INCREASING THE LEVELS AT WHICH UNDERGRADUATE STUDENTS ANSWER QUESTIONS IN A COMPUTER-AIDED PERSONALIZED SYSTEM OF INSTRUCTION COURSE

BY

Darlene Eleanor Crone-Todd

A Thesis
Submitted to the Faculty of Graduate Studies
In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

Department of Psychology
University of Manitoba
Winnipeg, Manitoba

January, 2002

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION PAGE

INCREASING THE LEVELS AT WHICH UNDERGRADUATE STUDENTS ANSWER QUESTIONS IN A COMPUTER-AIDED PERSONALIZED SYSTEM OF INSTRUCTION COURSE

BY

Darlene Eleanor Crone-Todd

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

Acknowledgements

I would like to express my sincere thanks to my advisor, Dr. Joseph Pear, for his patience, guidance, support, enthusiasm, and sense of humour. In addition to training in Behaviour Analysis, Learning, and Instructional Design, Joe has provided me with an incredibly rich environment in which to develop both personally and professionally. I consider myself fortunate indeed to have Joe as my graduate advisor.

I would also like to thank all of the members of my Doctoral Advisory Committee and members of my Examination Committee including, Drs. Joseph Pear, Stephen Holborn, Linda Wilson, Richard Bochonko, and the external examiner, Dr. George Semb (Department of Human Development and Family Life, University of Kansas). Their guidance, careful review of the manuscript, and helpful recommendations are much appreciated. I would also like to extend a special thanks to Ms. Kirsten Wirth and Ms. Sabrina Berry for their help with carrying out assessments on this project, along with their enthusiastic support. In addition, my thanks go to the members of the CAPSI lab and Pearlab who have provided time, friendship, and emotional support. These people include Kirsten and Sabrina, as well Mr. Jeremy Gawryluk, Mr. Toby Martin, Ms. Cynthia Read, Ms. Heather Simister, Ms. Tanya Walsh, Ms. Kerri Walters, Ms. Naomi Woodman. Also, thanks go to Dr. Marion Aftanas, Dr. John Whiteley, Dr. Helen Bochonko, and Dr. Jim Nickels for their time, support, and interesting dialogue.

My family deserves special thanks for their unwavering support and understanding. Thanks go to my mother (Marilyn), father (Fred), step-mother (Shirley), and my brothers (Glen and Bill) for their patience during this process. My heartfelt thanks go to my husband, Barrie, for his unwavering love, patience, understanding, help, and financial support. His support in these, and so many other ways, has provided me with a unique opportunity to realize a dream. I will be forever grateful.

This research was supported by funds from the Social Sciences and Humanities Research Council of Canada (SSHRC) grant to JJP, and a SSHRC Fellowship to DECT.

Table of Contents

List of Tables

List of Figures

Abstract

The development of critical, or higher-order, thinking skills in undergraduate students is considered to be the hallmark of post secondary education. Despite the importance of these cognitive skills, the majority of the literature does not provide reliable or valid measures to assess whether the complex behavior involved in higher-order thinking has occurred. The present study used a modified version of Bloom's (1956) taxonomy, which incorporates 6 behaviorally defined levels of thinking. Students in 2 undergraduate psychology courses (1 experimental and 1 control), using a computer-aided personalized system of instruction (CAPSI), answered various levels of questions to guided study question on the course material. The taxonomic levels were (a) Level 1 - Rote Knowledge, which involves phrasing of answers that are close to the text; (b) Level 2 - Comprehension, which involves correct answers are in the student's own words; (c) Level 3 - Application, which involves generating or identifying original examples; (d) Level 4 - Analysis, which involves a comparison or contrast or concepts, principles, or processes; (e) Level 5 - Synthesis, which involves putting together parts to form a whole, or generating a definition from examples; and (f) Evaluation, which involves cogently arguing a point of view. Higher-order thinking is considered to occur at Levels 3 through 6. While previous research on personalized systems of instruction (PSI) has demonstrated that PSI can produce higher examination scores and application-level responses, no previous research has studied whether PSI can be used to teach higher-order thinking beyond the Application level. Further, using behaviorally defined levels of critical thinking to study this question is also unique. This study examined whether strategies used in CAPSI-taught courses can result in higher-order thinking as measured by

students' answers to exam questions. The main independent variables were (a) feedback to students on the level at which they answered questions on unit tests and exams, and (b) provision of bonus points for higher-order answers above the identified minimum level at which the questions could be answered. Students in the experimental course scored an average of approximately 18% higher than students in the control course on independently scored final examinations, answered more higher-order questions correctly, and were more likely to answer at taxonomic levels above the question level. These findings suggest that the interventions were effective, and suggest that they should be applied ensure that higher-order thinking skills are developed in students.

Increasing the Levels at Which Undergraduate Students Answer Questions in a Computer-Aided Personalized System of Instruction Course

"When we study human thought, we study behavior. In the broadest possible sense, the thought of Julius Caesar was simply the sum total of his responses to the complex world in which he lived." (Skinner, 1957, pp 451-452).

The Problem

Post-secondary education is the hallmark setting in which higher-order, or critical, thinking skills are to be developed. It has also been argued (Halpern, 1998) that the teaching of these skills is an essential component of instruction in post-secondary institutions, and that these skills should be taught at all levels within education (Facione, 1997); however, this is not always the case. The challenge for educators, then, is to develop course materials that foster the development of students' higher-order thinking skills. This is complicated by the fact that the very skills we seek to foster are not operationally or behaviorally defined (Williams, 1999) in a way that facilitates objective assessment of whether such thinking has occurred. Therefore, assertions that higher-order thinking has occurred is subject to criticism as a subjective assessment; it is difficult to assert with confidence that higher-order thinking has indeed taken place.

As the quotation by Skinner above indicates, thinking is behavior. When defined behaviorally, thinking can be studied objectively in terms of whether it has occurred, and what its controlling variables are. Such an approach requires behaviorally oriented operational definitions, high inter-scorer agreement (ISA) on behavioral assessments, and demonstration that a particular intervention has the desired effect. The purposes of my

research were to (a) provide a background on critical, or higher-order, thinking, (b) to describe the problem with traditional approaches to studying it, (c) to provide a behavioral interpretation of what is meant by the term "higher-order thinking" and (d) to review the technology (personalized system of instruction) used in the current study, and how that technology is relevant to studying higher order thinking. More specifically the aim of my research was to use a computer-aided personalized system of instruction (CAPSI) to study the effects of a behavioral intervention, involving the use of targeted feedback and bonus points for higher-order thinking, which was targeted to increase the frequency of behaviorally defined higher-order thinking in a 2nd year undergraduate applied behavior analysis course.

Background on Higher Order Thinking

The literature includes many examples of higher-order thinking, and various definitions have been suggested. For example, the use of "reasoned argumentation" (i.e., the use of arguments that are supported by valid and sound premises, Newman, 1991a, b) is often considered to be the highest of the thinking levels. Other authors have suggested other dimensions are indicative of the process, such as considering the sameness (i.e., comparisons) of elements, principles, or concepts (Carnine, 1991), application of principles, processes, or concepts (Hohn, Gallagher, & Byrne, 1990; Semb & Spencer, 1976), or making effective judgements in light of the knowledge that one gains in the context of a specific discipline (Paul & Heaslip, 1995). Mayer and Goodchild (1990) have defined critical thinking (in psychology) as an attempt at an active, systematic process that is based on arguments that are understood and evaluated.

While all of these approaches to higher-order thinking discuss different aspects of what is perhaps best described as a set of behaviors that are more complex than memorization of text, none of them appear to capture the various types of conditions under which more complex behavior occurs, nor do they identify procedures to reliably or validly assess whether complex behavior has occurred. One set of definitions, *Bloom's Taxonomy of the Cognitive Domain* (Bloom, 1956; Bloom, Hastings, & Madaus, 1971), provides several different types, or levels, of behaviors that can occur which may be labeled as "critical thinking". These levels include knowledge, comprehension, application, analysis, synthesis, and evaluation. Bloom and his colleagues had as their goal to define behaviorally the cognitive processes involved in critical thinking, so that teachers could use the taxonomy to help ensure that, in fact, such thinking levels were being assessed.

In most of the literature, unfortunately, there is a dearth of evidence that any of the definitions provided are either reliable or valid. Two notable exceptions to this situation are the behavior analytic works by Semb and Spencer (1976), and by Johnson and Chase (1981), who were able to obtain high agreement (i.e., inter-rater agreement > 90%) on behaviorally defined levels of thinking behavior.

Semb and Spencer assessed recall (items where information has a point-to-point correspondence to text) versus more complex tasks, such as problem-solving (students identify principles or concepts in original, or novel, examples) and example-request questions (students must generate a novel example of a principle or concept). Semb and Spencer also suggested that, as a science, behavior analysis has the chance of reliably demonstrating effective methods for teaching higher-order thinking.

Johnson and Chase (1981) also achieved high levels of agreement on assessments according to their typology, which included elementary (i.e., lower-level) and conceptual (i.e., higher-level) forms of verbal behavior such as echoic, textual, transcriptive, intraverbals, tacts, or combinations of these categories of verbal behavior (see discussion of the relationship between verbal behavior and Bloom's taxonomy below). However, the forms of the response appear to mainly cover the use of examples, which would be indicative of Application-level answers in Bloom's taxonomy. It is also difficult to identify the amount of agreement between scorers at each of the various levels of the tasks, since agreement was reported as a combined measure.

Until recently, then, none of the literature reviewed suggests that reliable assessment, or demonstration, of higher-order thinking above the application level had been achieved. More recent research (Crone-Todd, Pear, & Read, 2000; Pear, Crone-Todd, Wirth, & Simister, in press), however, has established that a modified version of *Bloom's Taxonomy of the Cognitive Domain* (1956) yields relatively high inter-scorer agreement (ISA) values for within-group and between-group assessments of both short-answer questions and answers in second year undergraduate psychology courses. Consistent with the literature (e.g., Calder, 1983; Gierl, 1997; Kottke & Schuster, 1990; Roberts, 1976; Seddon, 1978; and Seddon, Chokotho, & Merritt, 1981), the researchers (Crone-Todd et al.) initially observed low ISA values when attempting to apply the taxonomy to assess study questions in three undergraduate courses taught using CAPSI. After they developed a modified, more stringent set of definitions for the six categories (i.e., Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation), the

overall ISA values were consistently at or above 80%. The use of the modified taxonomy for assessing answers (Pear et al.) produced similar results.

At the highest levels of the taxonomy (which were rare) in both studies, however, agreement between the scorers was lower. There are two possible reasons for the lower ISA values at the higher levels: (a) There was an initial lack of clarity in the distinction between Application and Analysis questions, which was later addressed by modifying the definitions for these categories; and (b) the highest of the levels, Synthesis and Evaluation, represented approximately 1% of the study questions, and thus there was infrequent practice in evaluating these types of questions using the modified taxonomy. The method also required that 100% agreement was arrived at through discussion after the initial IOR values were determined, which facilitated more discussion on the definitions used in the taxonomy, and a few modifications of these definitions at these higher levels.

This preliminary research (Crone-Todd et al., 2000; Pear et al., 2001) addresses some of the first steps identified by Williams (1999) for developing more precise measures of behavior typically studied in the cognitive area. By modifying the definitions and assessing agreement, the taxonomic levels can be used to identify whether an intervention produces any change in students' use of higher order thinking skills. By extending the modified taxonomy to the assessment of answers (Pear et al.), the operational definitions were analyzed in terms of reliability and validity. It seems possible that Bloom's taxonomy in its original form is too multifaceted to be manageable and reliable. It is also possible that the modified definitions used in the research reviewed are too restricted to adequately represent the constructs involved in higher-order thinking

(i.e., Rote Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation). The research presented here, carried out on students taking CAPSI-taught courses, extends the literature by using the operational definitions for assessment of questions and answers, and introduces a behavioral intervention to increase the proportion of students who can answer at the higher levels. Further, the use of answer keys arrived at by consensus between two scorers is consistent with arriving at agreement on what constitutes a particular behavior prior to observation; hence, the present research findings were expected to yield higher IOR values for answer assessment. The present research, in its use of the modified taxonomy, will also extend the research by helping to determine whether the definitions are too restrictive, too broad, or just right.

I turn now to a behavior analytic interpretation of the key concepts used in the present research; namely, "higher order thinking" and "feedback" (discussed later).

Behavioral interpretation of terminology. Perhaps one of the difficulties that lies with the assessment of "thinking skills" is that the term itself does not refer to an observable process or product that seems available for objective study. As mentioned above, however, Skinner certainly stated that thinking could and should be studied by behavior analysts. As has been pointed out (e.g., Semb & Spencer, 1976; Skinner, 1968; 1969), a behavioral analysis of the tasks involved in higher-order educational tasks involves describing (a) the stimulus situation, or context, in which a particular behavior (or set of behaviors) is required; (b) the target, or final desired, response for a given stimulus situation; and (c) the criteria used to assess, or judge, the behavior. In addition, the consequences following behavior need to be analyzed.

Bloom's Taxonomy. In a recent article, Crone-Todd and Pear (2001) made comparisons between Bloom's Taxonomy and behavior analysis. What follows is a summary of the comparisons of the levels, mainly based on Skinner's (1957) theoretical approach to verbal behavior.

Lower levels

Level 1 - Rote Knowledge

Here the question requires answers (i.e., textual behavior) that are word-for-word from the textual material. The answers can be in the form of intraverbal chains (i.e., verbal behavior in which the speaker's own verbal behavior controls subsequent verbal behavior, see Winokur, 1976), which may bear a point-to-point correspondence, or have similar phrasing, to the course material. Such answers involve echoic and intraverbal behavior: Echoics involve emitting the same verbal response as the verbal response presented as a stimulus. Hence, to answer the question, "Define positive reinforcer", a student would emit an echoic response "Positive reinforcer," which is combined in an intraverbal chain with other echoics: "is defined as 'an event that, when presented immediately following a behavior, causes the behavior to increase in frequency (or likelihood of occurrence)'" (Martin & Pear, 1999, p. 27). Verbal behavior in the form of echoics and intraverbals is likely a result of previous histories of reinforcement for emitting similar kinds of behavior when answering questions.

Level 2 - Comprehension

To put an answer in one's "own words", the learner must emit certain appropriate intraverbals from the text, and present (i.e., arrange) them in a manner consistent with the

language structure of the environment in which the person is emitting the behavior. While the form of the response differs somewhat from that presented in the text, there is an equivalence (Sidman, 1994) between the answer provided and the information in the text.

When a question such as "Describe an example of extinction from the text", certain key terms are required as part of the criteria for a correct response; however, a response that has the same configuration of intraverbals as that which is presented in the text would be unlikely. The answer, then, requires a response that involves "modified, yet thematically similar, arrangement of intraverbals" (Crone-Todd & Pear, 2001).

As the writer of the answer has many existing verbal responses available, there are many combinations of echoics and intraverbals that can be combined into autoclitic frames (i.e., putting the answer in one's own words), which are used to modify text.

Higher Levels

The following descriptions are of responses that are more complex; that is, they require elements that are combined in ways that are unique, or novel, given the material presented in the text. Hence, these are termed as higher levels.

Level 3 - Application

Similar to Semb and Spencer (1976), there are two forms of responses considered under this heading. First, lower-level knowledge may be applied to a novel situation or problem (Crone-Todd & Pear) in the sense that a principle, concept, or definition is identified in a novel situation. Second, the writer may be required to provide or identify a new exemplar in the form of intraverbal chains that are similar in form to a particular concept or

definition. The first form of response is similar to Semb and Spencer's (1976) problem-solving tasks; the second form of response is similar to their example-request tasks. In both cases, exemplars provided in the text provide potential responses in the form of autoclitic or relational frames[1] (Hayes, 1994; Barnes, Healy & Hayes, 2000); as such, these responses demonstrate an analogy between the principle, concept, or definition and the novel exemplar.

An example of the first type of question would be to ask a student to "Identify the US, UR, CS, and CR in the Little Albert (Watson & Rayner, 1920) experiment", when the text provides only a description of the experiment without reference to the respondent conditioning process involved. An example of the second type of question would be "Provide an example of respondent conditioning from your own life - one that is not provided in the textbook". In both cases, the student must provide a complex response that is based on a history that involves echoics, intraverbal chains, and autoclitic or relational frames related to respondent conditioning processes and definitions in order to emit a target response.

