hypothesis testing. To help characterize the process our subjects followed, we based our analysis on one way that these elements can be connected during a scientific investigation. Although this view of the process of science is supported by some philosophers of science (for example, Hempel 1966), inquiry-based practitioners (for example, the 'Scaffolded Inquiry Sequence' described by Hug and Krajcik [2002], the 'InquiryCycle' of White and Fredericksen [1998], and Lawson [2002]), and Introductory Biology textbooks (for example, Campbell and Reece 2002: 16–19; Purves et al. 1995: 7–9, Raven and Johnson 2002 7–8), much evidence shows that this is not always how practicing scientists proceed, at least in the context of major scientific discoveries (for example, Collins and Pinch 1993, Kuhn 1962, Latour and Woolgar (1986). For this reason, we call this sequence of moves the 'canonical' sequence — not because it is the only way or the best way to connect these elements in actual practice, but because it serves as a useful baseline to compare and contrast with students' actual practice. The steps in this 'canonical' sequence are as follows (numbers correspond to numbered arrows in figure 2; the passage of time proceeds from left to right).

1. The process begins with a *hypothesis*, defined as a mechanistic explanation or other testable statement about the phenomenon. Hypotheses were identified primarily by phrasing and context and often included statements like 'Let's hypothesize that …' or 'maybe …'. Occasionally, it was difficult to determine whether a statement was a conclusion or a hypothesis. In these cases, statements that students clearly intended to test were classified as hypotheses; the others were classified as conclusions.
2. The hypothesis prompts a *test* to evaluate it. Tests were identified by phrases like, 'Let's try …' or 'What if we …' as well as by descriptions of experiments as they carried them out. Any tests that were part of explicitly controlled experiments were linked by a dotted line.
3. The combination of a hypothesis and a test yields a *prediction* for the result of the test if the hypothesis were correct. Predictions were identified by phrases like 'then we'd predict …', 'if [a hypothesis] then we should see [a result] if we do [a test]', or other explicit statement of

expected experimental outcome. Predictions that are inconsistent with the corresponding hypothesis and test are marked with an asterisk.

4. One week later, each test generates *results*. Results were identified as descriptions of what students said that they saw on the plates, even if these observations may have been inaccurate. Inaccurate results, either due to poor technique or poor observation, are indicated with an asterisk.
5. These results reflect back on the original hypothesis to produce a *conclusion*. Conclusions were identified primarily by phrasing and context and often included statements like 'we conclude that …' or 'so it must be that …' that referred to particular pieces of data.
6. The hypothesis may then continue on for more trials.

Transcripts were converted to Reasoning Maps through several rounds of reading, revision, and discussion between two investigators. The final maps were reached by consensus with occasional reference to videotaped images. The maps were checked for accuracy by ensuring that every feature on the map corresponded to at least one part of the transcript. The maps were checked for completeness by making sure that every line of the transcript either corresponded to a feature of the map, was a side conversation, or did not contribute significantly to the map. Hypotheses, tests, and so on that were mentioned only once or did not have a connection to any other map elements were not included in the map. We also chose not to include any information on within-group dynamics; although different members of the group presented and championed particular ideas, we treated all ideas as coming from the group as a whole.

Reasoning Maps have been assigned a coordinate system similar to that used on road maps. The vertical axis is divided into rows A–D; the horizontal axis is divided into columns 1–14. Map coordinates are given as 'row column'; thus, 'C10' refers to row C, column 10.

In addition to the Reasoning Maps, we have developed three criteria for characterizing students' conclusions. Each conclusion reached by the group was scored by three independent criteria:

- *Validity*. Conclusions were scored as valid if they were consistent with the actual process underlying the red and white colour phenomenon as known to the investigators.

