1-1-2011

Improving Hypothesis Testing Skills: Evaluating a General Purpose Classroom Exercise with Biology Students in Grade 9.

Michael Gregg Wilder
Portland State University

Recommended Citation

Wilder, Michael Gregg, "Improving Hypothesis Testing Skills: Evaluating a General Purpose Classroom Exercise with Biology Students in Grade 9." (2011). *Dissertations and Theses.* Paper 427.

10.15760/etd.427

Improving Hypothesis Testing Skills: Evaluating a General Purpose Classroom Exercise

with Biology Students in Grade 9.

by

Michael Gregg Wilder

A thesis submitted in partial fulfillment of the
requirements for the degree of

Master of Science in Teaching
in
General Science

Thesis Committee:
Michael Flower, Chair
Liza Finkel
Cary Sneider

Portland State University

Abstract

There is an increased emphasis on inquiry in national and Oregon state high school science standards. As hypothesis testing is a key component of these new standards, instructors need effective strategies to improve students' hypothesis testing skills. Recent research suggests that classroom exercises may prove useful. A general purpose classroom activity called the *thought experiment* is proposed. The effectiveness of 7 hours of instruction using this exercise was measured in an introductory biology course, using a quasi-experimental contrast group design. An instrument for measuring hypothesis testing skill is also proposed. Treatment (n=18) and control (n=10) sections drawn from preexisting high school classes were pre- and post-assessed using the proposed Multiple Choice Assessment of Deductive Reasoning. Both groups were also post-assessed by individually completing a written, short-answer format hypothesis testing exercise. Treatment section mean posttest scores on contextualized, multiple choice problem sets were significantly higher than those of the control section. Mean posttest scores did not significantly differ between sections on abstract deductive logic problems or the short answer format hypothesis testing exercise.

Table of Contents:

List of Tables:

List of Figures:

Introduction

Over the last two decades, there has been a research driven shift towards *inquiry* in science education. Though the rationale for this shift is multifaceted, in simple terms it can be understood as an increased emphasis on procedural knowledge. In the 1996 National Science Education Standards (NRC), there is “more emphasis” on “activities that investigate and analyze science questions,” and on viewing “science as argument and explanation.” While declarative knowledge acquisition remains important, the inquiry model consistently stresses the linkage between declarative and procedural knowledge. As Olson and Horsely (2000) put it, the goal is to foster content acquisition as well as “the thinking strategies needed to use and inquire more deeply into [science] concepts.”

These “thinking strategies” find concrete expression in the Oregon State High School Science Standards of 2009 (ODE, 2009) (Table 1.) These standards make explicit reference to analysis, investigation, argumentation, and explanation. Importantly, hypothesis testing serves as the framework within which students are expected to demonstrate these reasoning abilities. If instructors are to adequately address these standards, it is therefore important to have an explicit conception of the components of hypothesis testing.

Lawson *et al*. (2000) argue that hypothetico-deductive reasoning is of paramount importance in scientific investigation. The authors clearly delineate the six considerations that are required to engage in hypothetico-deductive reasoning. Indeed, it is worth noting that these 6 questions map readily onto the Oregon High School Science Standards as shown in Table 1.

Table 1: Oregon's 2009 High School Science Standards and Lawson *et al.*'s (2000) analysis of hypothetico-deductive reasoning.

2009 Oregon High School Science Standards:	Lawson *et al.*'s analysis of hypothetico-deductive reasoning:
H.3S.1 Based on observations and science principles formulate a question or hypothesis that can be investigated through the collection and analysis of relevant information.	1. What is the central causal question? 2. What hypotheses can be advanced to answer this question?
H.3S.2 Design and conduct a controlled experiment, field study, or other investigation to make systematic observations about the natural world, including the collection of sufficient and appropriate data.	3. How can each hypothesis be tested? 4. What are the consequences or predictions of each hypothesis and/or test?
H.3S.3 Analyze data and identify uncertainties. Draw a valid conclusion, explain how it is supported by the evidence, and communicate the findings of a scientific investigation.	5. How do the results of the tests match the predictions? 6. What conclusion can be drawn based on these results?

Given this analysis of hypothesis testing, the question remains how one might foster these skills in students. *Prima facie*, there are two possible broad categories: one might address them with laboratory exercises, and/or one might address them outside the context of the laboratory. To be sure, laboratory exercises will be necessary, since standard H.3S.2 demands that students actually "conduct" an investigation. Nevertheless, laboratory exercises have numerous practical limitations--they can be prohibitively expensive in terms of money or time. It is therefore reasonable to ask whether laboratory exercises might be usefully supplemented with classroom work that explicitly targets hypothesis testing skills.

The *thought experiment* is a general purpose classroom exercise that targets hypothesis testing skills. (More details can be found in the **Treatment** section, below.) The purpose of the present study is to test the hypothesis that 7 hours of instruction using the thought experiment exercise will improve the hypothesis testing skills of biology students in grade 9. The treatment, or independent variable, is utilization of the thought experiment exercise. The dependent variable is hypothesis testing skill. The latter was measured in two ways. Treatment and control groups were pre- and post-assessed using a novel Multiple Choice Assessment of Deductive Reasoning. Both groups also individually completed a written hypothesis testing assessment ("the caterpillar exercise.")

Review of Literature

Theory

Inhelder and Piaget (1958) described and analyzed the development of logical reasoning abilities in children and adolescents. The authors distinguished between concrete and formal mental operations. A child exhibiting concrete operations will have explicit (but limited) awareness of the abstract logical "actions" he or she deploys when solving a problem. In contrast, a child exhibiting formal operations will make explicit appeal to the logical *necessity* of a deductive conclusion. The formal operational thinker can produce and evaluate explanatory hypotheses, because of his or her recognition that observations can be *explained* "in terms of the formal operations of implication, etc., which are the conditions of hypothetico-deductive thought." In other words, hypothetico-deductive thought requires an awareness that deductive reasoning can *justify* general conclusions about observations.

In an expository paper, Lawson (2000) argues that hypothetico-deductive reasoning is a hallmark of scientific thought. He defends this view by considering multiple examples from biology, chemistry, physics and geology. He demonstrates that in each discipline, hypotheses are evaluated using a very general logical form, which he terms "if…and…then…and/but… therefore…arguments." To illustrate, consider Harvey's efforts to support his theory of unidirectional blood circulation in the human body. *If* blood circulation is unidirectional, *and if* it is maintained by one way check valves in the veins, *then*, after applying a tourniquet, one should see a vein bulging only up to the location of a check valve. In contrast, *if* blood flow is bidirectional (as Galen had argued), *then* there should not be one way check valves, *and if* one applies a tourniquet, *then* one should see the entire length of the vein bulging. Given that the competing hypotheses make different predictions, one can thus test them using the tourniquet experiment. Harvey found that veins behaved in a manner consistent with his hypothesis, but inconsistent with Galen's. *Therefore*, he concluded, Galen's view is unsupported, while the unidirectional circulation hypothesis *is* supported. Lawson convincingly demonstrates that this general deductive reasoning pattern is employed not just in biology, but also in other disciplines.

Lawson *et al*. (2000) further clarified this conception of hypothesis testing by analyzing the production of "if…and…then…therefore…" hypothetico-deductive arguments. The authors analyze this ability in terms of the ability to pose and answer 6 questions (Table 1), and make the further claim that hypothetico-deductive reasoning is equivalent to the scientific method. Though this latter characterization is justified with numerous citations, others have objected that there is not in fact *one* monolithic scientific method. (Hatton and Plouffe, 1996; Lederman *et al*., 2001) The concerns expressed by Lederman and others focus principally on the descriptive claim that *theories* are sometimes accepted by scientists even in the absence of confirming evidence. This objection can be granted by simply stating that Lawson *et al*. have characterized the process of *hypothesis testing*, rather than 'the' scientific method. Irrespective of one's views on the existence of a single scientific method, Lawson *et al*.'s analysis does accurately describe the process of hypothesis testing. As noted previously, making such an argument requires one to pose and answer 6 questions, and these map neatly onto the 2009 Oregon High School Science Standards (Table 1.)