Level 4 - Analysis

An analysis question requires a response to "parts of a particular definition or concept" (Crone-Todd & Pear). In the vernacular, we might say that the writer must "compare and contrast" definitions or concepts. Comparison may also involve explaining how a novel

[1] Relational frame theory posits that relationships are learned between behavior and the world. One type of relationship, for example, would be recognizing "sameness" between two events/objects. Such a relationship is said to entail behavioral consequences such that the functions of one event/object are the same as the other. Relationships such as "sameness", "opposition", "bigger than", "less than", "equal", et cetera, are learned through the use of multiple exemplars. The relationships can be learned arbitrarily between objects/events, which include verbal and textual behavior.

exemplar fits the definition of a particular principle, concept, or definition. In order to engage in this complex behavior, all of the previous forms of response (echoics, intraverbal chains, and autoclitic or relational frames) are involved. Here, the writer uses autoclitic frames such as "X and Y are similar in that ... and they differ in that ..." , which is followed by a chain of intraverbal responses. The intraverbal responses are learned not from reading the text; rather, by emitting responses that are novel and have a desirable effect: Namely, a response that meets the criteria for verbal behavior that is recognized by the assessor (i.e., the marker) as being correct in terms of definitions, principles, or concepts.

Level 5 - Synthesis

A Synthesis question requires novel verbal behavior that is created by combining elements from various sources (i.e., multiple causation, Skinner, p. 422). Here, unique verbal responses, recognized by the assessor(s) as meeting the criteria in terms of definitions, principles, or concepts, are created by combining these various components in such a way that a desirable effect is achieved. Specifically, a particular problem may be solved (e.g., using behavioral principles to design a new course technology). In order for this complex behavior to occur, it presumably also involves analysis (e.g., identifying the situation in terms of its parts).

Level 6 - Evaluation

Considered the highest of Bloom's Taxonomy, answers at this level are typically what educators refer to as higher-level, or critical, thinking (e.g., Newman, 1991a, b). Skinner

stated: "An important part of scientific practice is the evaluation of the probability that a verbal response is 'right' or 'true' - that it may be acted upon successfully" (p. 428). In other words, the response involved in evaluation leads to further action, the evaluative response itself serves as both reinforcement for the behavior engaged in prior to the target response, and a stimulus for subsequent behavior consistent with the evaluation. An example of a question of this type would be "Given the controversy over the use of punishment, discuss whether there would ever be any circumstance under which electrical stimulation should be used to prevent harmful self-injurious behavior"? Here, the student will be required to "synthesize" the arguments presented in the text, and evaluate the probability of whether "yes" or "no" has the highest probability of being "correct".

Skinner also noted that the branch of philosophy concerning logic provides an "... analysis of the internal, and eventually tautological, relationships among autoclitic frames" (pp. 428-429). Hence, the criteria by which responses falling under this heading are assessed involves the degree to which the autoclitic frames correspond to the rules of logic. More recently, Hayes (1994) described this behavior in terms of relational frames that are complex and interconnected. As such, the novel verbal behavior that is emitted refers to elements that deal "with concepts, the world and the other" (p. 27).

Feedback. Peterson (1982) points out that "feedback" is not a precise term. That is, despite its ubiquitous use in behavior analysis, it can take on many forms that are not often specified. Feedback is a physical stimulus, which can take on the form of various types of stimuli to evoke operant (e.g., discriminative stimuli) or respondent (e.g.,

conditioned stimulus) behavior, provide conditioned reinforcement or punishment. As Peterson suggests, if feedback is effective, we must analyze why this is so. For the present study, "feedback" refers to an event in which textual information generated by a marker (i.e., the instructor, teaching assistant, or a student who has already passed the unit test) is presented to a student, usually within 24 hours, after submitting his or her written unit test. Additionally, feedback is provided within one week of the completion of a midterm examination. In this case, then, the feedback event occurs some time after the behavior of writing and submitting the test or examination. While it may be tempting to attribute the effectiveness of feedback to the process involved in reinforcement or discriminative stimuli, the fact remains that often feedback involves a long delay between the response and the stimulus; hence, we often cannot attribute the effect to these processes. Instead, some sort of verbal behavior - such as rules (Skinner, 1969) - are more likely to be operating. (An explanation in terms of operant principles is included in the methods section (below) in an effort to make the research presented more conceptually systematic with the field of behavior analysis.)

Rules are "description[s] … of a three-term contingency of reinforcement (antecedents-behavior-consequences" (Martin & Pear, 1999, p. 204). In essence, a rule specifies a behavior that will pay off in a given circumstance (or, alternatively, a rule specifies a behavior that will not pay off in a given circumstance). As Martin and Pear point out, rules are useful when a rapid change in behavior is desirable, the delay in consequences may be too long to directly reinforce a behavior, or when immediate inforcers are intermittent. The use of rules is an effective procedure in applied behavior modification programs with verbal individuals (e.g., Baldwin & Baldwin, 1986; Skinner,

1969; 1974). There are at least two reasons (see Martin & Pear) for why a student in a computer-mediated course may follow a rule such as "If you provide an original example on the midterm you will receive a bonus point". First, the student may make reinforcing statements to himself or herself while studying for, and writing, unit tests and exams. The student, after generating an original example, may verbally remind himself or herself that they will earn a higher grade in the course by using the original example. It is even possible that when they do not comply with the rule that this non-compliance will result in a form of self-punishment (e.g., "I *knew* I should have used an original example! Now I've lost points!").

Of course, another explanation could be that the student has a history of following rules simply because in the past they have been presented with reinforcers for following rules. In this case, then, the student will be likely to follow the rule provided by a marker because in the past following such rules has perhaps resulted in a higher grade (which is a desirable outcome for the student). Presumably, the use of feedback that specifies the desirable behavior, along with explicit information (i.e., verbal behavior) that describes the outcomes for the behavior will produce effective change in behavior if there is a good correspondence between the feedback provided and the instructions provided in the course (see Ribes & Rodriguez, 2001 for more information).

Description of PSI

To provide a context for the technology used in the current research, I turn now to a description of personalized system of instruction (PSI), the research carried out on its

various components, recent adaptations involving computer mediation, and its relationship to higher-order thinking.

Personalized system of instruction (PSI, Keller, 1966; 1968), or the Keller Plan, is a technology of instruction that had its roots in reinforcement theory and programmed instruction (Keller & Sherman, 1974). The idea for the technology had its basis in a meeting between Keller and his colleagues concerning the founding of the Department of Psychology at the University of Brasilia (Keller & Sherman). The combined factors of having the freedom and responsibility to create this new department, along with their general dissatisfaction with the traditional college lecture produced the necessary conditions for Keller and his colleagues to apply their experimentally based behavioral theories to teaching.

The main features of PSI-taught courses as explained by Keller (1968) included: (a) small units of study, which can usually be managed in a week or so; (b) mastery criteria, which require students to pass one unit before proceeding to the next; (c) student self-pacing, which allows students who learn at different rates to demonstrate mastery of the subject at different points in the course; (d) use of student markers (called proctors), who administer and grade unit tests and are available as a resource for students; and (e) material to be learned that is contained within written objectives, textbooks, and study guides, with lectures used for motivation or as a reinforcer for completing work (i.e., students could only attend if a given unit was already passed) rather than as a source of information.

Research on PSI. Since its inception, there have been a number of studies on the components involved in PSI-taught courses. To provide the reader with a background on the research concerning the main features of PSI, a review is provided here of several studies concerning unit size, mastery criteria, self-pacing, use of proctors, and objectives (for a more thorough review, see Semb, 1995). For example, by comparing sections in which the size of study units was determined by the students versus the instructor (or versus a traditional lecture section), Born, Gledhill, and Davis (1972) found that students performed highest on final examinations in a PSI-taught section in which the instructor determined the unit size. Hence, instructors appear to be better judges than students of what is manageable in terms of unit size for study in PSI-taught courses.

Student self-pacing was promoted by Keller (1968) as a way to allow students to learn at their individual rates; however, this aspect of the method is fairly controversial, and has therefore received a fair amount of attention in the literature. The controversy is probably due to administrators requiring that courses fall within an academic semester (Keller & Sherman, 1974) rather than allowing students to study at their own pace. Thus, any course taught within a specified time frame is considered to be a "quasi-PSI" (Keller & Sherman, p. 90) course. In these approaches, students who self-pace at a rate that does not allow for completion of all units are usually labeled as "procrastinators", and have been subject to a number of different attempts to change their study behavior. The findings, however, are mixed.

Lloyd and Knutzen (1969) found that when students were allowed to decide when and how many activities they would complete in a course, those who started later generally completed less work. Semb, Glick, and Spencer (1979) found that students who

withdraw from PSI-taught courses tend to have lower GPAs and to delay writing their first unit test. In addition, initial performance tended to predict later performance in the course. This led Semb et al. to suggest that intervention strategies be used on students who are performing poorly during the first few weeks of a course. In a follow-up study, telephone prompts and verbal agreements were incorporated to intervene with slow-moving students (Glick, Moore, Roberts, & Born, 1982). The majority of students increased their rate of testing with this intervention.

Several researchers (Crone-Todd & Pear, 1998; Henneberry, 1976) have found that students who are slow self-pacers tend to score lower on final examinations, and that students who withdraw from the courses typically do not complete much work (Pear & Crone-Todd, 1999). Some preliminary data (Crone-Todd & Pear) indicate that when students are provided with a unit completion graph (with suggested rate of completion), they complete more units when compared to previous sections of the same course.

Other research has studied the following aspects of pacing in PSI-taught courses: (a) contingency contracting for completion of unit tests (Brooke & Ruthven, 1984), and (b) entirely instructor-paced courses (e.g., Buerkel-Rothfuss, Gray, & Yerby, 1993; Caldwell et al., 1978), or (c) instructor-paced versus student-paced courses (Wesp & Ford, 1982). In all of these studies, it appears that some combination of instructor-paced study combined with flexibility for the student appears to accelerate movement through the course units of study and to improve performance on final examinations.

The mastery component has also been studied by a number of researchers. Keller (1968) originally intended the mastery criterion to be met by perfect or near-perfect performance on a given unit test. In other words, 100% correct answers must be provided

on a given unit before proceeding to the next one (McMichael & Corey, 1969). This criterion was not meant to be punitive: Students would not "fail" a unit test in the traditional manner, but would be permitted to take it over again as many times as necessary until mastery was achieved. Reviews and meta-analyses of PSI studies indicate that mastery is an essential feature for the efficacious use of PSI technology, if one's goal is to increase student performance on tests (Buskist, Cush, & DeGrandpre, 1991; Kulik, Jaska, & Kulik, 1978; Kulik, Kulik, & Bangert-Drowns, 1990; Kulik, Kulik, & Cohen, 1979; Lloyd & Lloyd, 1992).

It is presumed that if the mastery requirement were beneficial for students' learning of the course material, then they would perform at a higher level on a final examination. In fact, Bostow and O'Connor (1973) found that introductory educational psychology students enrolled in a mastery criteria section (little or no academic credit for performance at less than 90%, with extra credit available for taking a remedial quiz) performed higher on a final examination than students in a more stringent mastery requirement (as many tests as required for mastery), and had higher final examinations scores than students enrolled in a less stringent condition (only two tests allowed on a unit).

Another variation that includes a mastery requirement, but not necessarily rewarding it with academic credit, was used by Crosbie and Kelly (1993). In this research, students were required to obtain 90% on unit tests, and could retake tests as often as they liked to achieve mastery. However, only the first test score counted toward their final grade in the course. Crosbie and Kelly suggest that this approach may prevent students from simply taking a test without fully studying for it. While this is an

unorthodox approach, it may solve the problem of poor study habits suggested by Crist (1982) without resorting to lower mastery criteria.

Progression in PSI courses "involves a personal interaction between a student and his peer, or his better, in what may be a lively verbal interchange, of interest and importance to each participant. The job of the 'teacher' is to be a facilitator" (Keller, 1967, as reference in Keller, 1968, p. 84). PSI courses make use of proctors to facilitate interaction between students, and these proctors engage in the majority of interactions with students in the course. As Keller originally designed PSI, proctors were students who were recruited from the top of previously taught courses (i.e., "external proctors"). However, Keller and Sherman (1974) suggested that the use of students from within the course itself to serve as proctors (i.e., "internal proctors") makes a great deal of sense from both an administrative and a pedagogical viewpoint. That is, if a particular department does not have the resources to provide external proctors, or will not authorize an upper-level course to serve for credit as proctors, then an instructor may wish to use students in the course who have already demonstrated mastery at a given level.

As designed by Keller (1968), an external proctor would be responsible for 10 students in a course, thereby producing a student-peer ratio that is small enough to "personalize" the course. The proctor should support the goals of the course instructor (Johnson, 1977); however, maintaining quality control over proctors when they suffer disapproval, discontent, or special attention and appeals can be difficult (Caldwell, 1985; Hobbs, 1987; Johnson). Hence, Johnson suggested that proctors need training in such areas as evaluation, administration, tutorial instruction, and social interaction.

Methods have also been devised to select and train external proctors. Conrad and Semb (1977) moved from an intensive to a less intensive system of selection, training, and quality control which resulted in procedures that were simpler and more direct than previously used, while being more responsive to individual performance. Despite the attention paid to training external proctors, many authors state that little evidence suggests that proctors provide any observable benefit to student learning (Caldwell, et al., 1978; Croft, Johnson, Berger, & Zlotlow, 1976; Kulik et al., 1978). As mentioned previously, difficulties with funding and departmental support (Keller & Sherman, 1974) may also prevent an instructor from recruiting external proctors. These problems led Keller and Sherman to suggest that the use of internal proctors can be beneficial because internal proctors would presumably learn while helping fellow students, and would save administrative costs.

Keller and Sherman (1982) pointed out that internal proctors typically do not need as much time to review the material as external proctors do, since they have mastered the material more recently. Previous research (e.g., Johnson & Sulzer-Azaroff, 1975) indicated no significant difference in the performance of students who had feedback from external versus internal proctors. However, other research (Johnson, Sulzer-Azaroff, & Maass, 1976; Sheppard & MacDermot, 1970) found that when internal proctors were used student performance was higher. Johnson and Ruskin (1977) have suggested that this discrepancy is probably due to the difference in proctor selection: When proctoring was optional (Johnson & Sulzer-Azaroff), students were less likely to show differences in performance. Johnson and Ruskin further suggested that many interesting questions are available for future research, including (a) the amount of proctoring required to effect a

meaningful change in performance; (b) the change in students' performance relative to the amount proctored versus the type of quiz; (c) the amount spent proctoring relative to academic gain (and whether there is a ceiling effect on proctoring activity); and (d) the quality of students' evaluations (i.e., attitudes), and whether they change as a function of the amount of proctoring.

Work on CAPSI (Pear & Crone-Todd, 1999) has shed some light on these questions. First, the issue of how much proctoring is required is a difficult one to assess when the amount of proctoring is open-ended. That is, in CAPSI-taught courses the students can proctor or not, and can do so as long as the program assigns tests for marking. However, since the course grade is dependent in part on the amount of proctoring (i.e., 0.5 marks for each proctored test), students might use the extra gain in proctor points as a buffer against performing poorly on the final examination. One way to control for this behavior in future courses is to set an upper limit on the number of points that students can earn, while allowing unlimited proctoring opportunities.

As a partial answer to the second point raised by Johnson & Ruskin (1977), categorizing the unit and exam questions according to Bloom's taxonomy (e.g., Bloom, Hastings, Madaus, 1971) could identify the thinking level required by a question, and how proctoring certain levels of questions affects performance on similar or different levels of questions. As for the amount of academic gain relative to the time spent on the proctoring task, one way to investigate this question would be to devise a metric that assessed the gain in thinking level divided by the amount of time spent. However, this approach could be problematic for several reasons. First, the quantification of thinking levels (i.e., Bloom) might not be strictly hierarchical. Second, the amount gained by each

student is likely to vary and will not necessarily correspond to the time spent on proctoring. That is, some students might make very large gains relative to other students given the same amount of time spent proctoring. As with most of the findings for PSI-taught courses in general, those students who have the most to gain (i.e., initially low GPA) will likely show the greatest differences. While GPA and other pre-tests tap into various skills, they may not fully capture students' skills in a given discipline (Semb & Spencer, 1976). One alternative would be to use initial student performance on unit tests covering material early on in the course. This type of assessment could provide an investigator or instructor with information related to the skills students have mastered at the outset of a course, which can be used to direct research, teaching, or both.

Students who act as internal proctors tend to rate proctoring as very beneficial to their learning (Pear & Crone-Todd, 1999). Preliminary research indicates that the amount of proctoring students complete in our courses correlates positively with their final exam performance, which is consistent with earlier research (Gaynor & Wolking, 1974; Johnson et al., 1976; Sheppard & MacDermot, 1970).