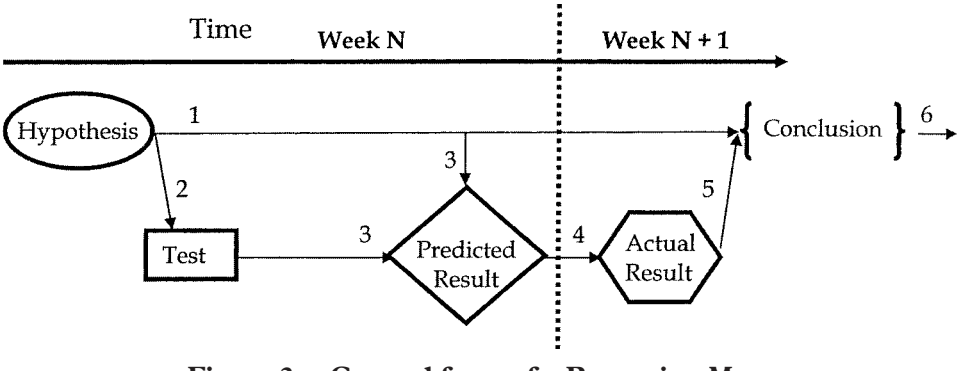



**Figure 2. General form of a Reasoning Map.**

- *Consistency*. We scored conclusions as 'consistent' if they were consistent with all the data available to that group at that time the conclusion was made, even if the data were based on incorrect techniques or observations. This criterion emphasizes the subjects' point of view. We do this both to respect the student-centred nature of inquiry as well as to take into account that, when drawing conclusions, students only have access to the data on hand at a particular time.
- *Type*. Conclusions were either *positive* (supporting the hypothesis), *negative* (refuting a hypothesis), or *NBH* (not based on a particular hypothesis).

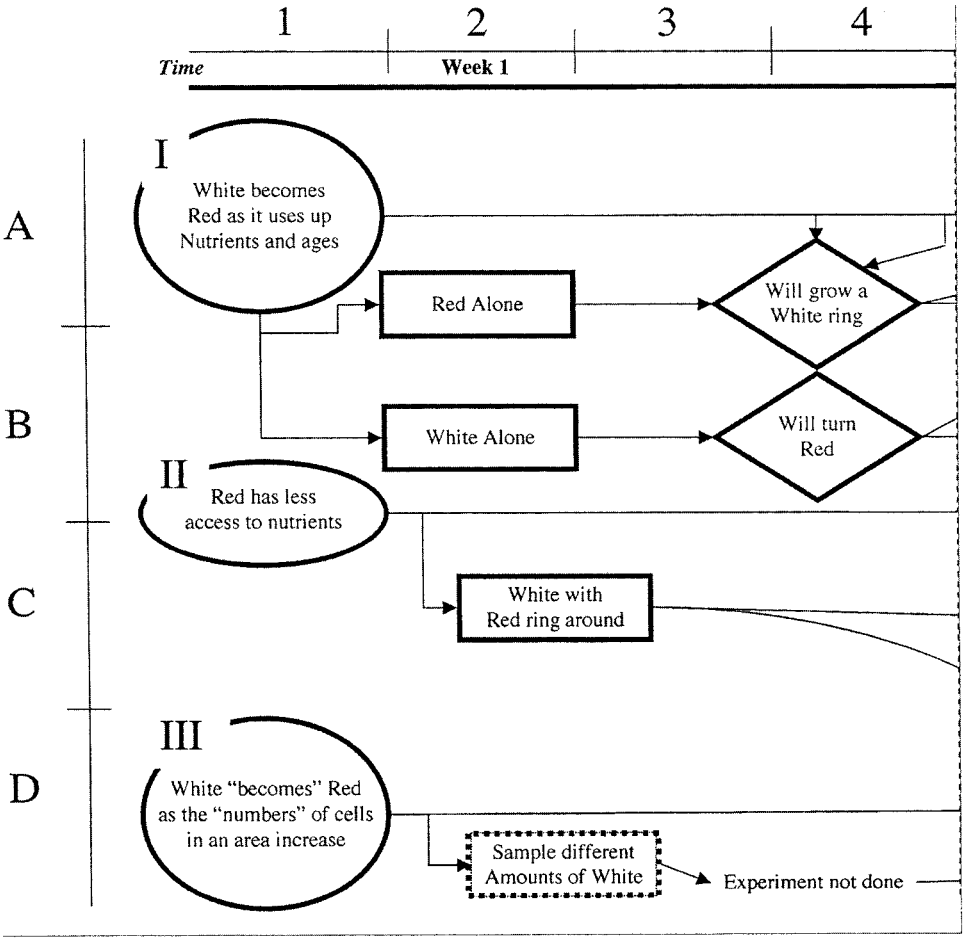
## Results

Reasoning Maps were produced from transcripts as described in Methods. Note that, while all the events shown on a given day occurred during that day and events in the same thread occur in the order shown, events in different threads are not necessarily shown in the relative order in which they occurred.

