Analysis of Proctor Feedback Accuracy in a Computer-Aided Personalized System of Instruction

Toby Martin

University of Manitoba

A Thesis Submitted in Partial Fulfilment

of the Master of Arts Degree at the University of Manitoba

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION PAGE

Analysis of Proctor Feedback Accuracy in a Computer-Aided Personalized System of Instruction

BY

Toby Martin

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 Master of Arts

Abstract

Information stored in computer files was analyzed to assess proctor feedback accuracy in a Computer-Aided Personalized System of Instruction. The effect of feedback on subsequent student responses to test questions also was assessed. Participants were students who in the Fall of 1996 completed an undergraduate psychology course at the University of Manitoba. Participants had no contact with the researcher; instead, records of past student-proctor interactions were examined. Overall proctor accuracy was 72.6%. However, accuracy dropped to 36.0% when proctors graded incorrect answers. Students complied at least partially with 61.3% of all instances of feedback; this value changed little as a function of feedback importance.

Analysis of Proctor Feedback Accuracy in a Computer-Aided Personalized System of Instruction

Computer-Aided Personalized System of Instruction combines several alternatives to traditional methods of teaching. Developed by Joseph Pear and colleagues at the University of Manitoba (Kinsner & Pear, 1988), CAPSI utilizes both computer-mediated communication (the students communicate with the instructor and teaching assistants via E-mail) and computer-managed instruction (quizzes and exams are generated by a computer program and e-mailed to the students). CAPSI also shares many features with Personalized System of Instruction (PSI), after which it was named, first described by Fred S. Keller (1968). In both systems, students are provided with clear learning objectives, which are based on the content of written materials rather than the content of lectures. The objectives are grouped into small study units, through which the students are free to move at their own pace. The students write frequent tests on the study units, and must demonstrate mastery of each unit before moving on to the next. If a student is unsuccessful on a first attempt at a unit test, he or she may reattempt that test as often as time permits.

Another feature central to both CAPSI and PSI is the use of student assistants called proctors, who score tests and provide rapid and personalized feedback. In CAPSI, students act as proctors to other students in the same course: a student is eligible to act as a proctor and mark a test on a particular study unit if he or she has already successfully completed that unit. CAPSI unit tests consist of several short essay questions, and there are typically 10 study units in a 12-week course. Students who complete all 10 units have

therefore emitted a considerable amount of verbal behaviour related to the study material. Since proctors are the primary source of feedback on this behaviour, it is desirable that the proctors be trained to provide the most effective feedback possible. To date, however, no study has assessed the quality or the content of the feedback currently being given by CAPSI proctors. This thesis remedies that deficiency. First, however, we look at (a) a number of efforts to combine computer-based instruction with PSI, and in particular how these efforts addressed the proctoring component of PSI; and (b) a more detailed description of the CAPSI method of instruction, especially in regards to the proctoring system.

The Personalized System of Instruction

In 1963, a group of four psychologists accepted an invitation to help create a Department of Psychology at the newly founded University of Brasilia. As part of their planning they developed a new method of instruction, one whose features were based upon principles of behaviour that had been established in the preceding decades through the experimental analysis of animal and human behaviour. One of the four innovating psychologists, Keller, is usually credited as the originator of the method. He summarized the application of reinforcement theory to the development of the new method in this way:

"The designers of the system had the goal of maximizing the rewards for educational behaviour, minimizing chances for extinction and frustration, eliminating punishment and fear, and facilitating the development of precise discriminations" (Keller & Sherman, 1982, p. 48).

Using these guidelines, the psychologists developed what came to be known as the

Keller Plan, or the Personalized System of Instruction. As originally conceived, a PSI course would include the following features:

(a) clear study objectives,

(b) the division of the material to be learned into small units that can be studied in a week or two,

(c) frequent tests on the study units in which students must demonstrate mastery of the study objectives,

(d) mastery criteria so that students must demonstrate mastery at a particular level before going on to the next level,

(e) a number of student assistants (called proctors) to immediately score tests and provide feedback,

(f) a "go-at-your-own-pace" feature, and

(g) lectures used primarily for motivation and demonstration rather than as a major means of presenting information. In contrast, the most common approach to university teaching at the undergraduate level involves lecturing in combination with two or three exams per semester (Terenzini & Pascarella, 1994). Since its development, PSI has been used to teach courses from virtually all subject areas, at hundreds of institutions in different parts of the world. A meta-analysis of many experimental comparisons between PSI and the more traditional approach to university teaching has demonstrated that PSI is considerably more effective at helping students to learn (Kulik, Kulik, & Bangert-Drowns, 1990).