While serving as internal proctors is related to student performance for the proctors, another issue is the quality of internal proctors' feedback. Several studies have included an assessment of and attempts to improve internal proctor feedback in PSI-taught courses (e.g., Robin & Cook, 1978; Robin & Heselton, 1977; Semb, 1975; Sulzer-Azaroff, Johnson, Dean, & Freyman, 1977). More recently a perusal of archived data (Martin, Pear, & Martin, 2001a, b) from a CAPSI-taught course sheds light on the accuracy and effectiveness of internal proctors in this environment. Martin et al. assessed the marking accuracy on correct and incorrect answers (i.e., pass provided for correct

answers, and restudy for incorrect answers), and the quality and accuracy of the content of the written feedback of those students who acted as proctors in the course. The reason the accuracy of the pass/restudy decision may differ from the accuracy of the written feedback is that students may provide accurate feedback in written form, yet still render a "pass" result when a restudy should have been assigned. When assessing correct answers, proctors were accurate 73% of the time; however, accuracy dropped to 40% when proctors assessed inaccurate answers. Since CAPSI-taught courses typically require two internal proctors to mark each test, the accuracy of the feedback on inaccurate answers increased to approximately 64%. In addition proctors were more accurate in their discrimination of incorrect answers when the answers had a higher degree of incorrectness compared to other answers. Interestingly, specific feedback on incorrect answers (e.g., giving the correct answer or suggestions on how to improve the answers) indicates a higher accuracy of 88%. Finally, the students who received feedback from either two proctors, the teaching assistant, or the instructor incorporated the suggestions in subsequent tests approximately 61% of the time. The analysis of proctor feedback and the accuracy of the decision when proctoring is important to study in these courses since students appear to be following the suggestions given by their within-course peers just as often as they would an instructor or teaching assistant. One question that can be examined is whether attempting to increase the higher-order thinking skills of all students in these courses would also raise the overall accuracy of internal proctors. One step in this direction would be to ascertain whether the thinking levels are increased as a function of feedback provided in the course. This is the purpose of the current study.

Recent Computer-Mediated Adaptations

Some writers (e.g., Buskist, Cush, & DeGrandper, 1991; Lamal, 1984) have suggested that PSI has fallen out of favor. This observation is based upon the decline in articles written on it since the mid-1980s. Lloyd and Lloyd (1986) sent out questionnaires to senior authors of PSI articles and psychology department chairs. Respondents indicated that fewer PSI courses were being taught, and that the ones in place were "SLIs" (Something Like It). Of the SLI courses, most incorporated study guides, frequent testing and immediate feedback. Problems identified with the use of PSI included time-consumption, difficulties with progression in one's career (e.g., merit pay, tenure, and promotion), and departmental pressure. As Lloyd and Lloyd point out, if a teaching technology is to survive, it will require positive feedback from students, colleagues, and administrators. Given these concerns, it was imperative to investigate methods of presenting PSI-taught courses in which technology (i.e., computers) can overcome many of the administrative difficulties.

Computer technology can be used to provide PSI as a component of an overall course (e.g., Pear & Crone-Todd, 1999; Crosbie & Kelly, 1993), or as an exclusive course method in itself (e.g., Pear & Crone-Todd, 1999; Skinner, 1990). Crosbie and Kelly (1993) used a computerized PSI program to test and provide feedback on fill-in-the-blank unit tests in a college-level course in applied behavior analysis. The findings indicate that (a) few students dropped out of the course, (b) most students met the instructor-paced deadline, and (c) a typical (for PSI courses) negatively - skewed distribution of grades resulted. Thus their utilization of computers to perform testing functions appears to be at least as effective as other PSI-taught courses.

Skinner (1990) used computer-based instruction to deliver tutorials to undergraduate students in a classroom management course. The findings of this study were that students typically performed at a higher level when the tutorials were available, and tended to use the tutorials even when they were not required. The students who were initially low performers gained the most in terms of academic progress. Hence, the use of corrective or self-selected computer-based tutorials may provide educators with an alternative to external proctors.

Joseph Pear and his colleagues (Crone-Todd et al., 2000; Crone-Todd, Holborn & Pear, 2001; Crone-Todd & Pear, 1998; Crone-Todd & Pear, 1999; Kinsner & Pear, 1990; Pear & Crone-Todd, 1999; Pear & Kinsner, 1988; Pear & Novak, 1996) have incorporated the use of computers in several CAPSI-taught courses at the University of Manitoba for a number of years. In general, the computer in these courses is used as a computer-mediated communication device which carries out most of the administrative functions of a PSI-taught course, such as (a) presenting short-answer essay unit tests and midterm and final examinations; (b) selecting either the instructor, teaching assistant, or two students who have volunteered to serve as proctors (i.e., students within the course who have demonstrated mastery on the unit to be marked); and (c) recording all work completed in the course.

Based on Keller's (1968) personalized system of instruction, CAPSI-taught courses have included the following components: (a) a mastery criterion; (b) a restudy is indicated when mastery is not demonstrated; and (c) a student who has already passed a given unit may volunteer to serve as proctors for other students' tests for that unit. Students type their answers into a response window directly below the question window.

Within a certain time limit students can edit their answers until they submit the test or exam (typically, 30-60 min for tests and midterms, and 120 min for examinations, respectively), and the student can cancel the entire test at any time. When the test is marked, a third window appears in which the marker(s) type their feedback. The test is marked as 'pass' or 'restudy' and electronically mailed to the student who wrote the test. The instructor of the course has access to all marked unit tests and examinations and can therefore carry out quality control interventions, such as providing information to proctors regarding their performance. In essence, in addition to serving instructional purposes, CAPSI may be viewed as a laboratory which combines teaching and research (Pear & Crone-Todd, 1999) to study the educational processes involved in student learning.

So far, using performance on a supervised final examination as the dependent variable (Crone-Todd & Pear, 1998), our correlational research has revealed the following: (a) students who start unit tests later in the term typically perform at lower levels than students who begin earlier; (b) the greater the number of proctored unit tests completed, the higher student performance is; (c) the greater the number of unit tests completed by the second midterm test, the higher performance is; and (d) the higher the score on the first midterm test, the higher performance is likely to be on the final examination. Interestingly, there was no correlation between the first and second term test, or between the second term test and the final examination, which may indicate that some students could be copying their answers on the second term test from the text or prepared materials. However, the students' performance on supervised examinations is consistent with findings in the literature concerning PSI components: Namely, students

who self-pace according to a quasi-imposed schedule tend to perform at a higher level (e.g., Brooke & Ruthvin, 1984; Wesp & Ford, 1982), and when students serve as internal proctors, they make academic gains (Gaynor & Wolking, 1974; Johnson et al., 1976; Sheppard & MacDermot, 1970).

Relationship between PSI and Higher Order Thinking

The use of CAPSI as an educational laboratory also lends itself to studying the effects of various pedagogical strategies aimed to increase higher-order thinking skills in students taking the courses. As mentioned, the research on teaching effectiveness demonstrates that PSI-taught courses result in examination scores that are higher than those in traditionally taught courses (Kulik, Kulik, & Bangert-Drowns, 1990). Reboy and Semb (1991) have also suggested that PSI is successful at generating higher order thinking, provided that the content of the course is designed in such a way to facilitate this process. As reviewed earlier, Semb and Spencer (1976) defined higher order thinking in terms of "low-level" and "high level" tasks. Low-level tasks were recognition and recall, whereas high-level tasks were comprehension, generalization, integration, and application. Interestingly, professors who were asked to estimate the ratio of higher- to lower-level tasks required in their courses tended to over-estimate by approximately 400% (i.e., 33% were estimated, and only 8% were assessed as being at the higher levels). Hence, a natural outgrowth of the PSI research literature is to use operationally defined indices of higher-order thinking to determine how effective CAPSI is in providing an environment in which such behavior can develop (see http://home.cc.umanitoba.ca/~capsi).

The Current Study

The present study used several reliably assessed measures of student learning. While examination scores provide some measure against which to compare students in different sections, they are at least somewhat subjective. For example, one instructor (or teaching assistant) may use a more stringent method of scoring examination questions, or may have relatively easier (or more difficult) examination questions, compared to another instructor. Hence, for this study, comparisons are made on the basis of independently scored (a) unit tests and midterms, (b) final examinations that are based on similar questions for two different sections of the same courses (c) levels at which students answer questions on midterm and final examinations according to the modified Bloom's taxonomy (Crone-Todd et al., 2000; Pear et al., 2001). The use of the levels identified in the modified taxonomy and the requirement that all inter-scorer agreement (ISA) values are consistently above 80% are required for determining the accuracy and validity of assessing whether differences exist in the proportion of higher order answers between two sections in which the course procedures changed.

The purpose of my research was to develop strategies to discover ways to optimize computer-mediated PSI-taught teaching and learning at the post-secondary level. That is, the strategies used in a quasi-experimental design were developed to encourage higher-order thinking. The independent variables were (a) feedback (i.e., prompts, praise, exemplars, and identification) to students on the level at which they answer questions on unit tests and exams, and (b) provision of bonus points for student answers on midterms and final exams above the specified minimum conceptual level. The dependent variables were (a) student performance on specific study questions,

relative to the minimum level at which the question can be answered, and (b) the score on test and exam performance.

METHOD

Participants

The participants for the main part of the study were 42 students who completed "Principles of Behavior Modification" courses (Course #017.244) in either the fall of 1999 (Control Group - Fall 1999, $n = 19$) or the fall of 2000 (Experimental Group - Fall 2000, $n = 23$) at the University of Manitoba. In addition, there were eight students enrolled in the same course at a small university in Southern California who interacted with the students in the Fall 1999 session; hence, their archived data are included for the purpose of determining the amount of feedback provided to the University of Manitoba students. Finally, archived data from 52 students enrolled in three courses during Fall 1999 and Fall 2000 were used for the purpose of comparing midterm examination scores under supervised versus unsupervised situations. The courses were: (a) Behavior Modification Application (Fall 1999: $n = 8$; Fall 2000: $n = 8$); (b) Foundations of Learning (Fall 1999: $n = 17$; Fall 2000: $n = 9$); and (c) Introduction to Systems of Psychology (Fall 1999: $n = 5$; Fall 2000: $n = 5$). All courses were used to investigate the principles and procedures involved in pedagogical effectiveness, and students were made aware of this fact in the course manual. In accordance with the University's Research Ethics Board (R.E.B.), students are protected against abuses of power in all courses. In particular, the University's policy of the "Responsibilities of Academic Staff with regard to Students" (R.O.A.S.S.) are strictly adhered to in the courses. The details of the study were explained to the students at the end of the course following the final examination (see Appendix A).

The mean (M) year in course of study for University of Manitoba students in Fall 2000 was 2.82 (SD = .18); for students in Fall 1999 it was 2.7 (SD = .18). A Mann-Whitney U non-parametric rank sums test demonstrated that there was no significant difference detected between the two groups in terms of year in course (z = - .645, p = .563; η^2 = .01). In addition, there was no significant difference in the amount of feedback provided by the California students versus the Manitoba students on any of the four measures of feedback (see Procedures section): (a) Identification (z = - .620, p = .547; η^2 = .02); (b) Praise (z = - .439, p = .697; η^2 = .01); (c) Prompt (z = - .584, p = .595; η^2 = .02); and (d) Exemplar (z = - 1.006, p = .336; η^2 = .05). Hence, the data reported for feedback on unit test includes feedback provided by these students, but does not distinguish it from the feedback provided by the Manitoba students.

Materials and Equipment

Study questions whose levels were assessed and agreed upon by independent scorers using a modified Bloom's Taxonomy (see Crone-Todd et al, 2000) formed the standard for the criteria, and minimum answer level, required to completely and correctly pass a given study question. The numeric values of the levels required to minimally answer the study questions were indicated in the respective course manuals. A copy of the summarized written instructions used to assess questions are included in Appendix B.

A computer-aided personalized system of instruction (CAPSI/PC) program was used to deliver all unit tests and examinations, and record all responses to student-requested unit tests and examinations; the data were archived for analyses after the course. All answers were scored by the author and by research assistants. There were two

research assistants, both of whom were third-year students who had completed the course under study, and who had been working on higher-order thinking research for at least six months. Score sheets (see Appendix C) were used to record the assessed levels of the questions and answers in the course. SPSS and Microsoft Excel © were used to perform all calculations, compute statistical analyses, and create graphs. A password-restricted Pentium II personal computer was used to store all of the archived data for analyses, and a program written for research purposes was used to view student information without access to identifying information. Data archived from the Fall 1999 and Fall 2000 courses which had all identifying information regarding course and scoring removed were used, and the midterm and final examination scores from six courses (two sections each, taught during Fall 1999 and Fall 2000, of Behavior Modification Applications, Foundations of Learning, and Introduction to Psychological Systems) also taught using the CAPSI method.

Procedure

Design

Due to the availability of sections taught and number of students available, a non-randomized quasi-experimental design was used. In this design, the students who were enrolled in the courses served in either the experimental (Fall 2000) or control course (Fall 1999) as a function of having registered for the course. No specific controls are used to ensure the equality of the groups from the outset; however, post-hoc scoring of the first two unit tests, and year in course, by term was conducted to provide a measure of equality of the groups.

During the Academic Session

Minimum levels. Prior to the first day of classes in the Fall 1999 session, the course manuals were available for students in the university bookstore. The minimum level at which a given question could be answered correctly was included in the course manuals, and was based on a modified version of Bloom's Taxonomy (Crone-Todd et al, 2000), which consisted of six levels:

1. *Rote Knowledge*, which involves answers that appear to be memorized or closely paraphrased from material in the text;
2. *Comprehension,* which involves answers that are in the student's own words, but still adhere to terminology that is correct and appropriate to the course;
3. *Application*, which involves the identification, or use of a particular concept or principle in a new situation or to solve a new problem;
4. *Analysis*, which involves breaking down concepts into their constituent parts, such as in contrasting or comparing concepts or explaining how an example illustrates a given concept;
5. *Synthesis*, which involves putting together parts in a unique way to form a whole.

   This level requires that a definition is generated from examples or descriptions, or the explanation of how to combine principles or concepts in a novel way; and
6. *Evaluation*, which requires that reasons are presented and evaluated for or against a particular position, and that some conclusion is reached about the validity of a given position.

A good discussion has no correct answer per se, but involves the use of all of the preceding levels.

Note that Rote Knowledge (Level 1, also referred to as "knowledge" at various points in this manuscript) and Comprehension (Level 2) answers may be directly obtained from the material in the text; however, Application (Level 3), Analysis (Level 4), Synthesis (Level 5), and Evaluation (Level 6) answers make an inference or extrapolation about what is presented in the text. Hence, for the purpose of this study, "higher-order thinking" is defined as any answer that is assessed at Level 3, 4, 5 or 6. Due to the small number of Level 5 and 6 questions, Levels 5 and 6 were scored as one category (i.e., Level 5-6). Similarly, answers at Levels 1 and 2 constitute "lower-order thinking", and were also scored as one category (i.e., Level 1-2).

The minimal level for each question in the course manuals was independently scored by two different groups of assessors (Crone-Todd et al., 2000). Inter-group agreement was over 80%. The high level of inter-group agreement therefore suggests that the questions are valid indicators of the level of thinking required to answer questions in the courses.

Direct teaching strategies. The instructor and teaching assistant for the Fall 2000 course provided feedback to students on midterm and final examinations, and on unit tests not assigned to proctors. The instructor and teaching assistant included statements in their feedback to prompt the student to answer the questions at or above the assessed levels. The inclusion of these statements controlled for variations in instructor - or teaching assistant - provided instructions that may differentially affect students, yet still

allowed for different statements that pertain to the question being marked. For example, if the student answered a question at Level 2 (Comprehension), then the feedback might contain a verbal prompt to provide an original example for this level of question on the next unit test written. In this case, the intended use of the feedback would be to provide a prompt to answer at Level 3 when a Level 2 question is asked. If the answer were above the minimum level specified, then the feedback would include praise regarding which part of the answer was exemplary. For example, "Excellent original example!" might be used when a question asked for a definition only (i.e., Levels 1 or 2), because giving an original example would raise the answer to a Level 3. This type of praise was also used when a Level 3 question was answered correctly. In these cases, the intended use of the feedback would be to act as praise for answering questions at levels which indicate higher-level thinking. Also, if the answer were below the assessed minimum level, then the feedback would contain comments that indicate the information needed to correctly answer the question. In the cases where an answer was not correct or complete at a lower level, feedback may not have indicated reference to higher-order thinking because the lower order question had not yet been mastered.