Figure 3 shows the Reasoning Map of the group's activities; the process they followed can be organized into four threads. The first thread begins with Hypothesis I, 'White becomes red as it uses up nutrients and ages' (A1). This hypothesis prompts two tests (A2 and B2) and corresponding predictions (A3 and B3). During the week 2 session, they apply a result from another thread to conclude that Hypothesis I has been refuted (Conclusion #1; A7). In spite of this conclusion, they later go on to apply the results of the White Only test (B5) to conclude that Hypothesis I is, in fact, supported (Conclusion #4; A9). In the third session, the group looked at their plates from week 1, now two weeks old, to conclude finally that Hypothesis I is incorrect (Conclusion #5; A13). Furthermore, data from this thread and another thread are combined to draw Conclusion #6 (B13), 'White doesn't become red but red becomes white', which is based on their observations rather than any particular hypothesis.

The second thread begins with Hypothesis II, 'Red has less access to nutrients' (B1). This hypothesis generates a test (C2) but no corresponding prediction. During the second session, the result of this test is used to draw three separate conclusions. First, Conclusion #1, that Hypothesis I was refuted, as was described previously. Second, in Conclusion #2 (B7), the group used this evidence to decide that Hypothesis II was supported. Hypothesis II is then dropped at this point for no clear reason. Third, they drew Conclusion #3 (C7), which was not based on any hypothesis. They then followed up on Conclusion 3 by converting the conclusion into a new hypothesis for testing (Hypothesis V, 'Red is alive'; C8). Hypothesis V prompts a test to confirm it (D9) and a prediction (C10). Interestingly, the students not only predicted the result if their hypothesis were correct, but they also predicted the result if it were incorrect. In the third session, they used their results to conclude that Hypothesis V was correct (Conclusion #7).

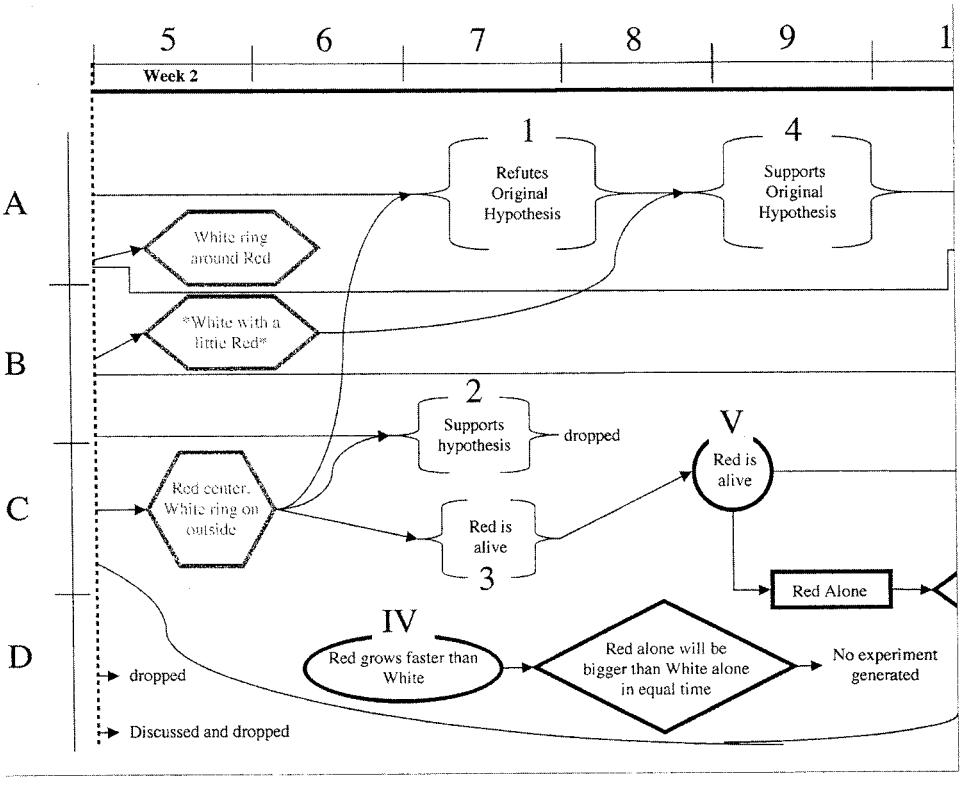
The remaining threads never generated any experiments that were actually performed. Although the group did discuss a test for Hypothesis III, 'White becomes red as the number of cells in an area increase' (D1), they never generated a corresponding prediction and never conducted the test. Hypothesis IV, 'Red grows faster than white' (D6), was discussed, and a result predicted, although no test was described; this thread was also dropped before any tests were conducted.