In the years following Keller's original exposition of the method, many efforts

were made to understand PSI's effectiveness by determining the relative importance of its various features. To achieve this, researchers compared outcomes in courses where one or more of the basic elements were modified or omitted. The results of this research indicate that a version of PSI that emphasizes mastery criteria, frequent testing on small units, explicit study objectives, and immediate feedback should perform at least as well as one that included all of the components specified by Keller (Kulik, Jaksa, & Kulik, 1978). Research by Cooper and Greiner (1971) and DuNann and Weber (1974) suggests that an alternative individualized instruction method developed by Jack Michael at Western Michigan is as effective as PSI, but includes only those features whose importance is supported by the research data.

Towards Computer-Based PSI

In recent years, a number of university instructors have integrated PSI with another powerful educational technology, computer-based instruction. A meta-analysis of studies evaluating the effectiveness of CBI (Kulik and Kulik, 1991) has shown that, like PSI courses, computer-based courses are generally superior to the traditional alternatives at helping students to learn. An effective synthesis of the two approaches would enjoy the advantages of both, resulting in a method of instruction with great potential.

The features of a computer-based PSI course will vary depending upon: (a) how closely the course designer adheres to Keller's original plan for PSI, and (b) what functions the computer is employed to perform. Designers of computer-based courses have most commonly assigned the computer to one or more of three functions. First, they have been used as a medium of communication within the classroom, an approach usually

referred to as computer-mediated communication (Harasim, 1989; Hiltz, 1986; McComb, 1994). Second, they have been used as a tool to structure and administer lessons and exams, and keep records of student performance (Crosbie & Kelly, 1993; Halcomb et al., 1989; Kinsner & Pear, 1988; Pear & Kinsner, 1988). This approach, which is usually called computer-managed instruction, does not entail an active teaching role on the part of the computer. In contrast, the third major way in which computers are used in education has the computer assume the role of tutor. Known as computer-assisted instruction, this method uses specially designed software to present study materials in ways that hopefully will promote mastery (Kulik & Kulik, 1991; Kulik, Bangert, and Williams, 1983; Tudor, 1995). The following section summarizes two examples of computer-based PSI in university settings, and describes the courses that were created in terms of both their PSI features and their computer-based features.

Successful Applications of Computer-Based PSI

Crosbie and Kelly (1993) described a successful effort to combine PSI with computer-managed instruction. Their method was used to teach 51 students who were enrolled in an eight-week course in applied behaviour analysis. The instruction was in PSI format to the extent that:

(a) The course material was divided into 18 study units, and students had to demonstrate mastery of each unit (by scoring 90% or higher on a multiple-choice unit test) before moving on to the next unit. Only the grade received on the first attempt, however, counted toward the final grade (a final grade breakdown is described below).

(b) The students were free to proceed through the study units at their own pace.

(c) The material to be learned was in written form, and there were no lectures.

(d) Clear objectives, in the form of study questions, were provided for each unit.

The instruction differed from the traditional PSI format in that:

(a) The students were required to write four multiple-choice review tests, spaced throughout the term. The scores from these tests made up the largest part of a student's mark (72% of the total grade) while the unit tests counted only for 18%. A grade received on a written report made up the remaining 10% of the final grade. The review tests were not truly self-paced: deadlines were in place for each test, and small bonuses and penalties were awarded contingent upon meeting or not meeting the deadline. Only a student's first attempt on a particular review test was counted toward their final grade, but students receiving less than 90% were required to retake the review test until they had satisfied the mastery requirement.