During the term for a given course, students in both courses could complete up to 10 mastery-criterion (i.e., must be complete and correct) unit tests worth 1 point each, consisting of 3 randomly-selected questions from the study questions (typically 20 questions per unit) provided in the course manuals. Students received feedback comments from an instructor, a teaching assistant, or two students who had already demonstrated mastery on the given unit test. Midterm examinations consisted of three randomly-selected questions in the following manner: (a) Midterm 1 consisted of one question each

from Units 2, 3, and 4; and (b) Midterm 2 consisted of one question each from Units 5, 6, and 7. Students received feedback from the instructor on marked examinations within one week of writing their examinations. The final examination for Fall 1999 consisted of 10 randomly selected questions from Units 2 through 10, inclusive; however, the final examination for Fall 2000 consisted of non-randomly selected questions. Specifically, the final examination questions for Fall 2000 were the same questions as those presented to students in Fall 1999. All tests and examinations were requested and submitted on the computer.

Students in the Fall 2000 courses were presented with 0.5 bonus points, and a positive statement (e.g., "Above and Beyond!") for each examination (on both midterms and the final) answer that was above the minimum level, as long as the answer was minimally at Level 3 (Application). The reason for using Level 3 as a minimum level for bonus points was to encourage higher-order thinking, or "thinking outside the text". Hence, the following indicates the rules by which students could earn points for their answers:

Assessed Question Level	Minimum Answer Level for Bonus
1-2	3
3	4
4	5
5	6

Note that if students answered a question assessed at Level 6, then no bonus would be provided for answering at Level 6.

Students in the Fall 2000 course were required to write their first midterm under supervision. The supervised midterm made it possible to investigate whether students' answers differed both in terms of level and total score during supervised versus

unsupervised exams. For example, this intervention allows an assessment of whether students answer at higher levels on lower-level questions, or at lower levels on higher-level questions, when under supervised versus unsupervised conditions. Hence, the relative level and accuracy of students' answers are compared during both conditions for each individual student. Accuracy was measured by two independent scorers assessing answers as complete and correct versus incomplete or incorrect (see Appendices D and E for instructions to scorers). These answers had been edited to the extent required to render the scorers "blind" to the course location of the particular answers to be scored.

To establish higher inter-scorer agreement (ISA) values, the research assistants were trained to assess answers to sample questions (see Pear et al, 2001). An answer key was developed and agreed upon by the scorers prior to the actual assessment, and a flowchart of the decisions for assessing answers according to the modified Bloom's taxonomy was used by the scorers to assess the levels of the answers (see http://home.cc.umanitoba.ca/~capsi to view the flowchart). To evaluate the ISA values on levels of answers, a point-to-point agreement ratio (Kazdin, 1982) was used. That is, for each component of a question, the answer was assessed according to the MBT, and each component (as previously established by Crone-Todd et al., 2000) was scored as "agree" or "disagree". The total number of agreements included agreements on both occurrences and non-occurrences of a given level, and was divided by the total number of components. A minimum of 80% point-to-point agreement was required by scorers on practice sets prior to evaluating student answers. Periodic checks (i.e., 40% of scored items) of ISA assessments were carried out, with the requirement that the value

consistently be at, or above, 80%. Scorers had to complete three practice assessments in a row that yielded ISA ≥ 80% prior to assessing the data used for this research.

In addition to point-to-point agreement a kappa statistic was calculated, which corrects for chance agreement. That is, since the number of assessed components is fairly large, for any given answer level there are also a high number of non-occurrences of that level. For instance, if all of the levels were equally represented, then there would be a 1/6 chance that a given level occurred. Further, there would be a 5/6 chance that it did not occur – hence, agreement on the non-occurrence of a given level is more likely than agreement on the occurrence. By chance alone, the point-to-point agreement is likely to be greater than 80%. (This logic also applies to the existence of four levels.) The kappa statistic corrects for agreement based on chance, and an interpretation of the kappa value provides a guidepost of how much above-chance the agreement is. Hence, all ISA values are reported with their associated kappa value and interpretation.

The operational definitions for different levels of answers according to the MBT are summarized as follows:

- Incorrect (Level 0)

The answer contains incorrect terminology, is incomplete, or both. Answers of this type cannot be scored as correct according to the following categories; hence, they are scored as "0".

- Rote Knowledge (Level 1)

The answer is wholly from the text, uses appropriate terminology, and has a point-to-point correspondence to the textual information. The answer can be paraphrased or

synonyms for words or phrases that are not key to answering the question are used (e.g., "and" for "also"), but there is no new information added, nor any interpretation provided.

- Comprehension (Level 2)

The answer is wholly from the text, uses appropriate terminology, but has some portion that is written in the student's own words. An answer of this type requires a grammatical shift which rephrases the answer. Key words or phrases may be substituted.

- Application (Level 3)

The answer is not wholly from the text, but is based upon concepts, principles, or processes provided in the text. Most often a correct example was provided of a particular concept (e.g., provides an original example of positive reinforcement), or was correctly identified from a description not provided in the text (e.g., the student identifies a general principle in a specific experiment).

- Analysis (Level 4)

The answer is not wholly from the text, but involves concepts provided in the text. Most often concepts are compared or contrasted (i.e., similarities or differences are identified), or an explanation was provided about why an example is one of a particular type.

- Synthesis (Level 5)

The answer is not wholly from the text, but involves concepts, principles, or processes provided in the text. The concepts are combined in a new way that was not covered in the text. This requires a novel combination of concepts or principles to solve a problem (e.g., the student describes how one might use computer technology to help shape the limb movement of a person in a rehabilitative setting). Synthesis may also involve deriving a definition of a concept from general examples or discussion.

- Evaluation (Level 6)

The answer is not found wholly in the text. Evaluation answers are identified through reasoned argumentation for or against positions that may have been (but are not necessarily) discussed in the text. The answer should contain relevant points discussed in the text, provide arguments for or against those points, and come to some conclusion (even determining that a conclusion is not warranted may be a satisfactory conclusion).

Answers that resulted in disagreement between scorers were discussed by the scorers, and final agreement was 100%. Following consensus, the value for each question was divided by the indicated number of components (e.g., 5.00 points divided by 4 components, resulted in 4 components worth 1.25 points each), and scoring of the unit tests and exams was completed by assigning the points for number of correct components. Keeping with the original scoring used in the courses, answers to the three midterm examination questions were scored out of a total of 5 points each (15 points in total) and answers to the 10 final exam questions were scored out of a total of 6 points each (60 points in total).

The examinations written by 42 students resulted in 132 answer components for inter-scorer assessment. The author scored all of the examination answers, which were divided into sub-components according to the scoring using the modified taxonomy (Crone-Todd et al., 2000), which yielded a total of 2199 components for assessment of the level at which an answer occurred. A random number table was used to select examinations from 40% of the students (i.e., 8 students from Fall 1999 and 10 students from Fall 2000), which yielded a total of 54 examinations. These 54 examinations, once subdivided, yielded a total of 706 components for assessment. Table 1 shows the number

and percentage of answer levels as assessed by each observer, along with the percentage of interobserver agreement and the kappa statistic along with the interpretation (Landis & Koch, 1977). Note that *all* point-to-point interobserver agreements are over 80%, and the kappa values are moderately to substantially above chance. Hence, there is a high amount of agreement between the two scorers.

Table 1

Number and Percentage of Answer Levels for Each Observer and Measures of Inter-Scorer Agreement

Answer Level	Observer A Number of Occurrences	Observer B Number of Occurrences	Number of Agreements	Percentage of Inter-Scorer Agreements	Kappa (Interpretation)
		Assessment (Number of Components = 706)			
0	378	417	623	88.24	.76 (Substantial)
1-2	177	156	663	93.91	.57 (Moderate)
3	108	107	653	92.50	.71 (Substantial)
4	36	22	684	96.88	.61 (Substantial)
5-6	7	5	699	99.01	.41 (Moderate)

To assess the feedback on unit tests and examinations, five categories of feedback were developed as follows:

- Identification

This type of feedback is a verbal statement about the level at which an answer is provided. This may occur in the following circumstances: (a) when the actual level, or definition of the level, is written (e.g., "This answer is at a Level 3" or "An original answer was provided"), or specific information about why an answer is not at a higher

level (e.g., "The answer requires that you state the similarities and differences" or "The answer requires an original example").

Presumably, identification feedback serves a cue for future responses on tests. That is, rule-governed behavior may operate, in which the student might repeat the statement to him/herself when writing future tests and examinations.

- Praise

This type of feedback is a written word that provides approval or a compliment about something specific in the answer (e.g., "Good example!" or "Nice analysis!"). Where "good example" or other feedback does not clearly indicate whether the answer is at a higher level, the answer level must be taken into account to determine the condition under which the feedback is provided (i.e., whether the praise is for higher-level answers or not).

Presumably praise serves as a delayed reinforcer, which is mediated again by rule-governed behavior. If the student were to repeat the praise to themselves when studying or writing tests and exams, it may also serve the function of a stimulus to respond at a higher level. For example, when a student receives praise for providing an original example on a test, the student may form the rule "It is good to use original examples, and I will earn bonus points on exams for doing so", which may be repeated while studying.

- Prompt

This type of feedback directs or provides hints for answering at or above the level provided in the answer (e.g., "Can you think of an original example?", or "Next time try to explain why the answer is one of this type"). Feedback was considered a textual

prompt if it was provided to help (a) raise the level of the answer, or (b) correctly answer a higher-level question.

Again, prompts likely serve as stimuli to evoke certain verbal responses while studying for, and writing, tests and exams. More generally, prompts may generalize to other learning situations where the student may repeat to themselves "How are these elements the same or different?", and so forth.

- Exemplar

This type of feedback provides a written specimen or model of what could be included in an answer. Such feedback is scored as present if the exemplar is provided to answer a higher level question (e.g., "One similarity between the two concepts is ... ") or to answer a lower level question at a higher level (e.g., '"One original example could be...").

An exemplar presumably provides an intraverbal chain that may be repeated on subsequent examinations. Presumably the marker would provide this form of feedback when the verbal response in an answer is weak in form. The student, as both speaker and listener (Skinner, 1957), may repeat the exemplar a number of times until it becomes a strong response - again, the exemplar serving as a stimulus for future test writing, and perhaps as a reinforcer for the correct response being emitted either privately to oneself (immediate reinforcement), or as a delayed reinforcer for answers on tests and examinations.

- General

This type of feedback is provided without reference to any specific part of the answer per se. The feedback may be positive ("Good answer") or not ("Needs work"). As a "global" form of feedback, it does not serve as a cue, or as a form of reinforcement, for any

specific verbal response emitted by the student. Hence, it is difficult to ascertain which aspects of the verbal behavior in an answer would receive reinforcement or prompts for behavior to be emitted on subsequent tests and examinations, unless this form is combined with one of the four types of feedback identified above.

The total number of unit tests and examinations written by the students in the Fall 1999 and Fall 2000 sessions of the Behavior Modification Principles course was 632, which resulted in 2608 opportunities for feedback (one opportunity for each question and one in the General Comments section). Each of the 632 tests and exams was considered as an opportunity for feedback, however, such that any instance of any category of feedback provided was counted as having occurred for a given test or exam. The author scored all of the test and examination feedback, and a random number table was used to select 25% of the instances (i.e., 158) for interscorer assessment by a paid research assistant. Table 2 shows the number and percentage of feedback types as assessed by each observer, along with the interscorer agreement and the kappa statistic along with the interpretation of kappa (Landis & Koch, 1977). Note that all of the point-to-point interscorer agreements are over 80%, and the kappa values are *all* interpreted as "almost perfect". Hence, the assessment of type of feedback yielded a very high amount of agreement between the two scorers.

Group analyses. The difference (*D*) between the levels of the answers and the questions was calculated by subtracting the answer level (as assessed by two independent observers) from the minimum required level for each component of a question. For example, if a given answer was assessed at Level 3, and the question component was assessed at Level 2, then $D = 3 - 2 = +1.0$. Alternatively, if the reverse were true (a Level

2 answer to a Level 3 question), then $D = 2 - 3 = -1.0$. The overall data from the courses was compared within sessions, and between sessions to determine whether there were any consistent differences between the courses. Since the participants in this study were not randomly assigned to the courses, tests of differences between the groups of students in both courses

Table 2

Number and Percentage of Feedback Types for Each Observer and Measures of Inter-Scorer Agreement

Feedback Type	Observer A Number of Occurrences	Observer B Number of Occurrences	Number of Agreements	Percentage of Between-Scorer Agreements	Kappa (Interpretation)
Practice Assessment (Number of Components = 51)					
Identification	10	10	48	94.12	.91 (Almost Perfect)
Praise	11	8	46	90.92	.83 (Almost Perfect)
Prompt	7	5	48	94.12	.84 (Almost Perfect)
Exemplar	5	4	50	98.04	.88 (Almost Perfect)
General	26	25	49	96.08	.85 (Almost Perfect)

were conducted to determine whether any significant differences existed on tests of initial performance as measured by scores on the answer to the first two unit tests. Since the data are nominal, rather than interval or ratio, (i.e., levels of Bloom's taxonomy are not necessarily hierarchical), the data were analyzed using the non-parametric Mann-Whitney

rank sum test. The rank sums test is analogous to the parametric *t* test, and is used when there are two independent samples of ranks and either group has a sample size of $n > 20$. The *z*-scores are reported, along with the *p* values and effect size (η^2) for each of the tests.

Results

Table 3 shows the mean percentage correct scores for the first and second unit tests, midterms, and final examinations, by course. Of particular note is that there are no statistically significant differences in scores for Unit tests 1 and 2, or for Midterm 1, between the two courses. Interestingly, the mean score on Unit 1 tests for Fall 2000 (M = 82.70) students was 2.80% higher than the mean score on Unit 1 tests for Fall 1999 (M = 79.90) students; however, Fall 2000 students scored (M = 77.53) 5.63% lower on average for Unit 2 tests than Fall 1999 students (M = 83.16). Hence, the students in the experimental course scored at, or slightly below, the level of the students in the control course on these initial measures. This suggests that there is no apparent difference between the two groups at the outset of the course. Note that the mean scores in the Fall 2000 course are significantly higher for the second midterm (z = -2.17, p = .03; η^2 = .11) than the mean scores in Fall 1999. Also, the final exams scores are significantly higher for the students in Fall 2000 than for those for Fall 1999 (z = -2.401, p = .016; η^2 = .14).

There are no significant differences (z = - .717, p = .473; η^2 = .01) between the courses on the scores on Midterm 1, or on either Unit test. Further, note that while there are no significant within-course differences between Midterm 1 and Midterm 2 scores for students in Fall 1999 (z = -.491, p = .624; η^2 = .02) and in Fall 2000 (supervised, z = -904, p = .366; η^2 = .04), students in the Fall 1999 session tended to score lower on Midterm 2, while students in the Fall 2000 session tended to score higher on Midterm 2 (which was unsupervised).

Table 3

Mean Percentage Correct Scores on First and Second Unit Tests, Midterms, and Final Examination

	Course		Difference
	Fall 1999 *M* (*SD*)	Fall 2000 *M* (*SD*)	(Fall 2000 – Fall 1999)
Unit Test 1	79.90 (17.61)	82.70 (16.24)	2.80
Unit Test 2	83.16 (17.41)	77.53 (22.53)	-5.63
Midterm 1	71.97 (24.34)	74.75 (25.29)	2.78
Midterm 2	65.52 (22.22)	79.59 (21.69)	14.07*
Final Examination	35.63 (25.48)	53.76 (23.55)	18.13*

* Significant at $p < .05$

Figure 1 shows the proportion of questions asked at the various levels on Midterm 1, Midterm 2, and the Final Examination, by course. Note that the questions for Midterm 1 in both courses contained a higher proportion of lower-order question levels (i.e., Levels 1 and 2) relative to the proportion of higher-level questions (i.e., Level 3 or Level 5) in the Fall 1999 course. Hence, a small portion of the exam for Fall 2000 required higher level answers. On Midterm 2 there were more higher-level questions for Fall 1999 than for Fall 2000; however, on the final examination, the proportions of higher- versus lower-level questions were approximately equal in the two courses. If students were responding only at the level of the questions, they would be expected to answer similarly on Midterm 1 and the Final Exam. Given the higher proportion of higher-order questions

on Midterm 2 for the Fall 1999 course, however, students in that course would be expected to provide slightly more higher-level answers on Midterm 2.

Figure 2 shows the proportion of students who provided at least one answer at each of the assessed levels on a given exam. Note that a higher proportion of students in Fall 1999 provided answers to questions at Level 0 (95%) on Midterms 1 and 2 than students in Fall 2000 (74%); however, more students in Fall 2000 provided at least one answer at Level 0 (100%) on the Final Exam than students in Fall 1999 (85%). Also, Students in Fall 1999, despite having no higher-level questions on Midterm 1 were just as likely to provide answers at Levels 3 and 4 as students in Fall 2000.