*Characterizing individual moves made by the students*

The threads described previously can be broken down into individual moves: elements of the reasoning pattern shown in figure 2 and the connections between them. These individual moves can be categorized in two ways. First, if the move corresponds to the pattern of moves shown in figure 2, it is 'canonical'; if not, it is 'non-canonical'. Second, if the move is logically defensible or reasonable, it is classified as 'logical' even if it comes from or leads to later 'illogical' moves. Thus, any move can be classified into one of four categories: canonical and logical, canonical but illogical, non-canonical but logical, and non-canonical and illogical. The following sections explore these four categories in detail.

*Canonical and logical moves.* These moves follow the pattern shown in figure 2 and are logically defensible. They represent the most appropriate moves the subjects made. The group designed tests based on hypotheses (A2–B2, C2, and D9). They next made predictions based on combinations of hypotheses and tests (A4–B4, and C10); all of these predictions are correct. Interestingly, these predictions almost




never crossed the boundary between experimental threads; for example, they did not make predictions based on Hypothesis II for tests designed for Hypothesis I. For most of these tests, students collected data in the following week. Finally, the group combined data and hypotheses to draw conclusions all of which were consistent with the data available at the time (#2, #3, #4, and #5); two of these are valid (#3 and #5) and two are invalid (#2, and #4) based on the actual mechanism underlying the red and white colour phenomenon.

In addition, both groups used results to evaluate one hypothesis even though the test had originally been designed for a different hypothesis. For example, the result of one test designed for Hypothesis II was used to refute Hypothesis I (Conclusion #1).

*Canonical but illogical moves.* Here, the students followed the canonical process but made errors of logic. These were of two major types: improperly observing a result, and drawing conclusions that were inconsistent with the data available. Although a sample of pure white cells is relatively easy to prepare and should yield a pure white patch after one week of growth, the students observed that their pure white sample produced at least some red colour in the centre (B2–B5 and again at B12). Whether this was the result of poor experimental technique or incorrect observation, the students used this result to draw two conclusions (#4 and #5). In addition, three of the groups' conclusions (#4, #5, and #6) were inconsistent with the data they had at the time. Conclusions #1 and #5 were based on Hypothesis I, 'White becomes red as it uses up nutrients and ages' (A1). Based on this hypothesis, the students