(b) The students acted as their own proctors. Test delivery and feedback (for both the unit and review tests) were administered by computer. When a student wished to write a test, he or she would bring a personal diskette to a microcomputer laboratory made available for purposes of the course. The diskette contained a copy of the course PSI program, and upon command from the student would generate a test of 10 questions for a unit test, or 15 questions for a review test. Test items were drawn from a coded database containing sufficient questions for the student to repeat a test on a particular unit several times without encountering a question more than once. When all of the required test items were completed, the computer prompted the student to self-score their test by showing the question, the correct answer, and the student's answer for each item. The student assessed

each answer as either correct or incorrect, and the results, the correct answer, and the student's answer were all written to an output file for immediate perusal by the instructor, who was on hand in the microcomputer laboratory. The instructor discussed any problems with the student, recorded their grade, and updated the student's personal diskette, so that it would generate a copy of the next required test.

Although there was no experimental comparison between this computer-based section and a non-computer-based control group, Crosbie and Kelly considered their program a success. The computer-based PSI produced:

(a) a low rate of student drop-out (i.e. no students withdrew from the course),

(b) low procrastination of test-taking (only 2 students failed to complete all tests by the final deadline),

(c) high student performance as measured by final grades (42% of the students received an A), and

(d) positive student reactions, as measured by a post-course questionnaire.

In addition to the positive effects on students, Crosbie and Kelly found that their program was effective at reducing paperwork, and at keeping the instructor's duties to a minimum.

An on-line article published by Alan Tyree at the University of Sydney describes a version of PSI that incorporates both computer-managed instruction and, in a later modification, computer-based tutorials. Tyree's earliest effort at computer-based PSI involved a 1990 course in Technology Law, involving approximately 130 students. In accordance with the traditional PSI structure, Tyree's course consisted of a series of 15

study modules, each with a number of specific learning objectives. Written materials were heavily emphasized, and although five optional lectures were given, they were poorly attended. The students worked at their own pace, studying a given module until they were prepared to attempt a test. If they demonstrated mastery of the learning objectives, they could then proceed to study for the test on the next module. If the student failed to demonstrate mastery, no penalty applied, and the student was free to attempt another test on that module after a period of restudying.

The major function of the computer was to assume the role of proctor. When writing a module test, the students worked at a local terminal, answering a series of questions based on the behavioural objectives for that module. The questions asked were a kind of extended multiple choice, or "tree questions", as Tyree describes them. For each tree question, the computer displayed a short paragraph describing a hypothetical legal situation, and then asked a series of related questions to test the student's understanding of the relevant legal concepts. The tree structure of the questions was intended to more thoroughly examine the reasoning process of the student as it related to the information from the short paragraph. Module quizzes averaged 10 questions in length, and generally required approximately 20-30 minutes to complete.

Tyree considered the computer-based PSI course a qualified success. While the computer testing system performed all of its duties adequately, some students were dissatisfied with the inflexible nature of the questions themselves. Tyree judged that these complaints were associated with the poor construction of some test items, and were due to the inexperience of the teacher, and not the computer-based nature of the course.

Despite their occasional complaints, the students performed well within the computer-based PSI structure, though as in the case of the Crosbie and Kelly study, no non-computer-based control group was available to provide an experimental comparison. Of the 132 students who attempted at least one of the module tests, 116 completed all 15 modules, and only four of the students failed to complete enough modules to pass the course. Of those students that passed, 92 of them achieved a final mark of "high distinction" (that is, a grade of 85 percent or higher).

CAPSI

The computer-based version of PSI developed by Joseph Pear and colleagues at the University of Manitoba is highly flexible, and has been continually modified and improved since its development in 1983. The following summary describes both the computer-aided aspects of CAPSI and its unique approach to proctoring, while focussing on those characteristics that have remained relatively constant over the years.

Computer-mediated communication features. In CAPSI courses, the principal medium of communication among the students, the instructor, and the teaching assistant is e-mail (actually, a file-access process similar to e-mail). Regular class meetings are sometimes scheduled in order to provide the students with a forum for questions about the course, but in general, lectures are de-emphasized in favour of textual materials. Often the instructor will meet face-to-face with the students only at the very beginning of the course (to explain the course procedures), and at the very end of the course (to invigilate the final exam). Recently, even the introductory class meeting has been made unnecessary by a carefully prepared manual that instructs the students in the use of the computer system.