Deductive Reasoning and Achievement

Interestingly, the theoretical position outlined above is buttressed by several studies that have found that reasoning ability is a very strong predictor of concept acquisition and science achievement. In a study of 314 high school biology and chemistry students, Lawson *et al*. (1991) found that reasoning skills, as measured by the modified Lawson Classroom Test of Scientific Reasoning (LCTSR), were an excellent predictor of student success with four concept acquisition tasks. The latter were “puzzles” that illustrated exemplars of fictitious creatures: Gligs, Skints, Mellinarks, and Quarks. These fanciful puzzles were chosen in order to eliminate the potentially confounding issue of student prior knowledge, as would be found in puzzles based on real biological or chemical concepts. Given examples and non-examples of each “creature”, students demonstrated concept acquisition by selecting other valid examples of each “creature” from a set. Only 3.3% of students who scored 0-3 (out of 12) on the reasoning instrument correctly acquired all four concepts. In contrast, 43.5% of students who scored 8-12 on the same instrument correctly acquired all four concepts. Though the LCTSR requires several discrete reasoning skills, including mathematical ability, deductive reasoning skill is *necessary* to complete each problem on this instrument.

Johnson and Lawson (1998) sought to determine whether prior content knowledge or reasoning ability might better predict achievement on course quizzes and examinations. They pretested 366 community college students in a nonmajors’ introductory biology course for their reasoning ability and content knowledge, using a multiple choice content test, and a simple, two question reasoning test. The first question demanded proportional reasoning, while the second demanded variable control. Intriguingly, prior content knowledge was not predictive of performance on quizzes and exams, while these reasoning abilities were. Using a traditional expository teaching style, a stepwise multiple regression analysis showed that 18.8% of the variance on final examination scores was explained by prior reasoning ability. In contrast, prior content knowledge did not explain a significant amount of variance.

More dramatically, Bitner (1991) found that reasoning ability was a very strong predictor of high school students' grades in science and mathematics. During the Fall semester, Bitner preassessed the reasoning abilities of 101 students in 9-12th grade at a rural high school using the Group Assessment of Logical Thinking (GALT) and the Watson-Glaser Critical Thinking Appraisal (WGCTA.) The students' final grades in their Math and Science courses were collected in the same year. The specific skill of deductive reasoning, as measured by the WGCTA, explained 65% of the variance in student scores on the GALT. GALT scores explained 62% of the variance in science grades, and 29% of the variance in math grades. This finding underscores the extent to which deductive logic predicts student achievement in math and science.

In sum, it is clear that deductive reasoning is not merely a key component of hypothesis testing. In addition, deductive reasoning is a strong predictor of success in concept acquisition tasks and math and science achievement. It is important to recognize that the correlational evidence just described is neutral on the question of *causation*. Improved deductive reasoning may cause improved concept acquisition and/or math and science achievement, or it may not. Even in the absence of clarity on the causal question, one can be confident that improved student deductive reasoning ability in biology should enhance a student’s ability to logically confront new biology problems. Still, the question remains how instructors can effectively develop student deductive reasoning skills.

Teaching Strategies

Recent research suggests that classroom exercises may be a useful way to further reinforce and develop these skills. Hurst and Milkent (1996) found that under certain specific conditions, guided practice with computer simulations improved sophomore biology students’ ability to make accurate predictions. A randomly selected sample of 30 students was preassessed using the GALT, and was asked to solve 10 prediction problems in genetics or ecology. For the latter, answers were scored correct only if the correct prediction was made, and a deductively valid argument was given in support of the prediction. The sample was then evenly divided into treatment and control groups. Both groups were exposed to 8 hours of biology-based computer simulations that required students to make predictions. The treatment group was given instructor feedback, supplemental worksheets, and group discussion was encouraged. Both groups were then post-assessed using the same set of ten genetics and ecology prediction problems. The control group mean score remained constant, while the treatment group mean score increased nearly 50%. (The researchers present their results in histograms that prevent precise computation of the mean score increase.) This result is noteworthy because it demonstrates that the guided practice only improved mean student prediction ability when supplemented by instructor feedback, worksheets, and student discussion.

In contrast to this targeted emphasis on prediction, Lawson (2000) proposed a general purpose graphic organizer that could be used to augment laboratory exercises. The exercise allows students to explicitly construct hypothetico-deductive arguments based upon their laboratory work. This exercise is anecdotally effective (Lawson, 2000), and has been incorporated into an inquiry-based Biology textbook (Lawson, 2008). However, as Lawson concedes (personal communication, 2010), the effectiveness of this exercise has never been experimentally evaluated.

In a 2000 study of a college introductory biology course, Lawson *et al*. (2000) documented substantial learning gains for hypothesis testing skills. The purpose was to test the hypothesis that two qualitatively different levels of hypothesis testing ability—concerning a) observable and b) unobservable entities—exist. A sample of 667 undergraduate students was preassessed using a modified LCTSR, as well as a brief content knowledge test. At the end of the course, all students were then asked to solve a hypothesis testing problem involving unobservable entities. The researchers' hypothesis was only moderately supported. However, the authors note that in one semester, the average score on the LCTSR increased nearly 50%. (Once again, the results are presented in histograms that prevent precise computation of the change in mean scores.) Since the purpose of the study was not to evaluate effectiveness, this substantial learning gain is unexplained. The authors hypothesize that the observed increase in hypothesis testing ability can be attributed to the fact that "the course professors and graduate teaching assistants made a very conscious and concerted effort to make alternative hypothesis testing the central theme of nearly every lecture and virtually all labs." Though reasonable, this hypothesis has also not yet been experimentally evaluated.

These studies suggest a clear path forward. Giving students guided practice with producing explicit hypothetico-deductive arguments has been anecdotally reported to be effective (Lawson, 2000). A consistent instructor emphasis on evaluating alternative hypotheses is the proposed, but untested, explanation for substantial documented gains in student hypothesis testing ability (Lawson *et al*., 2000). Instructor feedback, worksheets, and classroom discussion have been demonstrated to be the decisive factor in making practice with prediction problems effective (Hurst and Milkent, 1996). A synthesis of this information suggests that an appropriately designed classroom activity may be an effective way to improve student hypothesis testing skills.

A general purpose *thought experiment* exercise that is intended to augment a lecture or laboratory-based lesson plan is proposed. Taking into account the above findings, it gives students the opportunity to a) propose and evaluate alternative hypotheses, b) make predictions, c) interpret evidence, d) draw conclusions, e) make explicit hypothetico-deductive arguments, f) receive instructor feedback, g) record progress on a worksheet, and h) engage in classroom discussion. Further details will be given in the **Treatment** section, below. In light of the unresolved research questions discussed previously, the purpose of this study will be to evaluate the effectiveness of 7 hours of instruction using the thought experiment exercise in a biology class to improve students' deductive reasoning abilities.

Assessing Scientific Reasoning Ability

A key question in this study is how to validly assess the deductive reasoning skills necessary to engage in hypothesis testing. One commonly used instrument is Lawson's (1978) Classroom Test of Scientific Reasoning. This instrument has been demonstrated to have face validity and reliability (Lawson, 1978). Working within an explicit Piagetian framework, Lawson designed the test to measure the extent to which students exhibit formal operational thinking skills, such as proportionality, conservation, control of variables, and probabilistic reasoning. An additional key consideration was to avoid making the test unnecessarily dependant upon reading or writing ability. In its 1978 form, the instrument solicited very brief written responses, though it was later modified to a multiple choice format (Lawson *et al*., 2000.)