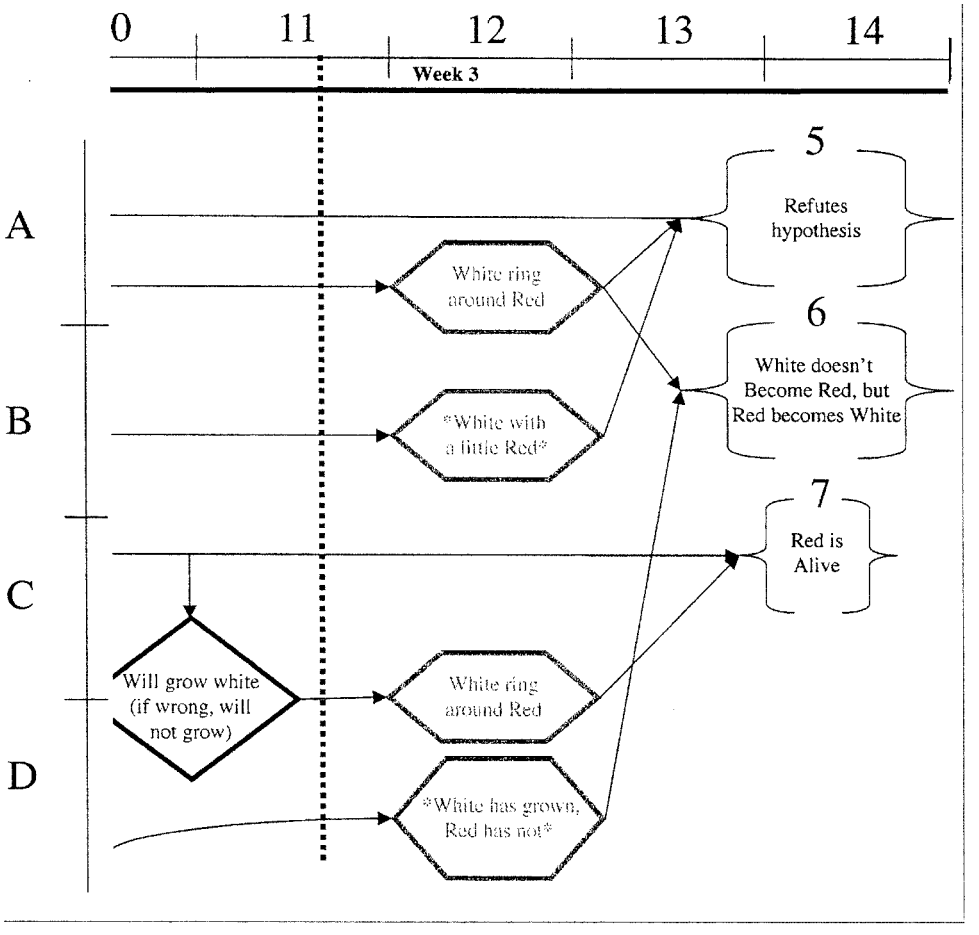



**Figure 3. RWYL Reasoning Map.**

correctly predicted that samples of red or white would grow to produce a patch with a red centre and a white edge; this result would also be expected for a mixture of red and white. All of this group's tests yielded patches with red centres and white edges. Thus conclusions #1 and #5, which rule out this hypothesis, are inconsistent with the students' results. Similarly, conclusion #6 'White does not give red but red gives white' is inconsistent with the result of the White Only experiment, which shows white growing to produce red. In spite of the logical errors, all three of these conclusions are valid

*Non-canonical but logical moves.* These moves do not follow the canonical progression but are logically defensible. These fell into two major groups, omitting intermediate steps or links and performing moves that represented novel connections not found in the canonical progression. In the canonical progression, conclusions result from a combination of a hypothesis and a result. On several occasions, the group drew valid conclusions from test results alone (Conclusions #3 and #6). These conclusions were inferences drawn from observations rather than simply descriptions

of test results. Finally, two separate lines of inquiry (D5 and D9) were dropped as a result of discussion; the students decided that these hypotheses were un-testable with the tools available or otherwise problematic.

*Non-canonical and illogical moves.* These moves were the most problematic; they involved omitting or ignoring important elements of their analysis without a logical reason or failing to make relevant connections between elements. Hypothesis III was dropped without discussion (C8). The result of the 'Red Alone' test was described but never correlated with any hypothesis (A5). In most cases, predictions were limited to tests in the same thread. In addition, the group did not predict a result for Hypothesis II and the test it prompted (C4). None of this groups' tests were designed as controlled experiments; the students never explicitly linked any tests to compliment each other in this way. Finally, the group never systematically examined all the data they had on hand to evaluate their hypotheses; data were only considered piecemeal. This led them to confirm a hypothesis that had previously been refuted (Conclusions #4 and #1) and confirm a hypothesis that is inconsistent with the available data (#6). Similarly, the group used the same result to confirm one and refute the other of two very similar hypotheses (Conclusions #1 and #2).

### *Conclusions the group reached*

During the last two laboratory sessions, the students drew seven conclusions. Each of these conclusions can be rated by three independent criteria: consistency, validity, and type. This is summarized in table 1. The students were expected to draw consistent conclusions; four out of seven were consistent with the available data. Confirmation bias would appear here as inconsistent conclusions of a positive type—confirming a hypothesis when one should not. Interestingly, for this group, the inconsistent conclusions were either negative or not based on a hypothesis; this does not reveal any confirmation bias. From the instructors' point of view, we would hope that students would draw valid conclusions; for this group, five out of seven were valid.