Computer-managed instruction features. The primary functions of the specialized software that is used to manage a CAPSI course are to generate unit tests and deliver them to students, assign proctors to grade a completed unit test, and deliver the completed test to the selected proctors via e-mail. Once the proctors (or the instructor or teaching assistant) have marked the test and provided written feedback, the test is mailed back to the student. This method provides students with more rapid feedback than would normally be possible in a traditional (non-PSI) university course: a CAPSI student can expect to receive their marked unit tests within 24 hours of submitting them.

The proctoring system. The unit test questions in CAPSI require student-generated (i.e. short essay) answers. These answers are evaluated not by the computer, nor by the student who wrote the test, but by other students within the course. When a student finishes writing a unit test, he or she submits the test to the CAPSI program, which then assigns the test to two eligible proctors. If two proctors are not available, the program sends the test to either the instructor or the teaching assistant.

When proctors are assigned a unit test to grade, their primary responsibility is to make a pass/restudy determination. A pass on a unit test should be assigned only when all answers demonstrate mastery of the relevant concepts. The proctors should also point out specific deficiencies in an answer, if it fails to demonstrate mastery. In other words, they should explain what the student would have had to do in order to receive a pass. Proctors are also expected to make comments that are positive and encouraging rather than critical.

An advantage of this method is that the test questions may be of the essay variety. In many computer-managed systems, the fact that the computer is responsible for grading

the tests sharply limits the kinds of questions that can be asked. In contrast, the CAPSI system requires students to demonstrate mastery of the course material through the construction of complete answers, a behaviour that constitutes an important learning objective in many courses.

Johnson, Sulzer-Azaroff, and Maass (1976) demonstrated another important advantage of employing student proctors to provide feedback. In their study, as in CAPSI, students who had successfully completed a particular unit could serve as proctors for other students in the same course who were attempting a test on that unit. This situation is a departure from Keller's original design (Keller, 1968), in which the proctors for a given PSI course likely would be individuals who had previously completed that course. The study found that those students who served as proctors performed significantly ($p < .05$) better on the final exam and on a generalized course achievement test than those who did not.

The CAPSI proctoring system has pitfalls as well as advantages. One concern is that the proctors assigned to mark a particular test may not do so promptly, a problem that is currently addressed by a bonus point system. Proctors ordinarily receive half a point toward their final grade for each instance of proctoring they perform. However, if a proctor has made him or herself available to mark tests, and fails to mark any test within 24 hours from the time it is assigned to him or her, half a point is deducted from the proctor's final grade total. This delay is still much greater than what students would experience in a traditional PSI course, but it is not clear that the difference has practical significance. While a meta-analysis by Kulik, Jaska, and Kulik (1978) concluded that

"delaying feedback in PSI courses interferes with student retention of course material", a subsequent article by Robin (1978) questioned their conclusion. Robin suggested that previous attempts to analyze the importance of immediacy were confounded by the simultaneous manipulation of several course features. He conducted a study to isolate the effects of immediate versus delayed feedback, and found that while students have a strong preference for immediate feedback in PSI, the feedback can be delayed by at least 24 hours without any detrimental effects on academic achievement.

Another concern is whether CAPSI proctors make accurate discriminations about the quality of the answers they grade. Accuracy is important for at least two proctor duties. First, proctors must correctly discriminate between adequate and inadequate answers on unit tests. Students are currently protected against unreasonably demanding proctors by a right to an appeal. A student who receives a "Restudy" result on a unit test because a proctor judged one or more answers to be inadequate can plead his or her case to the course instructor, who will review the student's answers and decide whether the test result should be changed to a "Pass". Unfortunately, students have less incentive to appeal a test result that they feel was too generous; an experimental analysis of proctor test-scoring accuracy by Sulzer-Azaroff et al. (1977) found that retention of course concepts is reduced when PSI proctors are too lenient in grading test items related to those concepts.

A second proctor behaviour for which accuracy is desirable is the identification of the specific deficiencies of inadequate answers. It is presumably of limited pedagogical value for a proctor to correctly identify an answer as inadequate, if he or she neglects to

give the student accurate information about what would have made the answer better. The potential importance of this second type of behaviour becomes clearer when proctor feedback is analyzed in terms of rule-governed behaviour.