Taking a different approach, Sieberg (2008) proposed the Experimental Design Ability Test (EDAT.) This instrument provides a simple prompt, in the form of a causal claim about e.g., ginseng's purported ability to increase endurance. Given the prompt, students are asked to design and describe an investigation intended to test this claim, in an open-ended short essay format. The validity and reliability of this instrument are unknown, though Sieberg (2008) did find that the test was "sensitive" to changes in instructional strategy.

For reasons to be detailed below, both of these instruments have important deficiencies. A novel Multiple Choice Assessment of Deductive Reasoning is proposed as a better alternative. Further discussion of the alleged deficiencies and the proposed alternative appears in the **Instruments** section below.

Another pertinent issue to consider is the possibility that the *content* on a deductive reasoning assessment may in principle affect the outcome. Linn *et al*. (1983) have criticized Piaget for primarily assessing formal operational reasoning in a single scientific "domain"—*viz.* physics, as opposed to a chemical or biological context. In their view, there is no *a priori* reason to assume that reasoning measured in one context can be generalized to another. This objection is reasonable. One could address this objection by removing content entirely, and assessing performance with purely abstract reasoning questions. However, if the presence of a specific context actually *facilitates* deductive reasoning, then this approach would be expected to underestimate reasoning ability in context. If one employs a specific content context in the assessment, one could address the objection by including content from at least two domains. As will be discussed further in the **Instruments** section, the proposed Multiple Choice Assessment of Deductive Reasoning employs both strategies in an effort to minimize content-specific effects.

Method

Research Question

Is the thought experiment exercise effective in improving the hypothesis testing skills of biology students in grade 9?

The investigation employed a quasi-experimental contrast group design, utilizing two pre-existing introductory biology class sections as its control and treatment groups. The treatment section was exposed to 7 hours of instruction using thought experiment exercises. The control section received instruction on the same content, but it was not delivered using thought experiment exercises. Both sections performed laboratory exercises, and because the goal was to quantify the *separate* effect of thought experiment exercises, the control and treatment sections were pre- and post-assessed using a multiple choice assessment, which exists in two forms. The treatment and control groups were divided into two subsamples. This permitted alternation of Forms A and B in pre- and post-assessment. Additionally, at the end of the study, both groups received an in-class written hypothesis testing assessment ("the caterpillar exercise") to be completed independently. The features of the investigative design are summarized in Table 2.

Table 2: Summary of the Investigation.

	N=	Pre-assessment with Multiple Choice Instrument:	Treatment with 7 hours of instruction using thought experiment exercises	Post-assessment with Multiple Choice Instrument:	Independent completion of the caterpillar exercise:
Section 1 Group 1	8	Form A	Yes	Form B	Yes
Section 1 Group 2	10	Form B	Yes	Form A	Yes
Section 2 Group 1	4	Form A	No	Form B	Yes
Section 2 Group 2	6*	Form B	No	Form A	Yes

*One participant was not pretested.

Participants

Two pre-existing sections of an introductory biology course were used as the treatment (n=18) and control (n=10) sections. All students in each section were invited to participate, but only those who returned a signed parent/student consent form were included in this study. The researcher was the instructor for both sections. Both sections used the same curriculum, with the exception of the treatment. The treatment section was selected randomly. Individual students in either section were randomly placed in Groups 1 or 2 (Table 2.) Attrition during the study period caused the group sizes to be unequal.

All participants were in grade 9, and attended an urban high school in the Pacific Northwest. The following data describe the overall student body of the high school: 47.1% of students are entitled to free or reduced price lunch, 14.9% of students have IEP’s, ESL/ELL students comprise 6.1% of the student body, and 11.7% of students are recognized as TAG. The racial and ethnic composition of the student body is also fairly diverse: Asian 15.6%, African American 8.2%, Hispanic 12.8%, Native American 1.6%, White 58.2%, multiple races 2.2%, unspecified 1.8%. For the class of 2009, the cohort graduation rate was 60.6%. In 2008, the latest year for which data are available, 49.5% of incoming freshmen met or exceeded 8th grade reading benchmarks, while 64.5% met or exceeded math benchmarks.

Treatment

The treatment group received 7 hours of instruction using thought experiment exercises. The thought experiment is intended to be a general purpose classroom exercise that can be deployed in lieu of a traditional lecture, or to augment a laboratory activity. The following example, utilized in the present study, can illustrate the idea.

It has been observed that *Drosera* have secretory glands, while many other plants do not. It has also been observed that *Drosera* trap and kill insects with these glands. Given this observation, one can ask why these plants trap and kill insects. The instructor presented this question to students, and asked them to generate hypotheses. It can be objected here that the students might fail to generate any hypotheses, which is undoubtedly true. Yet if the students are not explicitly asked to produce hypotheses, as in a typical lecture, one can be certain that the students won’t. Further, it is not at all obvious that shifting this exercise to a laboratory context would change the outcome, if one assumes that the students will fail to generate hypotheses. Finally, the instructor can in principle increase the likelihood of hypothesis generation by asking skillful questions: Why do spiders kill flies? (To eat them.) Is it possible that these plants are doing the same? Are there any plant predatory insects? (There are.) Would it benefit a plant to prevent attacks by such insects? (It would.) Depending on student responses, the instructor can volunteer two hypotheses: the behavior could be defensive (killing plant predatory insects), or it might be nutritive (the plants are carnivorous.)

Given these hypotheses, students were then asked to design experiments that could test these hypotheses. Once again, if no student designs are forthcoming, the instructor can ask further questions and/or propose some experiments. If it is true that the plants are carnivorous, shouldn't we expect plants deprived of insects to fare worse than those that captured insects? If it is true that trapping is defensive, shouldn't we find plant predatory insects in their traps? Both the defensive and carnivorous hypotheses have in fact been tested by scientists, so students can be asked to draw conclusions based on those results. Plant predatory insects are not found in the traps. Furthermore, despite the presence of the trap, a great deal of leaf surface area is fully exposed to attack from such insects. So given this information, is the defensive hypothesis supported? (It is not.) Darwin found that *Drosera* deprived of insects were smaller, and had fewer flowers and seeds than those that were provided with insects. Does this support the hypothesis of carnivory? (It does.) Thus, without undertaking any actual experimental work, this exercise specifically asks students to produce hypothetico-deductive arguments. This process was enriched by having students record the progress of a discussion on a worksheet (Sample 1.) The worksheet is a very general template that requires students to fill in the relevant results, arguments, etc., while the instructor facilitates discussion. In every case, the instructor provided the "Question," *viz.,* 'Why do these plants trap and kill insects?'

Exercises like the one described above were given to the treatment group 6 times over the course of the study period. In some cases, students were asked to interpret data from a laboratory or hands on activity they had conducted. The six treatment exercises are summarized chronologically in Table 3.

Table 3: Summary of Thought Experiment Exercises

Question:	Students interpreted data from a laboratory or hands on activity:	Duration of instructional period (minutes):	Date:
Does the frequency of a trait in a population change over time?[1]	Yes	90	4/21/11
Did all species come into existence at the same time?	No	50	4/22/11
Why do these plants trap and kill insects?	No	90	4/28/11
Did echolocation evolve just once in bats?	No	50	5/6/11
Are humans more closely related to chimpanzees or gorillas?	Yes	50	5/10/11
Will shaking a box of pennies at intervals simulate the predictable behavior of radioactive decay?	Yes	90	5/18/11

[1] The instructor provided the hypotheses to be considered, in an effort to familiarize students with this new exercise.

Control section students completed the same laboratory or hands on activities, but did not explicitly offer hypotheses, design experiments, or make predictions. In lessons where no laboratory or hands on data were interpreted, control section students either read material or received a lecture concerning the empirical question listed in Table 3. In these lessons, control section students did not offer hypotheses, design experiments, make predictions, or interpret data.