## Discussion

### *Hypothesis testing in the RWY Lab*

As shown in the Reasoning Map, students were able to generate hypotheses, design experiments, predict expected results, collect data, and draw conclusions. These

**Table 1. Conclusions reached by RWYL students.**

| *Conclusion* | | *Inconsistent with available data* | *Consistent with available data* |
|---|---|---|---|
| Valid | Type + | None | #7 |
| | Type – | #1, #5 | None |
| | Type NBH | #6 | #3 |
| Invalid | Type + | None | #2, #4 |
| | Type – | None | None |
| | Type NBH | None | None |

students also made a variety of procedural and logical mistakes, many of which have been observed by other workers. The students abandoned verified hypotheses (C7), and failed to seek disconfirming evidence; these were observed in Klahr's (2000) study of adults determining the function of a command in a computer simulation. In the Lawson (2002) study, his subjects, like ours, often failed to make predictions for experimental results (C4). Lawson also observed some students failing to consider alternative hypotheses when designing experiments and making predictions; our students also failed to do so (e.g. see C3). In Kuhn et al.'s (2000) studies of subjects exploring multivariate causality, they found that students frequently ignored or modified disconfirming evidence. Similarly, our students probably modified the result of their all white samples to 'observe' a reddish centre (B5 and B12). Additionally, our students never designed their tests in the format of a controlled experiment, nor did they follow a VOTAT or HOTAT strategy. This is most likely because the RWYL lacks the clearly-defined variables needed to implement these strategies. We have also identified two novel classes of erroneous moves. First, our subjects drop results (A6) without discussion. Second, they often fail to consider all relevant data when drawing conclusions.

### *Applying our methods to other hypothesis-testing activities*

Previous work has identified a set of important criteria for characterizing hypothesis-testing behaviour. Most, if not all, of these are accessible from the Reasoning Maps we have described. For example:

- *Hypotheses*. All of the previous studies examined hypotheses; these are explicitly part of the Reasoning Maps and therefore available for examination in detail.
- *Tests*. In previous studies tests took a variety of forms, each of which could be expressed in the Reasoning Map format. In addition, test strategies described previously are revealed by the Reasoning Maps: controlled experiments are indicated by dotted lines, VOTAT and HOTAT strategies are apparent by comparison of the tests conducted.
- *Predictions*. Whether required or not, these can be scored for presence/absence and consistency/inconsistency with the hypothesis and test to which they refer.
- *Results*. In Reasoning Maps, these are shown in detail and can be scored as accurate or inaccurate.
- *Conclusions*. In addition to showing these specifically, Reasoning Maps identify the particular pieces of data cited by the subjects when justifying each of their conclusions. In addition, the moves following a particular conclusion (retaining, abandoning, or modifying the hypothesis, etc.) can be tracked in detail.

Reasoning Maps provide a more revealing method for analysing the connections between elements of subjects' hypothesis-testing behaviour than those used previously. For example, Kelly et al. (1998) used Toulmin's (1958) argument framework in their analysis of students' behaviour when working with electric circuits. We have found that Toulmin's framework, which was originally designed to represent *argument*, does not capture all of the information necessary to understand our subjects' *reasoning*. As an example, figure 4 presents part of the group's process (Hypothesis

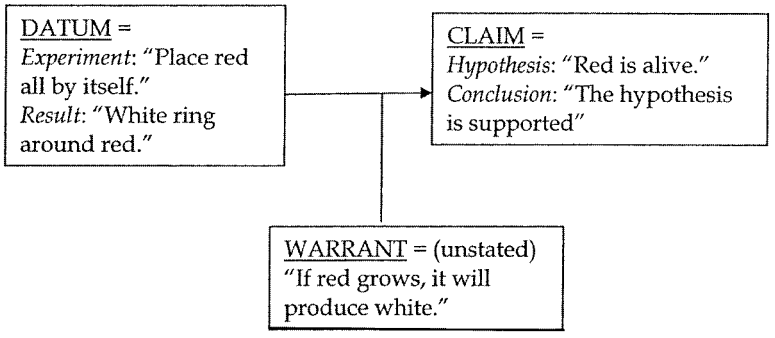



**Figure 4. Part of a Reasoning Map expressed in Toulmin's format.**

V; B8–C14) using Toulmin's argument format. Using Toulmin's format, the claim is a combination of the experiment and result conclusion. This is based on the datum; in our case, the combination of the experiment and result when combined with the warrant. Although this analysis does reveal the warrant and that it was left unstated, many of the temporal and causal links between elements are lost.

Other researchers have used diagramming schemes that more closely reflect the temporal sequence of subjects' actions. For example, Shute et al. (1989) used Student Procedure Graphs that diagram 'student actions and the resulting state of knowledge'; this is based on the Problem Behaviour Graphs used by Newell and Simon (1972). In these analyses, the 'actions' correspond to our 'tests', and the 'state of knowledge' to their conclusions. In addition to leaving out many other components of the reasoning process present in Reasoning Maps, this analysis assumes a one-to-one correspondence between results and conclusions that is clearly not present in the RWYL and other more complex hypothesis-testing tasks.