Note that students on average in Fall 1999 provided slightly more answers at Levels 3 and 4 for Midterms 1 (Level 3 = 85%; Level 4 = 30%) and 2 (Level 3 = 80%; Level 4 = 35%) than did students in Fall 2000 (Midterm 1: Level 3 = 87%; Level 4 = 22%; Midterm 2: Level 3 = 70%; Level 4 = 30%); however, only students in Fall 2000 provided answers at Levels 5 and 6 on these exams. While a small proportion of questions were presented at Level 5 on Midterm 1 for Fall 2000 students, the presence of answers at these levels on both midterms indicates that students were answering at a higher level than asked. In addition, many students in Fall 1999 did not provide Level 3 and 4 answers on Midterm 2, despite the fact that more of these questions were present on that exam. Finally, all students in Fall 2000 provided answers at Levels 3 and 4 on the Final examination, whereas approximately 75% of students in 2000 provided answers at Level 3, and approximately 46% at Level 4. Hence, more students in 2001 provide higher-level answers, despite the fact that both courses were approximately equal in terms of the question levels on the final examination.

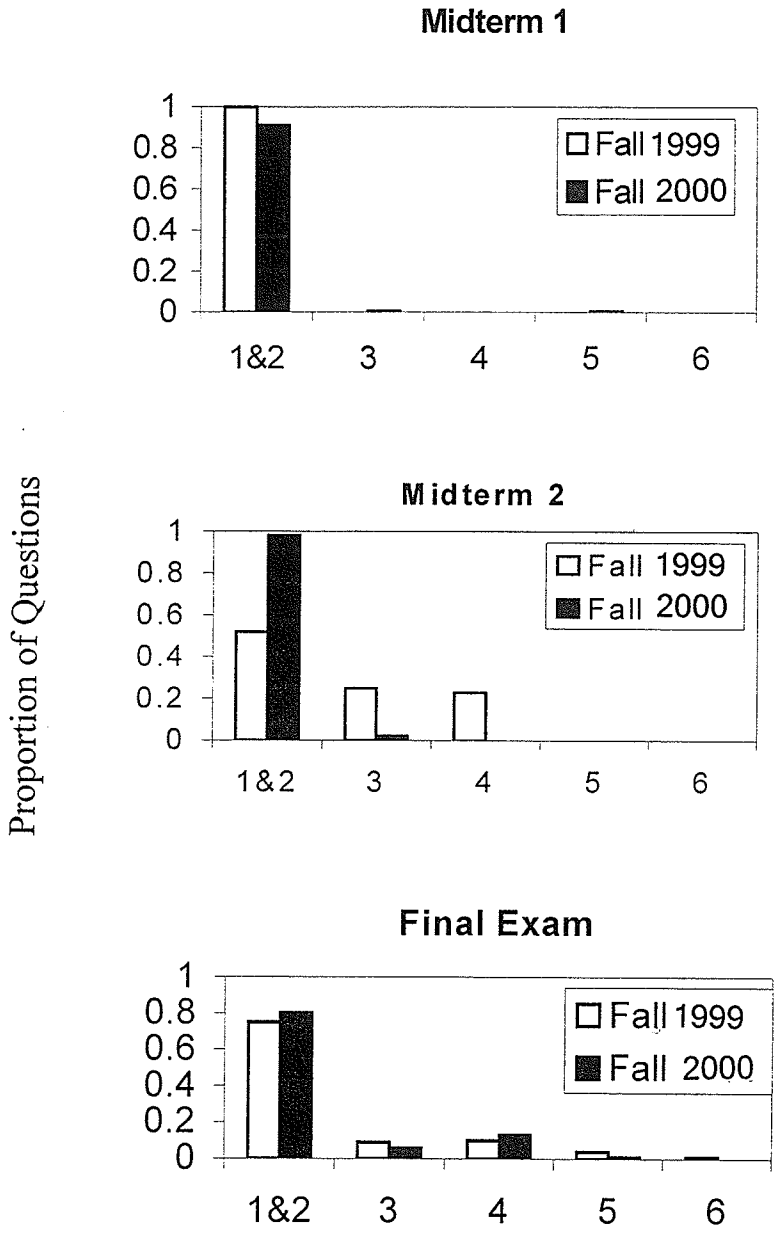

Figure 1. The proportion of questions asked, by level, on midterm and final examinations for Fall 2000 and Fall 2001 courses.

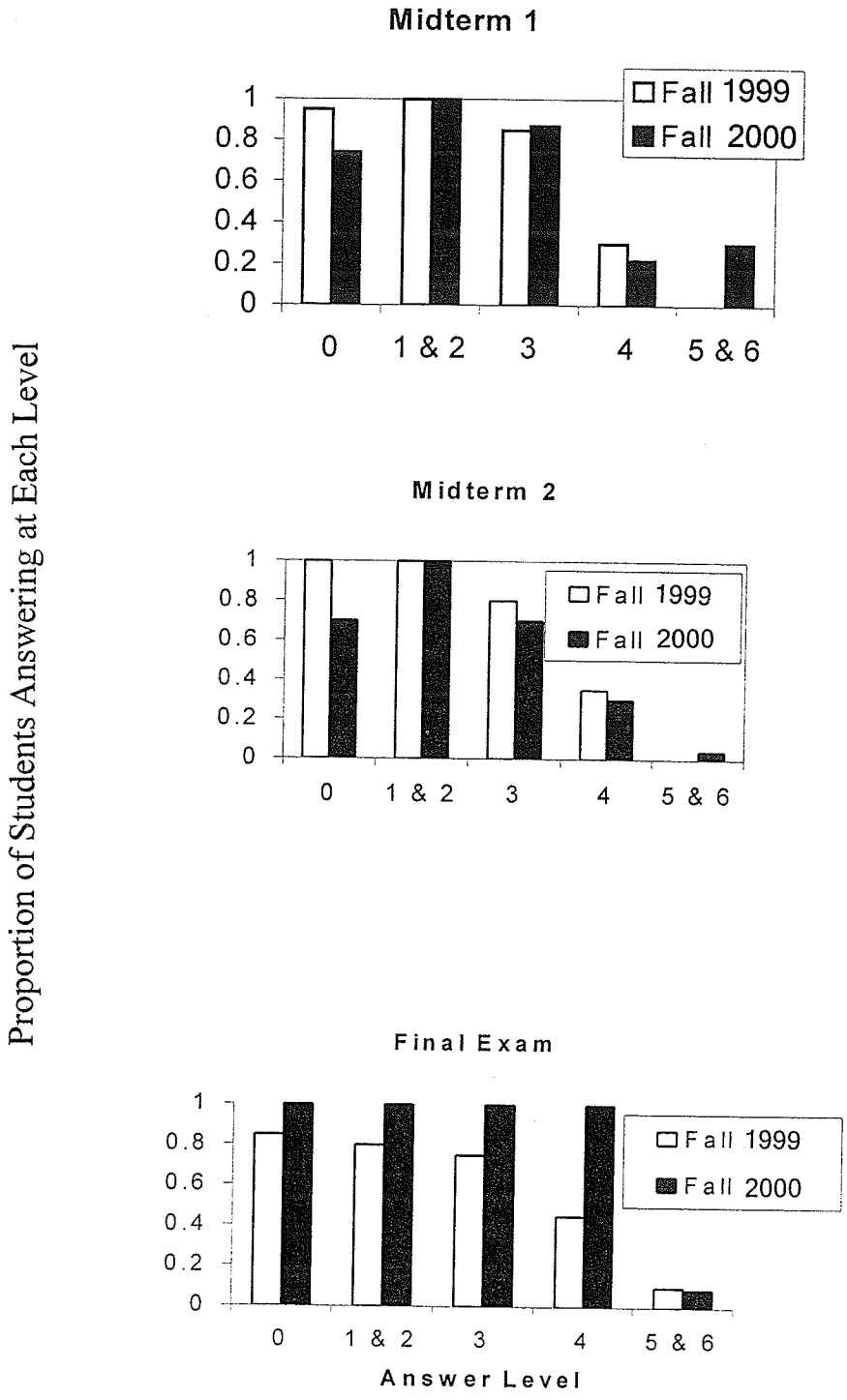

Figure 2. The proportion of students, by course, answering at each of the levels on both midterms and the final examination.

Figure 3 shows the mean proportion for the measure of *D* (i.e., the difference score calculated by subtracting the answer level from the question level) on the final examination. Note that the proportion of answers on the final examination that do not meet the minimum level (i.e., $D = 0$) for questions in Fall 1999 are generally approximately the same as or higher than those in Fall 2000. Of the answers that did not meet the minimum level, only the differences between courses in terms of incorrect answers to Level 5 questions (i.e., $D = -5$) are statistically significant ($z = -2.799$, $p = .005$; $\eta^2 = .19$). In terms of higher-level answers, the mean percentage of answers at one

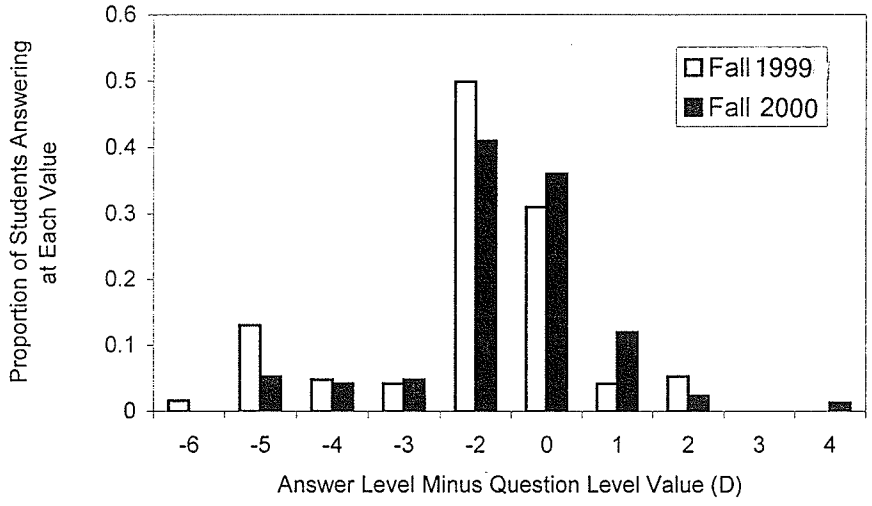

Figure 3. The proportion of students' answers on the final examination, by course, measured by *D*, or the level of the answer minus the level of the question. Students in Fall 2000 tended to provide more answers at higher levels than asked, and fewer answers at levels lower than asked.

level above the question on the final examination was higher for students in Fall 2000 (M = 11.96%) than for students in Fall 1999 (M = 4.26%), and the difference is statistically significant (z = -3.29, p = .001; η^2 = .26). Also significant (z = -2.992, p = .003; η^2 = .22), is the finding that the Fall 1999 students provided more answers at two levels above the minimum level (M = 5.26) compared with students in 2001 (M = 2.35). Finally, significantly more (z = -3.28, p = .001; η^2 = .25) answers to higher level questions were assessed as correct for students in Fall 2000 (52.43%) than for students in Fall 1999 (20.89%).

A chi-square goodness of fit analyses was conducted on the frequencies of correct answers to higher-level answers, by course. With an alpha level of .05, the observed frequencies were statistically significant at Level 3, χ^2 (1) = 8.19, p < .01, and at Level 4, χ^2 (1) = 7.43, p < .01. The observed frequencies were not statistically significant at Levels 5 and 6, χ^2 (1) = .40, p > .05. An additional analysis (see Table 4) of the type of answer that students provided above the minimum level asked on the final examination yielded the following results. Students in both courses were most likely to provide an answer above the level asked by giving a Level 3 answer to a Level 2 question (2001 M = 95.65%; 2000 M = 68.42%) or by giving a Level 4 answer to a Level 2 question (2001 M = 65.22%; 2000 M = 21.05%). Note that a relatively small percentage (approximately 13%) of students in Fall 2000 provided Level 5 or 6 answers to Level 2 or Level 3 questions; however, none of the students in Fall 1999 did so. Further, 22% of students in Fall 2000 provided a Level 4 answer to a Level 3 question. However, none of the Fall 1999 students provided answers at these levels for questions at similar levels. Hence,

some of the students in Fall 2000 provided a greater range of higher-level answers to various level questions, relative to students in Fall 1999.

Table 4

Number and Percentage of Students Providing Higher-Level Answers to Various Levels of Questions on Final Examination, by Course.

Answer Level	Question Level	Course	
		Fall 1999 (n = 19) Number of Students (%)	Fall 2000 (n = 23) Number of Students (%)
5 or 6	5 or 6	2 (10.52)	4 (17.39)
	4	0	0
	3	0	1 (4.35)
	1 or 2	0	2 (8.69)
4	4	8 (42.11)	19 (82.61)
	3	0	5 (21.74)
	1 or 2	4 (21.05)	15 (65.22)
3	3	4 (21.05)	15 (65.22)
	1 or 2	13 (68.42)	22 (95.65)

Figures 4, 5, and 6 show, by percentile rank on final examination scores, the percentage of tests or exams in which the various types of feedback were provided on all unit tests and midterm examinations by course. The data presented in Figure 4 show that all students in both courses received a high amount of non-substantive general feedback (e.g., ''Good'' or ''Incorrect'', without specific reference to elements of the answer) on

their tests, and that students in Fall 2000 were more likely to have received at least one form, or unit, of higher-order feedback. Figure 5 shows that students in Fall 2000 received more specific types of feedback on tests, specifically: identification, praise, prompts, and exemplars. Figure 6 shows that students in Fall 2000 also received more specific types of feedback on their midterm examinations, when compared with students in Fall 1999.

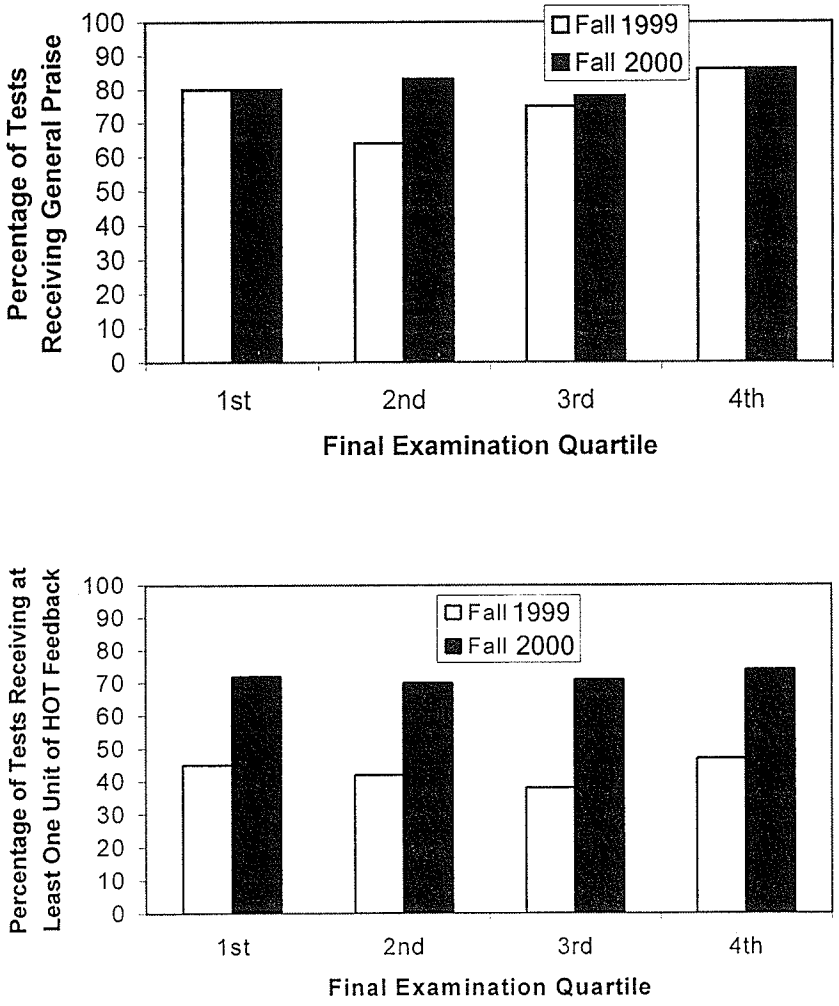

Figure 4. The percentage of students' unit tests, by course, in which students received feedback. The first graph shows the percentage of tests receiving general, non-specific, praise. The second graph shows that all students in Fall 2000 received more instances of higher-order thinking (HOT) feedback on tests.