Rule-governed behaviour and proctor feedback. In examining whether CAPSI proctors are providing feedback that can effectively refine the verbal behaviour of students, a survey of past analyses of PSI proctor behaviour provides little assistance, for several reasons. First, there is a tendency in PSI studies (e.g. Weaver & Miller, 1975; Robin & Heselton, 1977; Robin & Cook, 1978) to describe interactions between proctors and students in ways that only make sense in the context of face-to-face, real-time communication. Friendly greetings and closing comments (e.g. Weaver & Miller, 1975), attentive listening (e.g. Johnson, 1977), and prompting (e.g. Quigley, 1975) have all received attention from PSI researchers, but do not translate well to e-mail communication. Second, most of these behaviours, with the exception of prompting, have limited value as sophisticated tools for modifying verbal behaviour.

Feedback from CAPSI proctors is limited to written comments that will probably not be read by the student for at least several hours, and possibly not for several days. Proctors can explain what would have been needed to make an answer complete, or provide praise for any aspect of an answer that was especially well done. Proctors are also given the opportunity to provide general comments about the test or about the student's progress in the course. Unfortunately, it is doubtful that any praise provided by CAPSI proctors serves to reinforce test-taking behaviour. The praise is much too delayed to have a direct-acting effect, and a realistic descriptive analysis should probably identify course

grade as the major motivational factor (Michael, 1991). In other words, CAPSI students write and submit unit tests not to receive praise from proctors, but to receive points towards their final grade.

Could praise from proctors have an effect on other student behaviours? For example, if a proctor praises a student for using an example to clarify their answer, would the student be more likely to use examples in the future? Again, the praise being given is too delayed to have a direct-acting effect. If the student's subsequent example-giving behaviour increased in frequency, a different explanation would be needed. One explanation of the value of such feedback is that it is providing the student with a rule for responding, which can be applied to a wide range of situations (questions). Any analysis of student behaviour in this context is complicated by the fact that many CAPSI students use external memory aids when writing unit tests. Pen, paper, the course text, and study notes may all be used, although use of the text and notes during unit tests is discouraged by the instructor. Thus, whether they receive feedback applicable only to a specific question, such as, "Good explanation of this difficult concept!", or applicable to many questions, "Great use of examples!", responses to future questions may to some extent be controlled by what was recorded on paper while writing previous unit tests. The consistent presentation of accurate, relevant rules is no less important for the use of an external memory aid, however. In the absence of these rules, the students would be writing down and learning less complete answers.

Rule-governance also can explain the value (if any) of feedback provided for inadequate answers. When a proctor writes, "Good answer, but to make it complete, you

need to mention XYZ, found on p. 123 in the textbook", he or she is giving a rule, which provides an implicit antecedent, "when asked this question", a response, "write your answer plus XYZ", and a consequence, "your answer will be complete (and earn full marks)". An informal survey of proctor feedback suggests that comments of this sort are fairly frequent, and that proctors are more likely to provide students with rules when an answer is inadequate than when it is adequate.

Statement of the Problem

CAPSI can be considered a success in terms of student satisfaction. A study by Pear and Novak (1996) revealed that a majority of students (64%) were generally satisfied with CAPSI, and 77% percent of the students surveyed indicated that they would take another course that was taught using CAPSI. A potential concern with the CAPSI system is that the primary source of feedback to students on the quality of their answers to the study questions is the student-proctor interaction. Previous research suggests that accuracy is an important dimension of proctor behaviour, but no assessment has yet been made of CAPSI proctor accuracy, or of how proctor feedback affects student learning in a CAPSI course. The following studies address both of these deficiencies.

STUDY 1: ASSESSMENT OF PROCTOR GRADING ACCURACY

Method

Participants

The participants were 33 students who in the fall of 1996 completed "Principles of Behaviour Modification", an undergraduate psychology class at the University of Manitoba (Course #17.244). None of the participants interacted directly with the

experimenter. As each student worked through the 10 study units into which the course material was divided, the CAPSI program recorded the unit tests that were given, the answers written in response to those tests, and the feedback given in response to the answers. The information recorded in the CAPSI files constituted the data base of the research.