Beyond facilitating the types of analysis used previously, Reasoning Maps allow the examination of higher-order aspects of hypothesis-testing behaviour. First, Reasoning Maps allow characterization of the moves into four types (canonical and logical, etc.). The frequencies of each of these types of moves can then be compared between individuals and across hypothesis-testing tasks. These can then be correlated with the quality of conclusions reached and learning outcomes that result. Second, Reasoning Maps reveal differences in experimental approach or style that may also vary depending on individuals or tasks and may correlate with productive or un-productive outcomes.

As an example, we have applied our Reasoning Map technique to the published transcript of a different hypothesis-testing task. Figure 5 presents the first three minutes of one subject's work from one of Klahr's (2000) studies (this transcript is found on p. 127). The subject, an adult, was attempting to determine the function of the 'RPT n' command in a computer simulation. From the outset, it was clear to the subject that the 'RPT n' command repeated certain steps in the program; the subject's task was to determine which steps were repeated and how many times. The Reasoning Map of this process reveals some significant differences between this process and our subjects' process during the RWYL. In terms of individual moves, the subject of Klahr's study makes no explicit predictions of test results although he/she does design tests and draw correct conclusions from them. The flaws in his/her

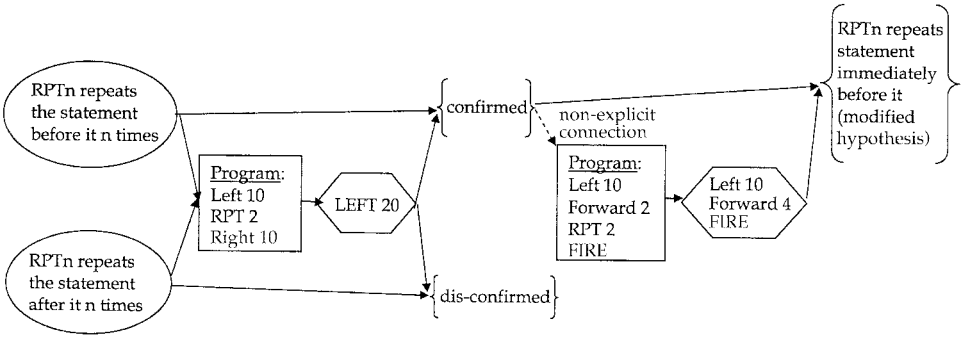



**Figure 5. Reasoning Map of three minutes of transcript from Klahr (2000): 127). Dashed lines indicate non-explicit connections.**

process, dropped intermediate elements, would be considered non-canonical and illogical moves. Perhaps the differences from what we have observed with the RWYL result from the different task environments; in the RWYL, results arrive slowly, allowing a more thorough analysis than is possible when using a computer simulation where results arrive instantaneously. Interestingly, although the process documented by Klahr contains illogical moves, the conclusions reached are consistent and valid. Further comparisons across a variety of tasks will help reveal which types of moves support or subvert the development of satisfactory conclusions. Knowledge of this sort will help understand hypothesis-testing and instructors to teach more effective hypothesis-testing strategies.

In terms of approach or style, in Klahr's study, the subject begins with two mutually exclusive hypotheses and follows a single thread of experimentation. After ruling out one hypothesis, the subject further refines the remaining hypothesis in response to an experimental result. In the RWYL, students investigate multiple threads with differing, but not necessarily mutually exclusive, hypotheses; often, the hypotheses are overlapping and highly similar (for example, Hypotheses I, II, and III). Perhaps the differences between what Klahr observed and the RWYL reflect individual stylistic differences of the subjects or some feature or features of the tasks. Here again, it will be revealing to examine the correlation between investigative style and differences between individuals, tasks, and outcomes.

Our analyses suggest that the Reasoning Maps described in this study can be applied to a wide range of hypothesis-testing situations beyond those shown here. The analysis of the resulting maps will probably reveal other trends and patterns in hypothesis-testing behaviour. It is our hope that more general application of the techniques described here will allow the development of a 'common language' for describing and evaluating hypothesis-testing behaviour across a wide variety of activities and situations. This will then facilitate a deeper understanding of this crucial component of science education.