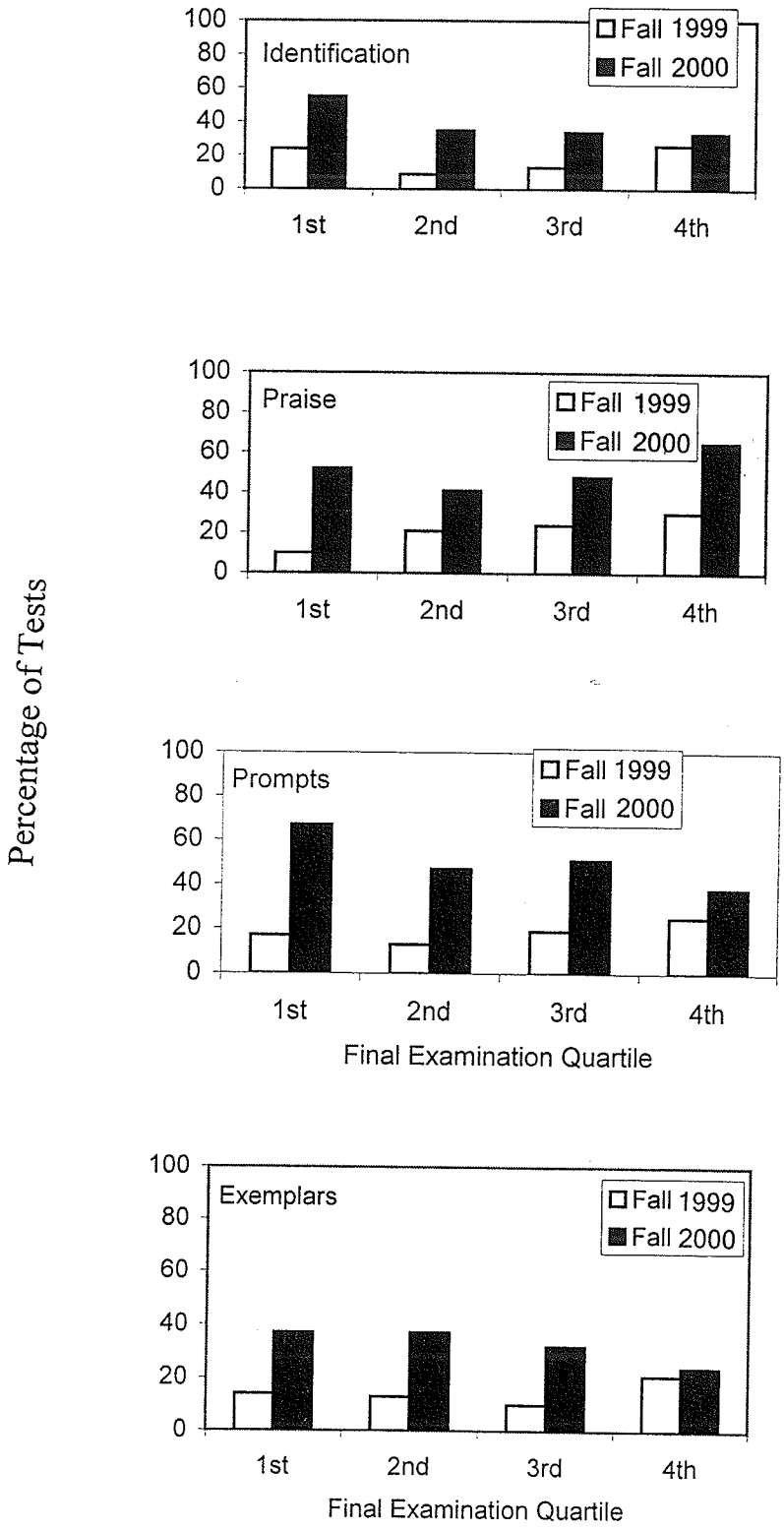

Figure 5. The percentage of students' unit tests, by course, in which students received specific feedback in the form of identification, praise, prompts, and exemplars.

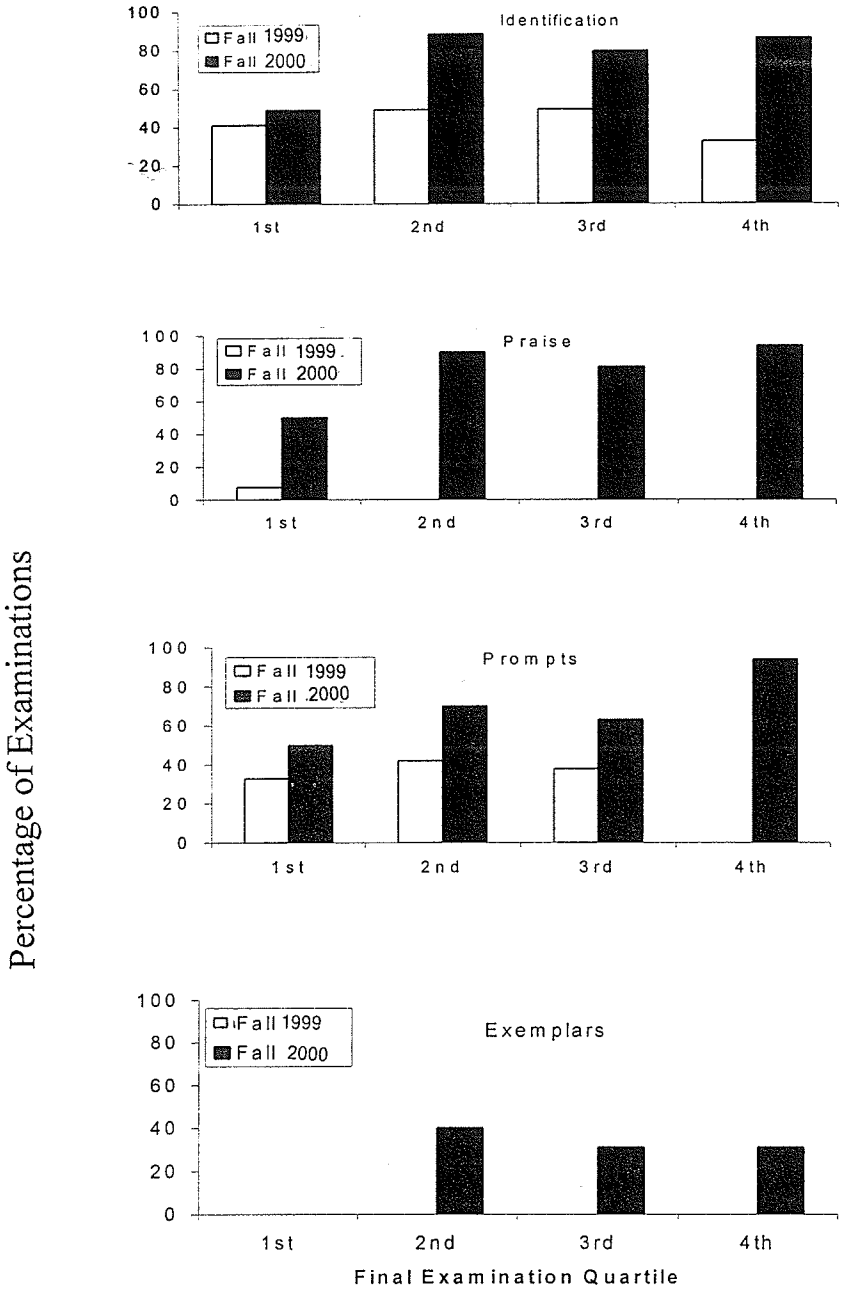

Figure 6. The percentage of students' exams, by course, in which students received specific feedback in the form of identification, praise, prompts, and exemplars.

Table 5 shows that on unit tests, the amount of all forms of feedback differ significantly between the courses. First, students in Fall 2000 were more likely to have received General Praise than students in Fall 1999. Second, students in Fall 2000 received a higher proportion of feedback in the form of identification, specific praise, prompts, and exemplars. Hence, a higher proportion of students' tests received substantive feedback (as shown in Figures 5 and 6) in the Fall 2000 course than did students in the Fall 1999 course.

The feedback that students received for the categories also differed on midterm examinations, with the exception of General Praise. Table 5 presents the specific feedback provided on midterm examinations. Note that the students in Fall 2000 received more substantive feedback related to higher order thinking (i.e., identification, praise, prompts, and exemplars) on their midterm examinations when compared with students in Fall 1999.

Table 6 shows the correlations between the types of substantive feedback that students received and the proportion of answers at each of the higher levels on the final examinations. Note that there are statistically reliable positive correlations between the presence of praise, prompts, and exemplars provided on both tests and examinations and Level 3 answers on the final examination. Further, the presence of praise and exemplars on unit test feedback is positively and reliably correlated with Level 4 answers on the final examination. Hence, these forms of feedback may play a role in reinforcing or prompting behavior involved in higher-order thinking.

Table 5

Mean Percentage and Effect Size for Unit Tests and Exams, by Course, Receiving Identification, Praise, Prompts, Exemplars, and General Praise Feedback

Course	Identification	Praise	Prompts	Exemplars	General Praise
			Unit Tests		
Fall 1999	17.00	20.00	17.00	13.00	68.00
Fall 2000	37.00**	53.00**	48.00**	31.00**	86.00*
(η^2)	(.28)	(.46)	(.46)	(.39)	(.14)
			Midterm Exams		
Fall 1999	43.00	3.00	38.00	0.00	90.00
Fall 2000	81.00**	83.00**	73.00**	31.00**	79.00
(η^2)	(.29)	(.79)	(.33)	(.39)	(.03)

* Significant at $p < .05$
** Significant at $p < .01$

Table 6

Correlations Between Substantive Feedback on Unit Tests and Midterm Examinations, and Higher Level Answers on Final Examination

	Unit Tests				Midterm Examinations			
Feedback Type	Final Exam Answer Level				Final Exam Answer Level			
	3	4	5	6	3	4	5	6
Identification	.30	.20	.02	.08	.29	.05	-.05	.28
Praise	.53*	.34*	.14	-.06	.62*	.27	.05	.16
Prompts	.31*	.21	.01	-.05	.43*	.30	-06	-.04
Exemplars	.32*	.33*	.05	.04	.39*	.08	.10	-.10

* Correlation is significant at $p < .05$ level (2-tailed)

Finally, while not the main focus of this research an additional analysis was carried out to determine the effect of having at least one supervised midterm. As indicated in Table 7, the general pattern in CAPSI courses is for mean Midterm 2 scores to be lower than, or equal to, average Midterm 1 scores. Note, however, that with the introduction of a supervised midterm that the average score is generally lower on the supervised midterm, regardless of whether it is Midterm 1 or Midterm 2.

Table 7

Mean Percentage Scores[1] *on Midterms and Final Exams, by Course and Supervision Situation*

Course	Midterm 1	Midterm 2	Final
Behavior Modification Principles			
Fall 1999 (n = 20)	93.50	85.50	69.12[s]
Fall 2000 (n = 24)	80.13[s]	82.08	61.70[s]
Behavior Modification Applications			
Fall 1999 (n = 8)	95.00	93.33	89.58[s]
Fall 2000 (n = 8)	92.27	72.33[s]	76.25[s]
Foundations of Learning			
Fall 1999 (n = 9)	84.33	84.93	64.52[s]
Fall 2000 (n = 18)	60.73[s]	71.47	54.57[s]
Systems of Psychology			
Fall 1999 (n = 5)	98.67	98.67	72.00[s]
Fall 2000 (n = 5)	82.00	74.67[s]	59.67[s]

[s] Indicates a supervised examination

[1] The scores in this table are taken directly from archived exams as graded by two different instructors (one for each session).

Discussion

As an educational intervention package, the results of this research suggest that the combination of the various types of feedback and bonus points was effective in terms of having students answer at higher levels of the modified taxonomy. First, the students in Fall 2000 performed at substantially higher levels compared to students in Fall 1999 both in terms of scores and levels of answers on the final examinations. The finding that there was no significant difference on Unit 1 or Unit 2 tests (Table 3), or year in course, suggests that the two groups of students were essentially equal; hence, it appears that there were no initial differences between the students enrolled in the courses that would account for the significant differences found in this study.

In addition, students' performance on Midterm 1 in Fall 2000 was at, or even above, the students in Fall 1999 (Table 3). One could argue that the scores on the first midterm (which was supervised for the Fall 2000 students) reflect that these students learned more than the Fall 1999 students, since there was no possibility for the students to make use of class materials during that midterm. Hence, it appears that the instructions regarding the presentation of bonus points may already have had a positive effect on the students in Fall 2000 course by the time of the first midterm.

One potential confound would be whether having a supervised midterm would have led to students answering at the higher levels of the modified taxonomy. While students may study earlier and more often for a supervised examination than for a non-supervised midterm, and therefore be better prepared to answer questions asked at higher levels, there is no reason to suppose that this would make them more likely to answer at levels higher than required. If students were likely to answer at higher levels than

required on a supervised examination, then evidence of this should have been observed in the Fall 1999 answers on the final examination. However, given that a smaller percentage of these students provided answers at levels higher than required (compared to the students in Fall 2000), there is no reason to assume that this would be the case.

There were higher proportions of the higher-order questions on Midterm 2 for the Fall 1999 students than for the Fall 2000 students (Figure 1); hence, if the students in both courses had learned the material equally well, one might have expected that significantly more students in Fall 1999 would have answered at the levels specified in the questions. However, there was no significant difference between the courses on the first midterm (Figure 2 and Table 3); further, students in the Fall 2000 course answered at the higher levels despite relatively few questions at Level 3, and no questions above that level (i.e., none at Levels 4, 5, or 6). Hence, students (on average) in Fall 2000 outperformed (in terms of higher level answers) students in Fall 1999 on both midterms and the final examination.

A higher proportion of students in Fall 1999 provided answers that were incomplete or incorrect (Level 0) on both midterms compared to students in Fall 2000. In contrast, a higher proportion of students in Fall 2000 provided at least one answer at Level 0 on the final examination than students in Fall 1999. Subsequent investigation might focus on the circumstances under which these Level 0 answers occur. For instance, one might speculate that the students in Fall 2000 were more likely than the students in Fall 1999 to be attempting higher level answers incorrectly, whereas the students in Fall 1999 were more likely to answer incorrectly at the level specified. Some evidence supports this suggestion. First, more of the Fall 1999 students' answers than of the Fall

2000 students' answers were assessed at $D = -2$, which would represent a lower-level answer (i.e., Level 1 or 2) answered incorrectly. Second, fewer of the Fall 2000 students' answers were assessed at $D = -4$ and $D = -5$, which represent incorrect answers to Level 4 and Level 5 questions. Finally, since the students in Fall 2000 were more likely than the students in Fall 1999 to answer higher-level question correctly ($D \geq 0$), the finding that students in Fall 2000 answered more questions at or above the level asked suggests that one reason for Level 0 answers might lie in lower-level questions that were answered incorrectly.

The fact that students in Fall 2000 received a higher proportion of feedback on their unit tests and exams does not alone suggest that the feedback was related to the performance in that course. The correlations, however, between praise, prompts, and exemplars with Level 3 and 4 answers suggest two things: (a) praise may have acted as an indirect delayed reinforcer, through the use rule-governed behavior (i.e., self-statements such as "I answered at a higher level, which will earn a bonus mark on an exam"), for an increase or maintenance of answers provided above the level asked, and (b) prompts and exemplars may have acted as discriminative stimuli (i.e., cues) for forms of answers or forms of responses to imitate, respectively, on future tests. However, with correlational analyses, it is difficult to know the direction of influence. While praise, prompts, and exemplars may have functioned as described above it is presumably the use of self-statements in the form of rules that increased the likelihood of students providing higher level answers, and the indirect reinforcement when praise or identification of a higher level answer occurred in the feedback from proctors, the teaching assistant, or the instructor, that helped maintain some of this behavior.

Students in Fall 2000 were more likely to receive the various forms of substantive feedback than were students in Fall 1999. Since students in Fall 2000 more often received these forms of feedback on midterms, it is possible that they were imitating the feedback that they received from the instructor. Another possibility is that the presence of specific statements related to higher-level answers paired with bonus points in the feedback to students in Fall 2000 served to strengthen the reinforcing qualities of the feedback. Perhaps through pairing feedback with bonus points the praise, prompts, and exemplars exerted more control over the targeted behavior of students than when they are provided without this pairing.