Instruments

Two types of instruments were used in this study. The Multiple Choice Assessment of Deductive Reasoning was used as pre- and post-test for both the treatment and control groups. At the end of the study period, both groups were asked to individually complete a written hypothesis testing assessment ("the caterpillar exercise.")

The Multiple Choice Assessment of Deductive Reasoning is proposed to remedy deficiencies present in other commonly used instruments. Though Lawson's (1978) test is a valid and reliable measure of formal operational thinking skills, it is important to note that these skills are not equivalent to the deductive reasoning that is, by his own definition, central to hypothesis testing. Though his test does require deductive reasoning, it also requires *additional* mathematical skills. For example, the probabilistic reasoning problems demand that the student demonstrate the ability to compute probabilities based on a dataset. Absent this specific mathematical ability, a student with a mastery of deductive logic would *fail* to answer the questions correctly. If deductive reasoning is the hallmark of hypothesis testing ability, it seems reasonable to utilize an instrument that does not conflate deduction with mathematical skills.

The Experimental Design Ability Test, or EDAT (Sieberg 2008) is problematic for two reasons. The first is that its open-ended, essay response format makes it an implicit test of the student's writing ability. Given that many students in American High Schools are English Language Learners, the researcher concurs with Lawson (1978) that it is imperative to employ an instrument that minimally tests reading and writing ability.

If the intention is to measure a student's deductive reasoning ability, as it is in this case, it simply should not *essentially* depend upon writing skills.

Furthermore, the EDAT scoring rubric makes it clear that the EDAT significantly measures declarative knowledge. The scoring guide for the EDAT gives students points for indicating awareness of the placebo effect, "that the larger the sample size or number of subjects, the better the Data", and that "the experiment needs to be repeated." While these are admittedly valuable experimental design concepts, they are declarative concepts. Once again, declarative knowledge of, e.g., the existence of the placebo effect is required *in addition* to deductive reasoning skills. It is the view of the present researcher that these declarative concepts are of secondary importance, because in the absence of deductive reasoning ability, one simply cannot usefully employ knowledge of e.g., the placebo effect.

In light of these considerations, the Multiple Choice Assessment of Deductive Reasoning is proposed. The instrument has an A form and a B form, and each is divided into two sections. The first section directly tests student knowledge of abstract deductive propositional logic. Question 1 addresses affirming the antecedent. Question 2 addresses transitivity with two conditional statements. Question 3 addresses denying the consequent, which is a critical skill for hypothesis falsification. Question 4 addresses affirming the consequent, which is critical for understanding that confirmation *does not necessarily* imply that the hypothesis is true.

Section 2 is modeled after quiz problems described by Lawson *et al*. (2000.) Form A uses a pendulum problem. This problem was chosen because it requires only an understanding of the simple concepts of weight, length, and speed, and though it requires the ability to *compare* numbers, it does not require advanced mathematical operations. Questions 5 and 6 ask the student to make predictions based on the assumed truth of a hypothesis. Questions 7 and 8 determine whether the student can identify proper variable control. Questions 9 and 10 focus on the student's ability to analyze data in light of hypotheses, and draw a conclusion. Form B uses a problem concerning differential grass growth on the North and South facing sides of a greenhouse. No special knowledge of ecology or biology is necessary. Questions 5-10 require the same skills as described for the pendulum problem.

One possible objection to this instrument is that it cannot determine a student's ability to propose hypotheses, or to design an experiment. This objection is fair, but it should be noted that it also applies to Lawson's Test. The EDAT can measure a student's ability to design an experiment, but it does not measure hypothesis generation ability, since the hypothesis is given in the prompt. On balance, then, while one must acknowledge the limitations of the proposed instrument, this objection is far from fatal. A multiple choice format can only determine whether a student can *recognize* a good experimental design. The cost of measuring the ability to *generate* designs is that it requires an open ended response format. As has been argued above, an open ended response format conflates deductive reasoning ability with writing ability. Further, the proposed instrument does address variable control, which is surely a key feature of any good experimental design.

The final instrument ("the caterpillar exercise"), given to both the treatment and control sections as a post-test, provided an opportunity to measure student ability to propose hypotheses and design experiments. This instrument is adapted from a problem set (“Mealworm Quiz”) designed by Lawson *et al*. (2000.) A brief prompt concerning the behavior of caterpillars in a box is provided. An empirical question concerning their behavior is given, followed by prompts that ask students to write hypotheses, design experiment(s), make predictions, and describe results that would suggest the falsity of their hypotheses, in a short answer format. The scoring guide for this exercise appears in the appendix.

The face validity of both instruments was affirmed by a panel of 4 university professors from two universities, all of whom have expertise in science instruction. Inter-rater reliability for the Multiple Choice Assessment of Deductive Reasoning is not a concern, as it is a multiple choice instrument. Inter-rater reliability for the written caterpillar exercise was tested by having another novice instructor apply the scoring guide to a sample of four student responses. Equal scores were given in 75% of the sample. The average difference in score was 0.5.

Procedure

In late April of 2011, both the treatment and control sections were pretested using the Multiple Choice Assessment of Deductive Reasoning. 16 minutes of class time were allotted for completion of this instrument. In the control section, 4 students took the A form, while 5 took the B form. (One student was not pretested.) In the treatment section, 8 students took the A form, while 10 took the B form. The treatment group performed thought experiment exercises during six different lessons, for a total of 7 hours of instruction time. The treatment exercises are summarized in Table 3. The final treatment occurred on May 18, 2011. On May 23, 2011, both sections were post-tested using the Multiple Choice Assessment of Deductive Reasoning. Forms were alternated relative to the pretest, such that all students who had been pretested with the A form were posttested with the B form, and vice versa. Again, 16 minutes of class time were allotted for completion of this instrument. On May 25, 2011, the control and treatment section participants were given 16 minutes of class time to complete the final written caterpillar exercise.

Student names were physically removed from the instruments and replaced with anonymous identification numbers. No instruments were examined or analyzed until the investigation and instruction were completed.

Given the small sample sizes, initial comparability of the treatment and control sections was assessed by computing the pretest mean scores for a) the abstract deductive reasoning questions (1-4), and b) the Form A *and* B contextualized problem sets (pendulum *and* grass growth; questions 5-10) pooled *within* each section. These means were tested for significant difference using Minitab statistical software, following this procedure: the data were first subjected to an Anderson-Darling Test for Normality. If there was insufficient evidence to reject the hypothesis that the data were normally distributed, two sample F tests were used to determine whether the samples exhibited unequal variances. If the hypothesis that the variances were equal could not be rejected, then control and treatment means were tested for significant difference using two sample two-tailed t tests, using pooled standard deviation, at 95% confidence. If there was sufficient evidence to reject the hypothesis that the data were normally distributed, then the control and treatment means were tested for significant difference using the nonparametric Mann-Whitney test.

The Form A and B contextualized problem sets were tested for context related effects by pooling the pretest scores on the Form A contextualized problem set from both control *and* treatment section participants, and comparing these with the pooled pretest scores on the Form B contextualized problem set from both control and treatment section participants. Means were computed, and these means were tested for significant difference using Minitab statistical software, following this procedure: the data were first subjected to an Anderson-Darling Test for Normality. If there was insufficient evidence to reject the hypothesis that the data were normally distributed, two sample F tests were used to determine whether the samples exhibited unequal variances. If the hypothesis that the variances were equal could not be rejected, then control and treatment means were tested for significant difference using two sample two-tailed t tests, using pooled standard deviation, at 95% confidence. If there was sufficient evidence to reject the hypothesis that the data were normally distributed, then the control and treatment means were tested for significant difference using the nonparametric Mann-Whitney test.