Materials

The primary materials were the CAPSI files in which all instances of proctoring are recorded, and a personal computer on which the data files were viewed. Additional materials included a course procedures manual which explains the course format and provides instructions on how to use the computers and the CAPSI program, the course textbook, Behaviour Modification: What it is and How to Do it, 5th Edition (Martin & Pear, 1996), and the instructor's manual for the textbook.

Procedure

The midterm and final exams for the course were inspected in order to determine which study questions were asked on each exam. All unit tests were then examined, in order to identify tests which met either or both of two conditions. The first condition was that the test contained a question which was asked again on a subsequent exam; the 55 tests which met this condition will be referred to as Sample One. The second condition combined two requirements. First, the test had to contain a question that was asked again of the same student on a subsequent unit test, or that had already been asked of that student on a previous unit test. This situation sometimes arises when a CAPSI student receives a restudy result on a unit test; his or her next attempt on the test for that unit may

include one or more study questions that they have already encountered on a test. The second requirement was that the student's answer to the first appearance of the repeated question had to have been given detailed feedback from at least one marker (that is, feedback other than a simple "correct" or "good answer"). The 82 tests that met this condition (including six that were also selected for Sample One) will be referred to as Sample Two.

The two samples taken together (but counting the six common tests only once) formed a combined sample of 131 unit tests, 25.2% of the total 523 unit tests that were written during the course. The criteria for selection made it possible to analyze how student responses to successive encounters with the same question changed following the presentation of detailed written feedback. This analysis, as well as an assessment of the content and accuracy of proctor feedback, will be described in Study 2.

All analyses in this study were restricted to the 101 unit tests from the combined sample described above which were graded by at least one student proctor. (Generally, a unit test is graded by either two proctors, the instructor, or the teaching assistant. Occasionally, however, when one of the proctors assigned to a test does not grade it within the allotted 24 hours, that proctor's duty will be performed by the instructor or TA.) For each test meeting all of these conditions, both the primary researcher (who had previously completed the course and subsequently worked as a TA for it) and an assistant with expert knowledge of the course material (one of the authors of the text used in the course) made an independent determination of the correctness of all the answers given on that test. The researchers applied the same criteria for mastery of the subject matter that

the proctors were instructed to use when grading unit tests. Specifically, an answer was judged to be correct only if all parts of the question were addressed correctly and completely, with appropriate reference to the relevant behavioural principles and procedures described in the course textbook. The researchers also identified in writing the specific deficiencies of any answer that was judged to be inadequate. An inter-observer-reliability (IOR) estimate was made of the researchers' assessments of the correctness of each answer. The IOR was calculated by dividing the total number of answers on which the observers agreed (251) by the total number of answers which were assessed by both observers (302), and multiplying by 100%; agreement was 83.1%.

In the event that the researcher and assistant disagreed on the adequacy of an answer, they discussed the answer's merits and deficiencies in order to reach a consensus about its correctness. Out of 302 questions graded, the researchers were unable to reach a consensus on only three questions, which were discarded from consideration in Study 1.

Results

Table 1 provides a comprehensive summary of what was learned about proctor accuracy. Proctors made errors (that is, disagreed with the observers) in 27.4% of all instances of proctoring. (One proctor evaluating one answer equals one instance of proctoring, or IOP.) 80.0% of all errors were of the sort where a proctor judged an answer to be correct while the researchers judged it to be wrong. This kind of error (for convenience, referred to here and in the tables as "+error") occurred in 64.0% of all instances in which a proctor graded a wrong answer. CAPSI requires two proctors to assign a pass to a unit test in order for the student to progress to the next unit. 56.7% of

all incorrect answers were detected by at least one proctor; having two proctors grade each test therefore reduced the percentage of undetected wrong answers from 64% to 43.3%.

Accuracy was also assessed separately for each month of the course. Total errors as a percentage of IOPs, +errors as a percentage of total errors, and wrong answers as a percentage of total answers all showed a slight but consistent decline from September through November (please refer to Table 1 for values). Chi-square values for these items: χ^2 (2, $\underline{N}$ = 160) = 1.33, $\underline{p}$ > .05; χ^2 (2, $\underline{N}$ = 128) = 0.57, $\underline{p}$ > .05; and χ^2 (2, $\underline{N}$ = 104) = 5.04, $\underline{p}$ > .05, respectively.