The potent variable, then, that appears to account for the differences between the two groups was the use of instructions related to, and presentation of, bonus points on midterms and final examinations in the Fall 2000 course. Certainly the results and statistical analyses of the types of feedback suggest significant findings; however, the fact remains that the bonus points likely acted as delayed reinforcers for answering at higher levels, and may also have acted as verbal prompts for students to answer at higher levels. In fact, since the bonus points on the midterm examinations were often paired with identification, praise, and sometimes prompts to take the answer even higher, these types of feedback became conditioned reinforcers through the pairing with the bonus points that may have acted as delayed conditioned, or even generalized, reinforcers. It is presumably through this process that students' answering at higher levels was maintained and increased in the Fall 2000 course as compared to the Fall 1999 course.

Clearly, the differences between Fall 1999 and Fall 2000 in terms of students answers to specific study questions, relative to the minimum level at which the question

can be asked, and scores on tests and exams, demonstrate that the use of substantive feedback to students on unit test and midterm examination performance, paired with bonus points for higher-level answers above the specified level is an effective means to encourage students to engage in behavior that is reliably assessed as higher-level thinking.

Limitations

As with any research, this study has limitations which future research can overcome. One limitation of this study is that students' scores are not statistically independent of each other. The students interacted with each other as within-course proctors, and with the instructor. From an inferential statistical standpoint, the use of parametric or non-parametric tests would not be appropriate because the nature of the design violates the assumption of independence of observations. However, the use of statistical procedures, interpreted with caution, provides information about effect sizes and significance values that descriptive statistics do not. In addition, these tests make the results more understandable to others who are more likely to use inferential statistics.

Another limitation is that the use of a non-randomized quasi-experimental design limits the statements that one can make about the strength of the findings. Specifically, future research should use stratified samples of students in the same year who are randomly assigned to two different course sections. One section can act as a control group, and the other as an experimental group. Such a design would improve the strength of the statements that can be made about the outcome of the statistical analyses.

Implications

Increasing higher-order thinking. This research demonstrates that the students in Fall 2000 answered more of the higher-order questions correctly, and provided more answers at above the level asked on the final (unsupervised) examination. This suggests that the course procedures combined with the intervention package were successful in increasing the proportion of higher level answers. As suggested by Reboy and Semb (1991), the use of a PSI-taught course in obtaining higher level thinking is dependent upon the content and pedagogical delivery of the course, rather than the technology used to deliver the course. As such, the use of CAPSI in this research to deliver an effective course that not only resulted in higher average scores than a previously-taught section (which in itself is impressive because it represents a gain in learning) but also resulted in a demonstration that all students could answer at the levels of Application and Analysis. Hence, critics who argue that PSI-taught courses may promote ''book-bound'' learning (see Caldwell, 1985; Hobbs, 1987) are not correct in their assessment; however, the use of the technology must be combined with quality materials and instruction.

Aside from a higher density of substantive feedback (Pear & Crone-Todd, 2001), the students in Fall 2000 did not receive any higher proportion of interaction with the instructor than the students in Fall 1999. This suggests that if an instructor (and the teaching assistant) provides instructions and a high proportion of substantive feedback, the within-course proctors tend to maintain this level of substantive feedback. Because all of the various forms of feedback were related to some higher level answers, instructors and teaching assistants who wish to increase the amount of substantive feedback in their courses would be wise to do so when providing their own feedback to students.

For instructors who wish to increase the answer levels of their students, there are several options. First, substantive feedback combined with bonus points (as in this study) is recommended. Second, since all of the students in Fall 2000 were able to provide answers at Levels 3 and 4, perhaps changing the questions in the courses to provide a higher proportion of these questions would result in more students answering at higher levels. By shifting the criterion for answering at higher levels on more of the questions, instructors provide more opportunities for students in the course to gain more experience in answering these questions, and to be provided with more feedback on their answers at these levels. The maximally effective proportion of higher- to lower-levels questions is an empirical question that deserves further research.

The use of the modified taxonomy. As Williams (1999) had suggested, the reliability and validity of operational definitions (in terms of behavior) need to be established for cognitive constructs. Earlier work on applying behavioral definitions up to Level 3 (Application) of Bloom's taxonomy (e.g., Johnson & Chase, 1981; Semb & Spencer, 1976) yielded impressive reliability, or agreement, measures. The use of the modified Bloom's taxonomy (Crone-Todd et al, 2000; Pear et al, 2001) has been demonstrated to be reliable and valid up to, and including, Level 4 (Analysis). As Crone-Todd et al. found, it is hard to establish the reliability of the taxonomy for Levels 5 and 6 without more exemplars for training and assessment. It is an empirical question whether the presentation of more exemplars of Levels 5 and 6 will increase the agreement on the assessment of these levels. While the previous research on this taxonomy required two groups of raters to obtain high agreement, the present research indicates that when the objectives for answers are well-defined, the assessment of answer levels using the

taxonomy yields high, above-chance, inter - scorer agreement. Further validation of the other taxonomic levels can be determined only by having others use the taxonomy in their research, and reporting on the agreement arrived at in their research.

Future Research

The first two unit tests served as an important marker for initial performance in the course. Future research could be carried out to determine whether a better measure might involve a supervised test of students' skills in writing and answering at higher levels. In an effort to compare the experimental course with a previous course, presenting another form of a pre-test would have presented an uncontrolled variable that may have had a different effect on students in one course versus the other. Hence, for the present study it would not have been appropriate to do so. It could be interesting to look at whether a more global measure of higher order thinking, such as the Watson-Glaser critical thinking test, would be correlated with students' higher-order answer levels in the courses under study in the CAPSI laboratory. Alternatively, a course-specific test could be prepared, which could present various levels of questions to students at the outset.

The goal for future research, then, will be to determine which variables (or combinations of variables) are most responsible for higher performance in terms of higher-level answers and scores on final examinations in CAPSI-taught courses. That is, would an increase in substantive feedback as indicated in this study, without the use of bonus points, produce such an effect? Would the presentation of bonus points, with proportions of substantive feedback similar to that in 2000, produce such an effect? Or, is it the combination of the bonus points and substantive feedback that produce the effect? Only by teasing apart these variables in future research can these questions be answered.

Future research should also address the question of maximally effective proportions of higher- versus lower-level questions in PSI- or CAPSI-taught courses. It is possible that study objectives in the form of questions underestimate the level at which students can answer the questions. Alternatively, perhaps incorporating more higher-level questions would make a CAPSI course too difficult, leading many students to drop the course. Another question to address is whether markers (instructors, teaching assistants, and proctors) provide accurate feedback on higher-level answers. While it may be true that having more higher-level questions in a course could provide more practice on these questions, if the feedback received on the tests is inaccurate, how will that affect the students' subsequent answers? Some impressive work has already been carried out on assessing the accuracy of proctor feedback (e.g., Martin, 2000; Martin et al., 2001a,b), indicating that students are more likely to provide inaccurate feedback on incorrect answers. Since students often demonstrate difficulty in answering higher-level answers, it seems likely that students will receive inaccurate information on higher-level answers from other students. Presumably inaccurate feedback would result in fewer correct answers to these questions; however, this is also an empirical question that can be addressed in future research.

References

Baldwin, J. D., & Baldwin, J. J. (1986). *Behavior principles in everyday life*, (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Barnes, D., Healy, O., & Hayes, S. C. (2000). Relational frame theory and the relational evaluation procedure: Approaching human language as derived relational responding. In J. C. Blackman (Eds.). *Experimental and Applied Analysis of Human Behavior* (pp. 149-150). Reno, NV: Context Press.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw Hill.

Born, D.G., Gledhill, S.M., & Davis, M.L. (1972). Examination performance in lecture-discussion and personalized instruction courses. *Journal of Applied Behavior Analysis, 5*, 33-43.

Bostow, D.E., & O'Connor, R.J. (1973). A comparison of two college classroom testing procedures: Required remediation versus no remediation. *Journal of Applied Behavior Analysis, 6*, 599-607.

Brooke, R. R., & Ruthven, A. J. (1984). The effects of contingency contracting on student performance in a PSI class. *Teaching of Psychology, 11*, 87-89.

Buerkel-Rothfuss, N. L., Gray, P. L., & Yerby, J. (1993). The structured model of competency-based instruction. *Communication Education, 42*, 22-36.

Buskist, W., Cush, D., & DeGrandpre, R. J. (1991). The life and times of PSI. *Journal of Behavioral Education, 42*, 215-234.

Calder, J. R. (1983). In the cells of the ''Bloom Taxonomy''. *Journal of Curriculum Studies, 15*, 291-302.

Caldwell, E. C. (1985). Dangers of PSI. *Teaching of Psychology, 12*, 9-12.

Caldwell, E. C., Bissonnettee, K., Klishis, M. J., Ripley, M., Farudi, P. P., Hochstetter, G. T., & Radiker, J. E. (1978). Mastery: The essential essential in PSI. *Teaching of Psychology, 5*, 59-65.

Carnine, D. (1991). Curricular interventions for teaching higher order thinking to all students: Introduction to the special series. *Journal of Learning Disabilities, 24*, 261-269.

Conrad, C. J., & Semb, G. (1977). Proctor selection, training, and quality control: A longitudinal case study. *Journal of Personalized Instruction, 2*, 238-240.

Crist, L. R. (1982). Preparing students for tests through the use of questions and answers. *Teaching of Psychology, 9*, 109-111.

Croft, R. G., Johnson, W. G., Berger, J., & Zlotlow, S. F. (1976). The influence of monitoring on PSI performance. *Journal of Personalized Instruction, 1*, 28-31.

Crone-Todd, D. E., Holborn, S. W., & Pear, J. J. (2001). *Increasing rates of class attendance and newsgroup participation in a computer-aided PSI course*. Manuscript submitted for publication. University of Manitoba.

Crone-Todd, D. E., & Pear, J. J. (1998). *The relationship between student pacing, feedback, and exam performance using a computer-aided PSI program*. Paper presented at the 3rd Annual EvNet Conference, Concordia University, Montreal, PQ, in February, 1998.

Crone-Todd, D. E., & Pear, J. J. (1999). *Computer-aided personalized system of instruction in higher education: Increasing in-course proctors' feedback.* Unpublished manuscript, University of Manitoba.

Crone-Todd, D.E., & Pear, J.J. (2001). Application of Bloom's taxonomy to PSI. *Behavior Analyst Today, 3 (2)*, 204-210.

Crone-Todd, D. E., Pear, J. J., & Read, C. N. (2000). Operational definitions of higher-order thinking objectives at the post-secondary level. *Academic Exchange Quarterly, 4 (3)*, 99-106.

Crosbie, J., & Kelly, G. (1993). A computer-based personalized system of instruction course in applied behavior analysis. *Behavior Research Methods, Instruments, and Computers, 25*, 366-370.

Facione, P.A., (1997) Critical thinking: What it is and why it counts. Located on the California Academic Press Web site at URL: *http://www.calpress.com/critical.html*

Gaynor, J. F., & Wolking, W. D. (1974). The effectiveness of currently enrolled student proctors in an undergraduate special education course. *Journal of Applied Behavior Analysis, 7,* 263-269.

Gierl, M. J. (1997). Comparing cognitive representations of test developers and students on a mathematics test with Bloom's taxonomy. *The Journal of Eduational Research, 91*, 26-32.

Glick, D. M., Moore, M. C., Roberts, M. S., & Born, D. G. (1982). A single subject analyses of the effectiveness of telephone prompts for students in a flexibly paced course. Reprinted from Journal of Personalized Instruction, 4 (1), 1980. In J. G. Sherman, R. S. Ruskin, & G. B. Semb (eds.), *The personalized system of instruction: 48 seminal papers,* 253-260.

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains. *American Psychologist, 53*, 449-455.

Hayes, S.C. (1994). Relational frame theory: A functional approach to verbal events. In S.C. Hayes, L.J. Hayes, M. Sato, & K. Ono (Eds.), *Behavior analysis of language and cognition: The fourth international institute on verbal relations* (pp. 11-30). Reno, NV: Context Press.

Henneberry, J. K. (1976). Initial progress rates as related to progress in a personalized system of instruction. *Teaching of Psychology, 3*, 178-181.

Hobbs, S. H. (1987). PSI: Use, misuse and abuse. *Teaching of Psychology, 14*, 106-107.

Hohn, R. L., Gallagher, T., & Byrne, M. (1990). Instructor-supplied notes and higher-order thinking. *Journal of Instructional Psychology, 17*, 71-74.

Johnson, K. R. (1977). Proctor training for natural control. *Journal of Personalized Instruction, 2*, 230-237.

Johnson, K. R., & Chase, P. N. (1981). Behavior analysis in instructional design: a functional typology of verbal tasks. *The Behavior Analyst, 4(2)*, 103-121.

Johnson, K.R., & Ruskin, R.S. (1977). *Behavioral instruction: An evaluative review.* Washington, DC: American Psychological Association.

Johnson, K.R., & Sulzer-Azaroff, B. (1975). PSI for first-time users: Pleasures and pitfalls. *Educational Technology, 15*, 9-16.

Johnson, K.R., Sulzer-Azaroff, B., & Maass, C.A. (1976). The effects of internal proctoring on examination performance in a personalized instruction course. *Journal of Personalized Instruction, 1*, 113-117.

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford University Press.

Keller, F. S. (1966). A personal course in psychology. In R. Ulrich, T. Stachnik, & J. Mabry (Eds.), *Control of human behavior: Expanding the behavioral laboratory (v. 1)*. Glenview, IL: Scott, Foresman and Company.

Keller, F. S. (1968). ''Good-bye, teacher''. *Journal of Applied Behavior Analysis, 1*, 79-89.

Keller, F. S., & Sherman, J. G. (1974). *The PSI handbook: Essays on a personalized instruction*. Don Mills, ON: W. A. Benjamin, Inc.

Keller, F.S., & Sherman, J.G. (1982). *The PSI handbook: Essays on personalized instruction*. Lawrence, Kansas: TRI Publications.

Kinsner, W., & Pear, J. J. (1990). A dynamic educational system for the virtual campus. In U. E. Gattiker, L. Larwood, & R. S. Stollenmaier's (Eds.) *End-User Training*, 201-238. New York: W. de Gruyter.

Kottke, J. L., & Schuster, D. H. (1990). Developing tests for measuring Bloom's learning outcomes. *Psychological Reports, 66*, 27-32.

Kulik, C. C., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research, 60*, 265-299.

Kulik, J. A., Jaska, P., & Kulik, C. C. (1978). Research on component features of Keller's personalized system of instruction. *Journal of Personalized Instruction, 3*, 2-14.

Kulik, J. A., Kulik, C. C., & Cohen, P. A. (1979). A meta-analysis of outcome studies of Keller's personalized system of instruction. *American Psychologist, 34*, 307-318.

Lamal, P. (1984). Interest in PSI across sixteen years. *Teaching of Psychology, 11*, 237-238.

Landis, R.J., & Koch. G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Lloyd, K. E., & Knutzen, N. J. (1969). A self-paced programmed undergraduate course in the experimental analysis of behavior. *Journal of Applied Behavior Analysis, 2*, 125-133.

Lloyd, K.E., & Lloyd, M.E. (1986). Has lightning struck twice? Use of PSI in college classrooms. *Teaching of Psychology, 13*, 149-151.

Lloyd, K. E., & Lloyd, M. E. (1992). Behavior analysis and technology in higher education. In R. P. West & L. A. Hamerlynck's (Eds.), *Designs for excellence in education: The legacy of B. F. Skinner*, 147-160. Longmont, CO: Sopris West, Inc.

Martin, T. L. (2000). *Analysis of proctor feedback accuracy in a computer-aided personalized system of instruction.* Unpublished master's thesis, University of Manitoba, Winnipeg, Manitoba, Canada.

Martin, G., & Pear, J. (1999). *Behavior modification: What it is and how to do it.* (6th Edition). Upper Saddle River, NJ: Prentice-Hall.

Martin, T.L., Pear, J.J., & Martin, G. L. (in press a) Feedback and its effectiveness in a computer-aided personalized system of instruction. *Journal of Applied Behavior Analysis*.

Martin, T.L., Pear, J.J., & Martin, G. L. (in press b). Analysis of proctor grading accuracy in a computer-aided personalized system of instruction. *Journal of Applied Behavior Analysis*.

Mayer, R., & Goodchild, F. (1990). *The critical thinker*. Santa Barbara, CA: Wm. C. Brown Publishers.

McMichael, J. S., & Corey, J. R. (1969). Contingency management in an introductory psychology course produces better learning. *Journal of Applied Behavior Analysis, 2*, 79-83.

Newman, F. M. (1991a). Promoting higher order thinking in social studies: Overview of a study of 16 high school departments. *Theory and Research in Social Education, 19*, 324-340.