The effectiveness of the treatment was assessed by computing the posttest mean scores for a) the abstract deductive reasoning questions (1-4), and b) the Form A *and* B contextualized problem sets (pendulum *and* grass growth; questions 5-10) pooled *within* each section, and c) the written hypothesis testing assessment. These means were tested for significant difference using Minitab statistical software, following this procedure: the data were first subjected to an Anderson-Darling Test for Normality. If there was sufficient evidence to reject the hypothesis that the data were normally distributed, then the nonparametric Mann-Whitney test was used to test the hypothesis that the control mean was lower than the treatment mean.

At present it is unclear how, or even whether, an effect size can be accurately estimated from the results. The predictor variable (treatment vs. control) is not continuous, so therefore Pearson's *r* cannot be used (Nakagawa and Cuthill, 2007.) The post test scores on the Form A and B contextualized problem sets were not normally distributed, so Cohen's *d* is not appropriate either (Nakagawa and Cuthill, 2007.) Whether one can compute a meaningful measure of effect size using non-normally distributed data is an unsettled matter of debate amongst statisticians (Nakagawa and Cuthill, 2007.)

Results

Mean pretest scores on the Multiple Choice Assessment of Deductive Reasoning are presented in Table 4. For the abstract logic portion (questions 1-4), the control section mean score was 2.11, and the treatment section mean score was 2.33. The maximum possible score was 4. These results are diplayed graphically in Figure 1. When all Form A *and* Form B contextualized problem set scores were pooled *within* each section, the control section mean score was 2.44, and the treatment section mean score was 3.44. The maximum possible score was 6. These results are displayed graphically in Figure 2. As summarized in Table 4, for all comparisons, the control section mean score was not significantly different than the treatment section mean score.

Table 4: Pretest Mean Scores with Results of Hypothesis Tests for Significant Difference in Means:

	Control:	Treatment:	*P* value:
Abstract Logic Portion, all participants:[1]	2.11 n=9	2.33 n=18	0.554[3]
Problem set A *and* B, all participants:[2]	2.44 n=9	3.44 n=18	0.730[4]

[1] Maximum possible score is 4.
[2] Maximum possible score is 6.
[3] From Mann-Whitney test.
[4] From two sample two-tailed t test.

A comparison for context effects was undertaken by pooling the pretest Form A pendulum problem set scores from control *and* treatment section participants, and comparing this mean score with the mean score of pooled pretest Form B grass growth problem set scores from control *and* treatment sections. As summarized in Table, the pretest Form A pendulum problem set mean score for *all* participants was 3.00, and the pretest Form B grass growth problem set mean score for *all* participants was 3.20. The maximum possible score on either problem set was 6. As shown in Table 5, no significant difference between these mean scores was found.

Table 5: Mean Pretest Scores on Form A and Form B Content Problem Sets for Pooled Participants in Control *and* Treatment Sections, with Results of Hypothesis Test for Significant Difference in Mean Score

Control *and* Treatment section mean pretest score on Form A pendulum problem set:	Control *and* Treatment section mean pretest score on Form B grass growth problem set:	*P* value:
3.00[1] n=12	3.20[1] n=15	P=0.773[2]

[1] Maximum possible score is 6.
[2] From two sample two-tailed t test.

Mean posttest scores on the Multiple Choice Assessment of Deductive Reasoning and caterpillar exercise are presented in Table 6. For the abstract logic portion (questions 1-4), the control section mean score was 2.00, and the treatment section mean score was 2.33. The maximum possible score was 4. These results are displayed graphically in Figure 1. When all Form A and Form B contextualized problem set scores were pooled *within* each section, the control section mean score was 2.11, and the treatment section mean score was 4.167. These results are displayed graphically in Figure 2. As summarized in Table 6, only this comparison yielded a significant difference in mean scores. The control section mean posttest score on Form A and B problem sets was significantly *lower* than the treatment section mean posttest score on Form A and B problem sets. The maximum probability of Type 1 error (*p*) was less than 0.0028. The control and treatment section means are illustrated in Figure 2. The control section mean score on the written hypothesis testing exercise was 4.10, and the treatment section mean score was 5.55. These results are displayed graphically in Figure 3.

Table 6: Posttest Mean Scores with Results of Hypothesis Tests for Significant Difference in Means:

	Control:	Treatment:	*P* value:
Abstract Logic Portion: [1]	2.0 n=10	2.33 n=18	0.2316[4]
Problem set A *and* B: [2]	2.11 n=10	4.167 n=18	<0.0028[4]
Caterpillar Exercise:[3]	4.10 n=10	5.55 n=18	0.1105[4]

[1] Maximum possible score is 4.
[2] Maximum possible score is 6.
[3] Maximum possible score is 8.
[4] From Mann-Whitney test.

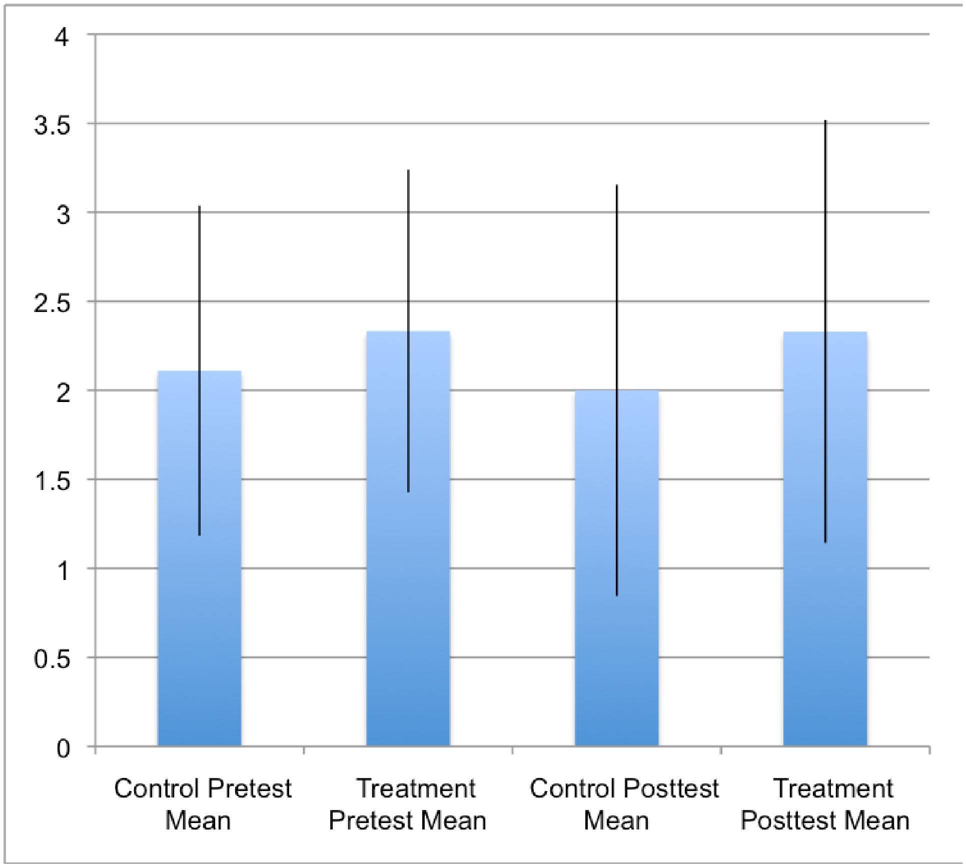

Figure 1: Control and Treatment Section Mean Scores on Pre- and Post-Test Abstract Logic Questions

Error bars are +/- one standard deviation.