**Acknowledgements**

Special thanks to Peter Rowinsky for help in preparing the Reasoning Maps. This work was supported by NSF CAREER Grant #9984612.

## References


BRANSFORD, J., BROWN, A. *et al.* (eds.) (1999). *How People Learn: Brain, Mind, Experience, and School* (Washington, DC: National Academy Press).

CAMPBELL, N. and REECE, J. (2002). *Biology* (San Francisco: Benjamin Cummings).

COLLINS, H. and PINCH, T. (1993). *The Golem: What Everyone Should Know About Science* (Cambridge: Cambridge University Press).

ERICSSON, K. and SIMON, H. (1984). *Protocol Analysis: Verbal Reports as Data* (Cambridge, MA: MIT Press).

HEMPEL, C. (1966). *Philosophy of Natural Science* (Englewood Cliffs, NJ: Prentice-Hall).

HUG, B. and KRAJCIK, J. (2002). *Students' Scientific Practices Using a Scaffolded Inquiry Sequence* (New Orleans, LA: National Association for Research in Science Teaching).

KELLY, G., DRUCKER, S., *et al.* (1998). Students' reasoning about electricity: combining performance assessments with argumentation analysis. *International Journal of Science Education,* 20(7), 849–871.

KLAHR, D. (2000). *Exploring Science* (Cambridge, MA: MIT Press).

KLAYMAN, J. and HA, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review,* 94(2), 211–228.

KUHN, T.S. (1962). *The Structure of Scientific Revolutions* (Chicago, IL: University of Chicago Press).

KUHN, D. and PHELPS, E. (1982). The development of problem-solving strategies. *Advances in Child Development,* 17, 1–44.

KUHN, D., BLACK, J., *et al.* (2000). The development of cognitive skills to support inquiry learning. *Cognition and Instruction,* 18(4), 495–503.

LATOUR, B. and WOOLGAR, S. (1986). *Laboratory Life: The Construcion of Scientific Facts* (Princeton, NJ: Princeton University Press).

LAWSON, A. (2002). Sound and faulty arguments generated by preservice biology teachers when testing hypotheses involving unobservable entities. *Journal of Research in Science Teaching,* 39(3), 237–252.

MOSHMAN, D. (1979). Development of formal hypothesis-testing ability. *Developmental Psychology,* 15(2), 104–112.

MYNATT, C., DOHERTY, M., *et al.* (1978). Consequences of confirmation and disconfirmation in a simulated research environment. *Quarterly Journal of Experimental Psychology,* 30, 395–406.

NEWELL, A. and SIMON, H. (1972). *Human Problem Solving* (Englewood Cliffs, NJ: Prentice-Hall).

NATIONAL RESEARCH COUNCIL (1996). *National Science Education Standards* (Washington, DC: National Academy Press).

PURVES, W., ORIANS, G., *et al.* (1995). *Life: The Science of Biology* (Sunderland, MA: Sinauer).

RAVEN, P. and JOHNSON, G. (2002). *Biology* (New York: McGraw-Hill).

SCHAUBLE, L. (1990). Belief revision in children: the role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology,* 49, 31–57.

SCHAUBLE, L., KLOPFER, L., *et al.* (1991). Students' transformation from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching,* 28(9), 859–882.

SCHAUBLE, L., GLASER, R., *et al.* (1992). The integration of knowledge and experimentation strategies in understanding a physical system. *Applied Cognitive Psychology,* 6, 321–343.

SHUTE, V., GLASER, R., et al. (1989). Inference and discovery in an exploratory laboratory. in P. Ackerman, R. Sternberg and R. Glaser (eds). *Learning and Individual Differences: Advances in Theory and Research* (New York: WH Freeman).

TOULMIN, S. (1958). *The Uses of Argument* (Cambridge: Cambridge University Press).

TSCHIRGI, J. (1980). Sensible reasoning: a hypothesis about hypotheses. *Child Development,* 51, 1–10.

WHITE, B.T. (1999). The Red & White Yeast Lab: an introduction to science as a process. *American Biology Teacher,* 61(8), 600–604.

WHITE, B.Y. and FREDERIKSEN, J.R. (1998). Inquiry, modeling, and metacognition: making science accessible to all students. *Cognition and Instruction,* 16(1), 3–118.