Newman, F. M. (1991b). Classroom thoughtfulness and students' higher order thinking: Common indicators and diverse social studies courses. *Theory and Research in Social Education, 19*, 410-433.

Paul, R. W., & Heaslip, P. (1995). Critical thinking and intuitive nursing practice. *Journal of Advanced Nursing, 22*, 40-47.

Pear, J. J., & Crone-Todd, D. E. (1999). Personalized system of instruction in cyberspace. *Journal of Applied Behavior Analysis, 32*, 205-209.

Pear, J.J., Crone-Todd, D.E., Wirth, K., & Simister, H. (2001). Assessment of thinking levels in students' answers. *Academic Exchange Quarterly*.

Pear, J.J., & Kinsner, W. (1988). Computer-aided personalized system of instruction: An effective and economical method for short- and long-distance education. *Machine-Mediated Learning, 2*, 213-237.

Pear, J.J., & Novak, M. (1996). Computer-aided personalized system of instruction: A program evaluation. *Teaching in Psychology, 23*, 119-123.

Peterson, N. (1982). Feedback is not a new principle of behavior. *Behavior Analyst, 5*, 101-102.

Reboy, L. M., & Semb, G. B. (1991). PSI and critical thinking: Compatibility or irreconcilable differences? *Teaching of Psychology, 18*, 215-218.

Ribes, E., & Rodriguez, M.E. (2001). Correspondence between instructions, performance, and self-descriptions in a conditional discrimination task: The effects of feedback and type of matching response. *Psychological Record, 51*, 309-333.

Roberts, N. (1976). Further verification of Bloom's taxonomy. *Journal of Experimental Education, 45*, 16-19.

Robin, A. L. and Cook, D. A. (1978). Training proctors for personalized instruction. *Teaching of Psychology, 5*, 9-13.

Robin, A.L., & Heselton, P. (1977). Proctor training: The effects of manual versus direct training. *Journal of Personalized Instruction, 2*, 19-24.

Seddon, G. (1978). The properties of Bloom's taxonomy of educational objectives for the cognitive domain. *Review of Educational Research, 48*, 303-323.

Seddon, G. M., Chokotho, N. C., & Merritt, R. (1981). The identification of radex properties in objective test items. *Journal of Educational Measurement, 18*, 155-170.

Semb, G. B. (1995). The personalized system of instruction (PSI): A quarter century report. *Revista Mexiscana de Psicologia, 12*, 161-175.

Semb, G., Glick, D. M., & Spencer, R. E. (1979). Student withdrawals and delayed work patterns in self-paced psychology courses. *Teaching of Psychology, 6*, 23-25.

Semb, G., & Spencer, R. (1976). Beyond the level of recall: An analysis of complex educational tasks in college and university instruction. In L.E. Fraley & E. A. Varas (Eds.), *Behavior Research and Technology in College and University Instruction* (pp. 115-126). Gainesville, FL: Department of Psychology, University of Florida.

Sheppard, W.C., & MacDermot, H.G. (1970). Design and evaluation of a programmed course in introductory psychology. *Journal of Applied Behavior Analysis, 3*, 5-11.

Sidman, M. (1994). *Equivalence relations and behavior: A research story*. Boston: Authors Cooperative.

Skinner, B.F. (1957). *Verbal behavior*. Acton, MA: Copley Publishing Group.

Skinner, B.F. (1968). *The technology of teaching*. New York: Appleton-Century-Crofts.

Skinner, B.F. (1969). *Contingencies of reinforcement: A theoretical analysis*. New York: Appleton-Century-Crofts.

Skinner, B.F. (1974). *About behaviorism*. New York: Knopf.

Skinner, M. E. (1990). The effects of computer-based instruction on the achievement of college students as a function of achievement status and mode of presentation. *Computers in Human Behavior, 6*, 351-360.

Sulzer-Azaroff B., Johnson K.R., Dean M.R., and Freyman D.R (1977). An experimental analysis of proctor quiz-scoring accuracy in personalized instruction courses. *Journal of Personalized Instruction 23*, 143-149.

Watson, J.B., & Rayner, R. (1920). Conditioned emotional reactions. *Journal of Experimental Psychology, 3*, 1-14

Wesp, R., & Ford, J. E. (1982). Flexible instructor pacing assists student progress in a personalized system of instruction. *Teaching of Psychology, 9*, 160-162.

Williams, R.L. (1999). Operational definitions and assessment of higher-order cognitive constructs. *Educational Psychology Review, 11*, 411-427

Winokur, S. (1976). *A primer of verbal behavior*: An operant view. Prentice-Hall

Appendix A - Debriefing Letter to Students in B. Mod I, B. Mod II, and Systems

Feedback to Students in CAPSI-Taught Courses During the 2000 Regular Session

As mentioned in the General Manual for your course, all data in CAPSI-taught courses are subject to later analyses. Your participation in this course will help to advance knowledge of the educational process through research on these courses, and hopefully has been of benefit to yourself.

During this year, the specific variables under consideration include the use of bonus points and supervised midterms. The effect of these variables on students' unit test and examination performance will be analyzed over the next few months. The following questions will be addressed: (1) Do students answer below, at, or above the levels required by questions on midterms when bonus points are available for going above the required level? (2)What level of answers do students provide when exams are supervised versus unsupervised? The expectation is that through the use of bonus points the answer levels should increase above what is required. In addition, supervision of exams may also change the level of answer; however, it remains to be determined through research precisely how this change may occur.

Another variable in this study included either providing answer levels to students, or having them identify the levels in their answers. In both cases, we expect that answers on tests and midterms should be at higher levels than in previous years.

All of your data will be kept confidential, and no identifying information will be provided to third parties. The researchers (Ms. Darlene Crone-Todd and Dr. Joseph Pear) thank you heartily for your participation in the course. If you have any questions or concerns about the study, or would like further information on any of the course material, please contact Ms. Darlene Crone-Todd at 474-8258 or Dr. Joseph Pear at 474-8777.

The outcome of this study will be advertised outside of your instructor's office (P435F Duff Roblin) within approximately six months from the end of your class. If you would like to receive the results via email or postal mail, please fill out the following and return it to your instructor.

Please send me information about the results of the study on CAPSI-taught courses from the 2001 Regular Session.

Name:_____

Email:_____ or: Postal Address: _____

_____

_____

Appendix B - Summary of Question Levels for Students in 17247 Learning Foundations of Psychology

Here are the question level summaries (please note that this material is copyrighted, and you may not reproduce, distribute, or quote without permission

Levels 1 and 2

The questions at these levels will always be directly from the course textbook, including examples, analyses, discussion, and arguments, et cetera.

Level 1: Knowledge

At this level, answers may be memorized or closely paraphrased from the textbook.

Level 2: Comprehension

At this level, answers must be in your own words, while still using terminology appropriate to course material.

Levels 3 Through 6

The answers for these levels go beyond the text material in that they must be inferred or extrapolated from the information in the text.

Level 3: Application

At this level, answers require you to recognize or apply a concept or principle you have learned at Level 2 in a new situation, or to solve a new problem. Questions at this level may present or require examples that are not found in the textbook.

Level 4: Analysis

At this level, questions require the breaking down of concepts into their constituent parts. Questions at this level may also require identification or explanation of the essential components of concepts, principles, or processes. In addition, this level may require you to compare and contrast (i.e., state the similarities and differences), or explain how an example illustrates a given concept, principle, et cetera.

Level 5: Synthesis

Synthesis is putting together parts to form a whole (i.e., the opposite of Level 4). Questions at this level may require you to generate definitions that are not identified in the textbook, or to explain how to combine principles or concepts to produce something new.

Level 6: Evaluation

An evaluation question requires you to present and evaluate reasons for and against a particular position, and (ideally) to come to a conclusion regarding the validity of that position. In some cases you may be able to support one particular position over another; in other cases, your conclusion might be that several positions are correct or that the evidence does not permit you to take a particular position. In this type of question the most important part of the answer is the justification or rationale for your conclusion. Note that at this level, there is no correct answer per se; rather, the answer is evaluated in terms of how well you argue your position, given the facts at your disposal. A good discussion at this level involves the use of all of the preceding levels.

Appendix C - Score Sheet for Student Answer Assessment

Subject Number:_____________ Course:________________ Year:___________

Unit	Test (T) or Exam (E)	Question	A: Level Assessed	B: Level Answered	Difference (B - A)

Appendix D - Instructions for Assessing Answers as Correct, Mostly Correct, or Incorrect

In differentiating between whether an answer is 'Correct', or 'Mostly Correct', we will use the following guidelines:

CORRECT: entire answer is correct - there is no doubt, based on the text material, that the student's answer is correct. The answer is fully correct (no ambiguity) according to the answer key provided.

MOSTLY CORRECT: Overall the answer is right - but if any one part of the answer is questionable in terms of being fully correct (yet not incorrect), then it would be marked as mostly correct. For example, even if the question had five points but the student gets one of the points 'mostly correct' then the entire answer would be marked as mostly correct. Also, if all five of the points were basically/mostly correct, then the entire answer would be marked "mostly correct".

Another example of "mostly correct" would be if the definition of a particular concept is weak or poorly worded, but an example given is bang-on. In this case, the student has demonstrated through the use of an example that they "understand" the concept. However, it is not fully "correct".

Note that in the above examples, errors can be ones of omission or commission. The "line" which determines "mostly" versus "incorrect" is fine - it will depend how far the answer deviates from the course material. For example, if one gives a good definition and example, but then also uses mentalistic terms to explain the example, this would probably be incorrect (e.g., the person wanted to run more often because they really liked getting praise from others).

INCORRECT:
Answers which do not use the correct terminology, have not reasonably addressed the question, or use errors of omission/commision that detract too far from the correctness of the answer are some of the types that will fit in this category.

For example, using the mentalistic terminology above would not work for most of these courses (although in 017.252 it might be acceptable under some circumstances which require such answer about other areas of Psychology).

Also, if an answer is to a different question than what is asked, it is incorrect. Some examples would include merely giving definitions when asked to "compare and contrast", or giving personal anecdotes when asked for whether something is effective or not.

Again, omission/commission can occur in a variety of ways. It might be useful if when these answers are being assessed if we can keep some record of the types of omission/commision/whatever errors we come across. It is possible that similar problems are encountered by various students, and we could address these in future studies.

Appendix E - Answer Assessment Instructions

NOTE: These instructions are imperative. The flowchart is a basic summary of these instructions, so be sure to refer back to these instructions when assessing answers.

Assessing Answers: General Comments and Guidelines About the Decision Boxes

Terminology/Phrasing

First, you need to assess whether the appropriate (or reasonable) terms and phrasing are used in answering a particular question correctly. To assess this, you need to determine which terms are specific to answering the question. For example, when asked to talk about reinforcement, an answer may use "reward" instead of "reinforcement", and be reasonably (although not technically) correct. However, if they explain the answer by appealing to "motivation" or some other mentalistic term, then the appropriate phrasing would not have been used - i.e., since reinforcement is determined by looking at an antecedent, the behavior and consequence, and the likely outcome in similar situations.

Answer Complete and Correct

If parts of an answer seem reasonable, but not fully correct, then the answer is assessed at Level 0. However, if the student goes on a "tangent" (e.g., the question asks for a definition, and they go onto argue for one system versus another that is not addressed in the text), then they have gone "beyond" the question. In this example, the student answered at a Level 6. If, however, the "going beyond" components include errors of commission or omission, then the answer is at Level 0.

Is Answer Wholly in Text?

Basically, here you are looking to see if what the student has answered is in the text. If it is, it can either be in their own words, or not. Specifically, if the answer is in the text, then you need to assess whether it has been summarized or rephrased in such as way to end up at a Level 2. Alternatively, if it appears to be "memorized" word-for-word (or paraphrased) from the text, then it is assessed at Level 1.

Be careful to distinguish between merely explaining, rephrasing, or translating what is already in the text. If you can place the phrase "In other words" in front of a student's own words for answering a question, and find that the answer is basically saying the same thing (but in "other words"), then it is a Level 2. However, if the student is providing a new example, comparing and contrasting concepts, principles, or processes, generating something new, or making an argument, then it is at a higher level.

Own Words

Is the answer in the student's own words, or from the text? We have currently operationally define this in the following way:

Basically, to put an answer in one's own words, there needs to be some kind of grammatical manipulation made to the structure of a sentence or paragraph. Hence, ''comprehension'' here involves at the minimum an ''understanding of grammar'' with respect to the material in the text. Students demonstrate that they have comprehended given material by rephrasing it in some way. The following rules to follow are:

- Merely rephrasing (i.e., *switching*) an answer that includes a restatement of the question does not count
    - E.g., when asked ''What are the various ways in which one could drive to Vancouver?'', the student may answer ''The various ways in which one could drive to Vancouver include...'' Here, the text may never state ''there are several ways to drive...'', and merely repeating/paraphrasing the question in the answer would not count as demonstrating Comprehension.
    - E.g., to answer ''What is a cat?'', the student's answer bears the same surface features as the answer in the text. That is, the answer in the text is given as ''A cat is a mammal with four legs, paws, claws, fangs, pointy ears, emits a 'meow' sound, and usually weighs under 20 pounds''. Simply *switching* the order of the elements would not count as a Level 2 answer: ''A mammal with four legs, paws, claws, fangs, pointy ears, that emits a 'meow' sound, and usually weighs under 20 pounds is a cat.''
- If one *switch* occurs between at least two parts of a sentence that are not restatements of the question, then it would count as Level 2.
    - E.g., From Martin & Pear (1999), a positive reinforcer is defined as: ''...an event that, when presented immediately following a behavior, causes the behavior to increase in frequency (or likelihood of occurrence)''. The student could provide a definition that states, ''If a behavior is immediately followed by an event that results in that behavior increasing in frequency, then the event is called a reinforcer''.
- If ''cutting'' (i.e., taking out unnecessary text in brackets or intervening sentences) or ''pasting'' occurs (i.e., sentences are taken from various parts of the text to answer a question, which do not include intervening text), then there must be at least one *switch* that conforms to the immediately preceding rule. That is, the switch must occur in at least one of the sentences, and cannot be merely a restatement of the question.

Position Defended

Generally speaking, this is just like the question assessment. The student has made a logical argument, based upon sound reasoning, that is at least partly based upon material presented in the text, but where the argument itself is not contained within the text. Answers of this type are assessed at Level 6.

Definitions, Processes, or Concepts Combined in a New Way

Here the answer indicates that the student has combined things in a way that is not present in the textbook. For example, a textbook may show a number of different Lego pieces, and how each one operates. The text may also show how one constructs, say, a house using Lego. Next, the student puts together all of these pieces to create a tractor (which is not in the text), and shows how each piece contributes to the overall operation of the tractor (on the basis of how each operates in relation to the overall tractor).

Here, the pieces are all in the text, but have not been put together in quite the same way before. This answer might also include comparing and contrasting the pieces, but not necessarily.

General Principle or Definition Created?

Here, one might be presented with a number of scenarios or examples, and then they come up with a definition or principle that is not given in the text. Consider when Skinner created principles such as reinforcement, extinction, and punishment: All of the examples exist in nature, but the principles had not been identified/defined in this way before. In addition, he provided operational definitions for each of the principles. In this way, he was synthesizing the observations made into a coherent taxonomy of behavioral principles that affect behavior (Level 5).

Comparison or Contrast Made?

Here, the answer provides similarities or differences that are not found in the textbook. For example, one might state how the length from the inside of the elbow to the wrist is related to the length of one's foot from the heel to the end of the big toe (e.g., they are usually the same length). One might also state how punishment and negative reinforcement are similar or different. If the answer is not in the text, then it is a Level 4. (However, if the answer is given in the text, then it should be a Level 2, in which case a mistake was made when deciding whether the answer was wholly from the text.)

Example Explained

Here, the answer provides some detail about why a particular example fits a definition, process or concept. It is one thing to provide an example and another to detail why it is correct in terms of the component(s) of the definition, process or concept. For example, one might provide an example of reinforcement, but when they explain it in terms of an

antecedent, behavior, consequence, and the probability of the behavior in future similar circumstances, then it is assessed at Level 4.

Identify or Provide an Original Example

Here, the answer provides an example that is not in the textbook as it pertains to the concepts being covered. The answer may identify ''mammal'' when presented with ''What is a human?'', and if that was never mentioned in the textbook, then it is a Level 3. Alternatively, one might provide several examples of mammals that are not in the textbook. In the former case, they simply identify the classification, while in the latter they generate examples based upon the classification. Examples can be of definitions, principles, or concepts. (Level 3).