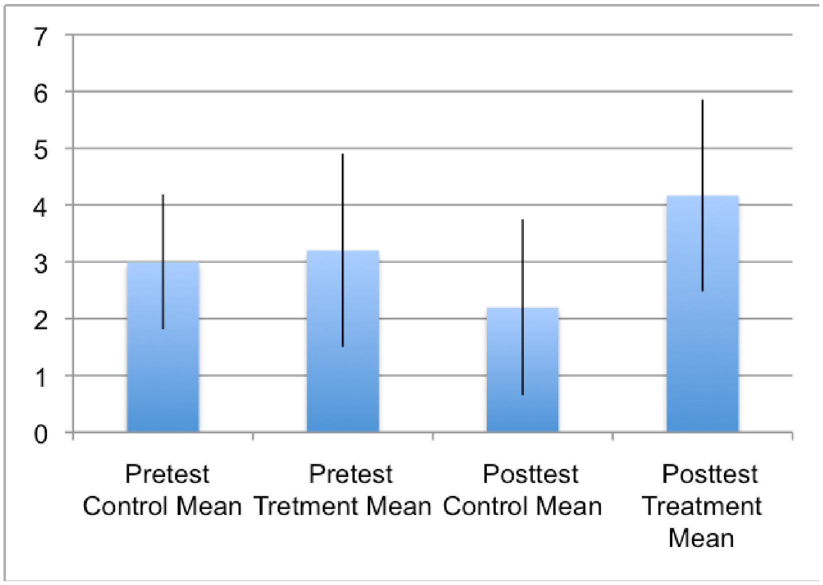

Figure 2: Control and Treatment Section Mean Scores on Pre- and Post-Test Form A *and* B Contextualized Problem Sets

Error bars are +/- one standard deviation.

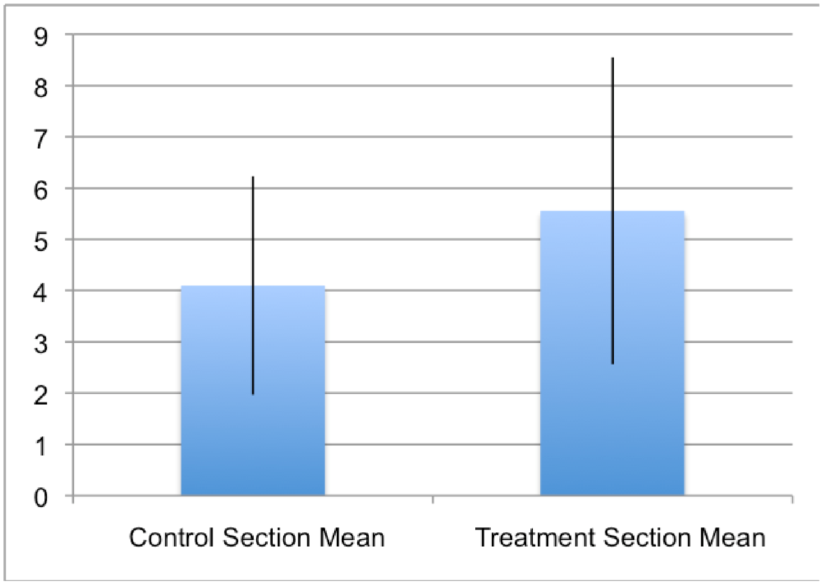

Figure 3: Control and Treatment Section Mean Scores on Written Hypothesis Testing Assessment (Caterpillar Exercise)

Error bars are +/- one standard deviation.

Discussion

Given the quasi-experimental contrast group design employed in this study, it was of great importance to determine whether the control and treatment sections exhibited initial significant differences in hypothesis testing ability. This concern was addressed in two ways, as summarized in Table 4. Control and treatment section participants' mean pretest scores on a) the abstract logic problems, and b) the Form A pendulum problem set *and* the Form B grass growth problem set did not exhibit significant differences. Based on these measures, no evidence was found to rebut the assumption that the control and treatment sections were comparable with respect to initial hypothesis testing ability.

A second key initial question was whether there was evidence of a content effect causing differential performance on the Form A pendulum problem set and the Form B grass growth problem set. Though neither problem set requires expert knowledge, in Linn et al.'s (1983) terminology, they differ in "domain"—the former utilizes physics content, while the latter utilizes biology content. This concern was addressed by pooling the control *and* treatment section pretest scores on the Form A pendulum problem set, computing the mean, and testing this for significant difference with the mean score computed from the pooled control *and* treatment section pretest scores on the Form B grass growth problem set. As shown in Table 5, these mean scores did not exhibit significant difference. There was therefore no evidence of a content effect arising from differences in problem set domains. These problem sets can be regarded as equivalent measures of hypothesis testing ability, irrespective of their difference in domain.

At the conclusion of this investigation, hypothesis testing ability was measured in three ways: a) the Form A and Form B problem sets, b) the abstract logic questions, and

c) the caterpillar exercise. The purpose of these measures was to determine whether 7 hours of instruction utilizing the thought experiment exercise would improve students' hypothesis testing ability.

The posttest Form A *and* Form B contextualized problem set scores were pooled *within* each section, and the section means were computed. The control section mean score was 2.11, while the treatment section mean score was 4.167. These mean scores were significantly different, with a maximum probability of Type 1 error (*p*) less than 0.0028. This result is consistent with Lawson's (2000) anecdotal report that practice with the explicit production of hypothetico-deducutive arguments effectively augments in class laboratory exercises.

In contrast, no significant difference in posttest control and treatment mean scores on the abstract logic portion of the Multiple Choice Assessment of Deductive Reasoning was found. This result is interesting, but not particularly difficult to explain. The thought experiment exercise consistently asked students to engage in hypothesis testing within a specific experimental context. At no time was there explicit instruction on purely abstract deductive reasoning, nor was there any practice with this skill. It is therefore unsurprising that the treatment had no effect on purely abstract deductive reasoning performance.

Mean scores on the posttest caterpillar exercise (control section=4.10, treatment section=5.55) were not significantly different. This result is somewhat unexpected, because like the Form A and B problem sets, the caterpillar exercise does ask students to engage in hypothesis testing within a specific experimental context. However, there are two pertinent differences between this assessment and the Form A and B problem sets. The first, as discussed previously, is that the caterpillar exercise utilizes a short answer format, and therefore conflates writing ability with hypothesis testing skill. The second (though related) difference is that all four of the caterpillar exercise prediction questions critically depend on the student's experimental design. A poorly designed (or ambiguously described) experiment in question 2 has the consequence that no points are awarded on the four subsequent prediction and falsification questions. In contrast, the Form A and B problem sets do not contain this interdependency between questions—a student who failed to identify appropriate variable control could still correctly predict e.g., a falsifying result. These defects in the caterpillar exercise may explain why no significant difference in control and treatment mean scores was found.

Contextual factors pertaining to the employment of the treatment may also explain this result. Utilization of the thought experiment exercise occurred during a unit on evolution, population genetics, and classification. The fundamentally historical and contingent nature of much of the unit content had important consequences. Only two of the six treatment exercises, namely those concerning *Drosera* and the radioactive decay simulation, permitted students to design traditional, repeatable, controlled-variable experiments. In contrast, the other four treatments permitted students to consider either the results of a simulation, or what could be best described as *data collection*, rather than a traditional controlled-variable experiment. For example, the empirical question concerning the origin of all species was considered in the context of the fossil record. Though students correctly proposed interpreting the fossil record as a data collection strategy, macroevolution simply is not susceptible to inquiry through traditional, repeatable, controlled-variable experimentation. Similarly, students proposed that the question concerning the closest living relatives of humans could be addressed by comparing DNA sequences and constructing a phylogenetic tree. Though this procedure *is* strictly speaking repeatable, it is again disanalogous to the traditional variable manipulation called for in the caterpillar exercise. Given that only one third (two) of the treatment exercises actually permitted students to design traditional, repeatable, controlled-variable experiments, it is plausible that the treatment had little or no effect on this skill in treatment section participants. By this hypothesis, it would be expected that control and treatment section mean scores on the caterpillar exercise would not differ significantly.

