IPESL Project Summary Report

1. Title of Project

Did we learn anything? Evaluating the effectiveness of physics, psychology and philosophy courses designed to improve critical thinking in the sciences.

Faculty: Jeffrey Buchanan, Ph.D. (Psychology), Steven Kipp, Ph.D. (Astronomy), and Melanie Frappier, Ph.D. (Philosophy)

2. Purpose

While students are asked to take courses to improve their critical thinking skills, little information is available as to these courses' effectiveness. The aim of this research is to evaluate the effectiveness of three critical thinking courses, namely PHIL 112: Logic of the Scientific Method, PHYS 115: Life in the Universe, and PSYC 103: Psychology Today.

Assessing the development of critical thinking skills is a difficult task. Following Paul and Nosich (1993), we understand critical thinking as "the intellectually disciplined process of actively […] conceptualizing, applying, analyzing, synthesizing, or evaluating information gathered from […] observation, experience, reflection […] as guide to belief and action" This means that critical thinking assessments should not focused on students' correct answers, but on how they achieve these correct answers. As Facione writes: "The challenge of CT assessment is not to let what is easily measured restrict our sense of the fullness of CT. It would be shameful if those assessment instruments […] drove our CT curricular design and caused the dispositional components of good CT to be neglected" (1990, p.17).

Yet, evidence suggests that among the different instruments proposed to evaluate critical thinking skills (e.g. The California Critical Thinking Skills Test, the Watson Glaser critical thinking test, The Cornell Critical Thinking Test, etc.), few capture both cognitive *and* dispositional critical thinking abilities. This research will therefore initially focus on developing an appropriate psychometric instrument to evaluate both cognitive and dispositional attitudes of students, before evaluating and comparing the effectiveness of each course.

* Note: we originally were unsure if we would use an existing measure or design our own. We decided that the existing measures did not capture critical/skeptical thinking, so we decided to design our own measure.

3. Results

This project involved the development of a critical thinking measure that could potentially be used as a pre-post measure in classes taught by the three project faculty.

We attempted to design a relative brief and ecologically valid (i.e., content reflects claims that are often made in the mass media) measure of critical/skeptical thinking.

Over the course of the last several months, many different versions of the test were developed that included a variety of different questions and response formats (e.g., open ended questions requiring essay answers, completing rating scales). Eventually, four separate forms of the test were developed, two being formatted like an advertisement that might be seen in a magazine or newspaper and the other two being formatted as a brief newspaper article. Four questions are included that measure the student's ability to: 1) identify claims, 2) identify evidence, 3) evaluate the quality of evidence, and 4) develop questions for claimants to further assess the validity of claims. All questions require the student to provide a short essay answer or a list of points. One question requires the student to provide an overall rating of the quality of the evidence provided. Specific scoring criteria were developed for all tests to increase the likelihood that different scorers could reliability grade the test and come up with similar scores.

After the four separate versions of the test were developed, we then piloted them with a group of students from Dr. Kipp's Astronomy 101 classes in the spring semester (IRB approval was obtained to do this). The purpose of this initial pilot study was to see if the test made sense to students and to see if the tests could be scored reliably. A total of 41 tests were completed and scored. Inter-rater reliability correlations ranged from .804 to .908, indicating that the scoring criteria were well-developed and different raters could scores the tests similarly. After the initial pilot study, the content of the tests were changed/simplified and scoring criteria were modified so as to be clearer.

Next, another group of students in Dr. Kipp's Astronomy 101 class (summer 2007) were asked to complete the revised versions of the test (IRB approval was obtained to do this). A total of 29 students completed the tests. Inter-rater reliability correlations ranged from .787 to .857, again indicating that the scoring criteria were well-developed and different raters could scores the tests similarly. Scoring criteria have now been revised and clarified once again with the hopes the inter-rater reliability can be improved. It is hoped that this test can be used as a pre-post measure of critical/skeptical thinking skills in classes. However, reliability must be improved before the test can be used in this manner.

As of June 2007, plans are to have Dr. Buchanan and his research team continue data collection in the fall of 2007 with continued input from Drs. Kipp and Frappier. This work will involve training 2-3 graduate students to use the scoring criteria, attempting to further improve inter-rater reliability, and making the four versions of the test as equivalent as possible (e.g., produce similar mean scores and standard deviations). Once adequate reliability has been achieved, the test will be used as a pre-post measure in Drs. Buchanan and Kipp's classes that explicitly focus on teaching critical thinking (Psychology 103 and Astronomy 115).

4. Issues

Many challenges arose during the completion of this project. First was determining what we were actually trying to measure and how to go about measuring it. Although we wished to measure "critical thinking", this is a very broad construct and can mean many different things. Through our discussions, it became clearer that we all wanted our students to more skeptical of claims that sound scientific, but appear questionable. We agreed that after taking our classes, we want students to be better at asking questions and be better at evaluating evidence that is used to bolster claims that they might see/hear in the mass media. We decided we wanted to see if students could use "critical thinking skills" to evaluate real life scenarios. This is what lead us to develop the test in the manner that we did (i.e., having the test stimuli formatted as newspaper articles and advertisements).

Another challenge was developing a brief, but comprehensive test of "skeptical thinking." Again, through many discussions we decided that we mainly wanted students to be able to specify claims, evaluate evidence allegedly supporting a claim, being able to evaluate the quality of the evidence and then be able to ask questions of the claimant to determine how valid a claim actually is.

Finally, developing scoring criteria for the essay-style questions was quite challenging. We want others to be able to use this test, thus making it necessary to specify criteria by which to grade the exams (instead of grading based on our gut instinct or professional opinion). This proved more challenging than we initially thought. For example, my scoring criteria made perfect sense to me, but did not always make sense to the other investigators. We resolved this by first having all of us attempt to use each other's grading criteria. Once grading was complete, we then identified tests where two or more of us significantly disagreed about the final score. We then discussed how we graded the test and identified sources of disagreement (e.g., unclear scoring criteria, scoring mistakes). This allowed us to make scoring criteria more specific, thus requiring less judgment or inference when grading tests. We believe these measures have improved the scoring criteria and they have produced good inter-rater agreement. We now need to attempt to train other individuals (i.e., persons who did not develop the test) to use the scoring criteria and determine how well the scoring criteria work. As stated above, the plan is to do this in the fall of 2007.

5. Dissemination

Until we can establish better psychometric properties of this test there are no immediate plans for dissemination. However, if we can establish the reliability of the test and show that it is valid for some purposes (e.g., scores improve after taking a critical thinking course, scores correlate with scores on similar measures of critical thinking), then dissemination at a regional or national conference (such as the Midwestern Conference on Professional Psychology or the National Institute on the Teaching of Psychology) would be feasible. Publication in a journal devoted to the teaching of psychology (e.g.,

Teaching of Psychology) or the teaching of Philosophy (e.g., *Teaching of Philosophy*) may also be feasible in the future if adequate psychometrics can be established.