One might object that the Form A and B pendulum and grass growth problem sets concern controlled-variable experimentation that is fully analogous to the caterpillar exercise. This objection is sound. However, two points bear repeating. The first is that the Form A and B problem sets did not permit students to design experiments. Rather, they tested a student’s ability to *identify* proper variable control in a multiple-choice format. Secondly, on the caterpillar exercise, a poorly designed (or ambiguously described) experiment has the consequence that no points are awarded for subsequent questions that target prediction or hypothesis falsification. The Form A and B problem sets do not suffer from this interdependency between questions. Therefore, though it is true that all of these exercises concern traditional, repeatable, controlled-variable experimentation, there are important differences that do not support the expectation that scores on all three exercises should be comparable.

On balance, it is fair to conclude that no evidence was found in support of the thought experiment exercise’s effectiveness in improving student ability to *design* traditional controlled-variable experiments. However, this result should be interpreted with caution, given that the unit content during the investigation permitted only two opportunities for practice with this skill using the thought experiment exercise.

Conclusion

7 hours of instruction using the thought experiment exercise did significantly improve the hypothesis testing ability of introductory biology students in grade 9, when measured by the ability to identify proper variable control, predict confirming and falsifying results, and analyze data and draw conclusions, in a simple physics or biology context. The exercise did not significantly improve purely abstract deductive reasoning, nor did it significantly improve the ability to design controlled-variable experiments.

Limitations

This investigation utilized relatively small samples, in only two class sections, at one high school. In any investigation, one computes sample means in an effort to estimate means for a larger *population*. One can therefore reasonably ask what population these study participants represent. At a minimum, they represent introductory biology students in grade 9 at one high school. It is entirely unknown whether they are representative of students at other high schools, or whether they would be representative of a student population that had a considerably different racial, ethnic, or economic profile. Utilizing larger samples, drawn from multiple high schools, with different racial, ethnic, and economic profiles would offer a better way to test the effectiveness of the thought experiment exercise.

There is some support for the face validity and inter-rater reliability of the instruments used in this investigation. In the researcher's view, the caterpillar exercise, or any assessment that utilizes a short answer format, is irretrievably flawed, because it conflates writing ability with hypothesis testing ability. However, the Multiple Choice Assessment of Deductive Reasoning does not suffer from this defect.

Recommendations

This investigation provides evidence that 7 hours of instruction using the thought experiment exercise can improve the ability to identify appropriate variable control, predict confirming and falsifying results, and interpret data to draw a conclusion, in some students. Given that control section mean posttest scores were significantly lower than treatment section mean posttest scores on an assessment of the ability to identify appropriate variable control, predict confirming and falsifying results, and analyze data and draw conclusions, it appears worthwhile to further test the effectiveness of the thought experiment exercise. As described above, such an investigation should utilize larger samples, from multiple high schools, with different racial, ethnic, and/or economic profiles.

Given the evidence obtained in this study, high school biology instructors can utilize the thought experiment exercise to improve student ability to identify appropriate variable control, predict confirming or falsifying results, and interpret data to draw a conclusion. The exercise is not particularly time consuming, and it is flexible enough to be integrated into a preexisting lecture or laboratory-based lesson plan. The potential benefits can therefore be argued to outweigh the costs in terms of class time or implementation difficulty. However, it should be emphasized that this study did not produce evidence that the thought experiment exercise improves student ability to *design* experiments. It has been argued previously that this may be a contingent consequence of the study context, rather than an inherent weakness of the exercise. Nevertheless, the federal and Oregon State High school Science Standards explicitly emphasize the importance of student ability to design experiments (NRC, 1996, ODE, 2009.) At present then, instructors should not rely on the thought experiment exercise to develop this key ability.

One possible way to enhance the effectiveness of the thought experiment exercise has been proposed (Cary Sneider, personal communication, 2011.) In this study, skill development was expected to occur by means of repeated explicit production of hypothetico-deductive arguments. Sneider proposes that this process could be enriched by explicitly asking students to reflect on the reasoning strategies used within each thought experiment exercise. More specifically, they could be prompted to look for formal similarities between the reasoning patterns used as alternative hypotheses are evaluated. The time cost of this explicit reflection should be minimal, yet the researcher agrees and suggests that this practice is likely to enhance the value of the thought experiment exercise.

The results of this study point to a broader practical consideration for instructors. In the simplest terms, participants showed improvement only in those skills with which they received explicit instruction and repeated practice. The thought experiment exercise has been demonstrated to be a useful strategy for meeting some, but not all, of the expectations embodied in the national and Oregon high school science standards. In light of this result, instructors must allot sufficient time to activities that provide students with explicit instruction and repeated practice in the full range of hypothesis testing abilities, including experimental design.

References

Bitner, B. L. (1991). Formal operational reasoning modes: predictors of critical thinking abilities and grades assigned by teachers in science and mathematics for students in grades nine through twelve. *Journal of Research in Science Teaching, 28*(3), 265-274.

Hatton, John and Plouffe, Paul B. (Eds.) (1996). *Science and its Ways of Knowing*. New York, NY: Benjamin Cummings.

Hurst, R.W., & Milkent, M.M. (1996). Facilitating successful problem solving in biology through application of skill theory. *Journal of Research in Science Teaching*, *33*(4), 541–552.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures*. (S. Milgram & A. Parsons, Trans.) New York, NY: Basic Books, Inc.,

Johnson, M and Lawson, A. (1998). What are the relative effects of reasoning ability and prior knowledge on biology achievement in expository and inquiry classes? *Journal of Research in Science Teaching*, *35*(1), 89–103

Lawson, A. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching*, *15*(1), 11-24.

Lawson, A., McElrath, C., Burton, M, James, B, Doyle, R, Woddward, S, Kellerman, L, Snyder, J. (1991). Hypothetico-deductive reasoning skill and concept acquisition: testing a constructivist hypothesis**.** *Journal of Research in Science Teaching, 28*(10), 953-972.

Lawson, A. (2000). The generality of hypothetico-deductive reasoning: making scientific thinking explicit. *The American Biology Teacher*, *62*(7), 482-495.

Lawson, A., Clark, B., Cramer-Meldrum, E., Falconer, K., Sequist, J.M., Kwon, Y. (2000). Development of scientific reasoning in college biology: do two levels of general hypothesis-testing skills exist? *Journal of Research in Science Teaching*, *37*(1), 81-101.

Lawson, A. (2008). *Biology: An inquiry approach*. New York, NY: Kendall Hunt Publishing Co.

Lederman, N.G., Fouad Abd-El-Khalick, R.L., Bell, R., & Schwartz, M. (2001). Views of nature of science questionnaire: toward valid and meaningful assessment of learners' conceptions of nature of science. *Journal of Research in Science Teaching*, *39*(6), 497–521.

Linn, M.C., Clement, C., & Pulos, S. (1983). Is it formal if it's not physics? (the influence of content on formal reasoning). *Journal of Research in Science Teaching, 20*(7), 755–770.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, *82*(4) pp. 591-605

National Research Council (1996). *National science education standards*. (p. 113) Washington, DC: National Academy Press.

Olson, S., & Louks-Horsely, S. (Eds.) (2000). *Inquiry and the national science education standards: A guide for teaching and learning.* **(**p. 120). Washington, DC: National Academy Press.

Oregon Department of Education (2009). *Standards by design: High school for science 2009*. Retrieved from http://www.ode.state.or.us/teachlearn/real/standards/sbd.aspx

Appendix: Instruments and Scoring Guide

Sample Worksheet for Thought Experiment Exercises: Student writing appears in italics.

Observations: *All of these plants have trapped insects.*

Question: *Why do these plants trap insects?*

Hypotheses: *1. The trapping may serve a defensive purpose.*
2. The plants may be carnivorous.

Predictions: *If hypothesis 1 is true, we would expect the plants to capture plant predatory insects. If hypothesis 2 is true, we would expect plants that were deprived of insects to be less successful.*

Experiments: *One could conduct a field study of these plants, to determine whether plant predatory insects are trapped. One could conduct a controlled growth experiment, in which some plants were deprived of insects, while others were provided with insects.*

Results: *Plant predatory insects were not found in the traps. Plants provided with insects did grow larger, and had more flowers and seed.*

Conclusions: *Since plant predatory insects were not found in the traps, there is no support for hypothesis 1. Since plants provided with insects were more successful, there is support for hypothesis 2.*

Multiple Choice Assessment of Deductive Reasoning

Pre-PostTest Form A

Section 1: Propositional Logic

1. Suppose we know that:
If A is true, then B is true.
And
A is true.
What can you conclude?
a. B is false.
b. B could be either true or false.
c. B is true.
d. There is not enough information to answer.

2. Suppose we know that:
If A is true, then B is true.
If B is true, then C is true.
A is true.
What can you conclude:
a. C could be either true or false.
b. B and C are true.
c. Only B is true
d. There is not enough information to answer.

3. Suppose we know that:
If A is true, then B is true.
B is false.
What can you conclude?
a. A is false.
b. A is true.
c. A could be either true or false.
d. There is not enough information to answer.

4. Suppose we know that:
If A is true, then B is true.
B is true.
What can you conclude?
a. A is false
b. A could be either true or false.
c. A is true.
d. None of the above.

Section 2

A swinging string with a weight at the end is called a pendulum. Amy has found that with one foot of string, and a one pound weight, it always takes 2 seconds for the pendulum to swing. She wonders what causes pendulums to swing fast or slow. Amy has two hypotheses:
Hypothesis 1: A change in the weight at the end of the string will cause a difference in the swing speed. The lighter the weight, the faster the swing.

Hypothesis 2: A change in the length of string will cause a difference in the swing speed. The shorter the string, the faster the swing.

5. If hypothesis 1 is true, then if Amy uses a one foot string and a two pound weight, the swing should be:
a. Slower than 2 seconds.
b. Faster than 2 seconds.

c. Exactly 2 seconds.
d. There is not enough information provided to answer the question.

6. If hypothesis 2 is true, then if Amy uses a 6 inch string and a one pound weight, the swing should be:
a. Slower than 2 seconds.
b. Faster than 2 seconds.
c. Exactly 2 seconds.
d. There is not enough information provided to answer the question.

7. Amy decides to make a pendulum with a two foot string and a two pound weight. If she times the swing, this experiment will be:
a. a good test of both hypotheses.
b. A good test of hypothesis 1.
c. A good test of hypothesis 2.
d. An uninformative (bad) test of both hypotheses.
e. None of the above.

8. Amy makes a pendulum with a 6 inch string and a two pound weight. If she times the swing, this experiment will be:
a. A good test of both hypotheses.
b. A good test of hypothesis 1.
c. A good test of hypothesis 2.
d. An uninformative (bad) test of both hypotheses.
e. None of the above.

9. Using the pendulum with a three foot string and one pound weight, Amy measures the swing. Suppose that it takes 4 seconds. This result suggests that:
a. Hypothesis 1 is probably true.
b. Hypothesis 2 is probably false.
c. Both hypotheses are probably true.
d. Both hypotheses are probably false.
e. None of the above.

10. Using a pendulum with a one foot string and a 5 pound weight, Amy measures the swing. Suppose that it takes 3 seconds. This result suggests that:
a. Both hypotheses are probably true.
b. Both hypotheses are probably false.
c. Hypothesis 1 is probably true.
d. Hypothesis 2 is probably true.
e. None of the above.

Pre-Post Test Form B

Section 1 is exactly as in Form A.

Section 2:

Amy was studying grass growing in a greenhouse. She discovered that there was more grass growing on the North-facing side than on the South-facing side of the greenhouse. The temperature is held constant in the greenhouse. She wonders what causes this difference in grass growth. Amy has two hypotheses:

Hypothesis 1: The soil moisture on the South side is lower than on the North side. Because the moisture is lower on the South side, the grass doesn't grow as well.

Hypothesis 2: The South side receives more sunlight than the North side. The light is too intense on the South side, and therefore grass doesn't grow as well.

Assume that changes in light intensity do not affect soil moisture.

5. If Hypothesis 1 is true, then if Amy increases the soil moisture on the south side, so that it matches the North side, the grass should:
a. grow less.
b. grow more.
c. grow the same as before.
d. there is not enough information to answer.

6. If Hypothesis 2 is true, then if Amy partially shades the grass on the South side (decreasing the sunlight intensity), so that it matches the North side, the grass should:
a. grow more
b. grow less
c. grow the same as before.
d. there is not enough information to answer.

7. Amy decides to increase the soil moisture and partially shade the grass on the South side. If she later measures the grass growth, this experiment will be:
a. a good test of both hypotheses.
b. a good test of hypothesis 1.
c. a good test of hypothesis 2.
d. an uninformative (bad) test of both hypotheses.
e. none of the above.

8. Amy decides to decrease the soil moisture and provide additional light to the grass on the South side. If she later measures the grass growth, this experiment will be:
a. a good test of both hypotheses.
b. a good test of hypothesis 1.
c. a good test of hypothesis 2.
d. an uninformative (bad) test of both hypotheses.
e. none of the above.

9. Amy decides to increase the soil moisture on the South facing side, so that it matches the North side. She allows the light intensity to remain at normal levels. Suppose she finds that the grass grows more than before. This result suggests that:
a. Hypothesis 1 is probably false.
b. Hypothesis 2 is probably true.
c. both Hypotheses are probably true.
d. both hypotheses are probably false.
e. none of the above.

10. Amy decides to partially shade the grass on the South side, decreasing the light intensity so that it matches the North side. She allows the soil moisture to stay at normal levels. Suppose she finds that the grass grows more as a result. This result suggests that:
a. both hypotheses are probably true.
b. both hypotheses are probably false.
c. Hypothesis 1 is probably true.
d. Hypothesis 2 is probably true.
e. none of the above.

Written Caterpillar Exercise

Read the following, and then answer the questions below.
Amy recently placed some caterpillars in a rectangular box to observe their behavior. She noticed that the caterpillars tended to group at the right end of the box. She also noticed that the right side had some leaves in it and that the box was darker at that end. She wondered what caused them to group at the right end.

1. In the space below, write at least one hypothesis that could explain why the caterpillars move to the right side of the box.
Hypothesis 1: That they moved to the right end of the box because it was dark.

Hypothesis 2: That they moved to the right end because it was leaves in it.

2. How could you test your hypotheses? Describe an experiment that could help Amy know whether the hypotheses are false.

3. If hypothesis 1 is true, what results would you expect in your experiment?

That they will go right because of darkness.

4. If hypothesis 2 is true, what results would you expect in your experiment?

5. What results would show that hypothesis 1 is probably false?

6. What results would show that hypothesis 2 is probably false?

Scoring Guide for Written Caterpillar Exercise

1. Student proposes one reasonable hypothesis: 1 point.
2. Student proposes an additional reasonable hypothesis: 1 point.
3. Student describes an experiment that could reasonably test hypothesis one, and that appropriately controls variables: 1 point
4. Student describes an experiment that could reasonably test hypothesis two, and that appropriately controls variables: 1 point
5. Student accurately predicts the expected result if hypothesis one is true: 1 point.
6. Student accurately predicts the expected result if hypothesis two is true: 1 point.
7. Student provides a result that would suggest the falsity of hypothesis one: 1 point.
8. Student provides a result that would suggest the falsity of hypothesis two: 1 point.