Randomness of the samples. Sample One was quasi-random, since the CAPSI program pseudo-randomly selected the three questions that made up each unit test. Thus, the exam questions will have previously appeared on a random sample of unit tests. However, there was no guarantee that all units were equally represented. An exam question taken from Unit 2, for example, will have previously appeared only on tests that were written on this unit. In the CAPSI course used for these studies, none of the exams contained questions from Unit 1, which is an introductory unit on the course procedures. Also, on the first midterm exam, one question was taken from Unit 2, one question was taken from Unit 3, and two questions were taken from Unit 4. Similarly, two out of four questions on the second midterm were taken from Unit 7, with Units 5 and 6 providing one question each. (Interestingly, it is actually Units 3 and 6 that were over-represented in Sample One, but in principle the problem stands. These units were also over-represented in Sample Two, suggesting that students just made more attempts at these units, which in

turn suggests that they were the most difficult.)

Sample Two was not random. Repeating unit tests is correlated with the production of poor answers; hence, the tests that made up Sample Two probably had a higher rate of wrong answers than the overall population of tests. Indeed, a test was only selected for Sample Two if the repeated question received some detailed feedback, and detailed feedback is almost certainly more likely to be given when the answer is actually wrong. I would expect these factors to skew the statistic of overall proctor accuracy, since the data show that proctors are much more likely to make discrimination errors on wrong answers. Sample Two also suffers the same problem as Sample One with respect to how well each unit is represented.

To address these concerns, the accuracy of each sample was assessed separately. Sample One, though smaller than Sample Two, should be relatively representative, and it is therefore informative to compare the data from that sample to the data from Sample Two.

Tables 2 and 3 provide a comprehensive summary of what was learned about proctor accuracy in Samples One and Two, respectively. As expected, wrong answers as a percentage of total answers were much higher for Sample Two (41.4%, compared to 23.4% for Sample One). Total errors as a percentage of IOPs were also higher for Sample Two (31.3%, compared to 20.8% for Sample One). However, +errors as a percentage of total errors was higher for Sample One than for Sample Two (88.9% versus 76.5%), as was +errors as a percentage of proctoring instances on wrong answers (78.4% versus 59.1%).

When interpreting these statistics, it is helpful to remember that the figures expressed are totals and percentages across all IOPs, and therefore they do not reveal the variation that exists among individual proctors. A proctor-by-proctor analysis of grading accuracy (based on the combined sample of 101 unit tests) is provided in Table 4.

The Table shows that in terms of total errors as a percentage of total IOPs, proctors ranged from 0% to 50%. However, proctors differed widely in their ability to detect wrong answers. The proctor with the most instances of proctoring (51) in the combined sample encountered 13 wrong answers, and failed to identify any of them. The 3rd busiest proctor (42 IOPs) encountered 16 wrong answers, and only failed to identify two of them. (However, this same proctor made 10 -errors: over 30% of the 32 -errors in the entire combined sample.) Out of all 33 proctors, 7 had a 100% error rate at identifying wrong answers, while 14 had an error rate of 80% or higher on this task. It might be objected that perhaps these values are being inflated by proctors who only encountered one or two wrong answers and failed to identify them. However, the mean number of wrong answers encountered by these 14 proctors was 5.9, and some of the busiest proctors showed high error rates. Of the 15 proctors who graded 17 or more questions in the sample (17.7 being the mean), 5 had a 80% or higher error rate at identifying wrong answers.

Discussion

The limited randomness of Sample One and the non-randomness of Sample Two pose only a minor threat to the number and quality of conclusions that may be drawn regarding proctor accuracy. Specifically, if a non-random sample (or even a random one,

for that matter) has a greater proportion of wrong answers than the overall population, it will tend to depress measures of overall proctor accuracy, since proctors have been shown to make more errors on wrong answers than on correct ones. However, accurate measures of proctor accuracy on incorrect answers can be obtained from such a sample, and these statistics are at least as interesting and important from a pedagogical standpoint.

It is clear, for example, that stronger measures should be taken in CAPSI-taught courses to insure that wrong answers on unit tests are detected. A possible solution would be to require three proctors to grade each unit test instead of only two; this approach would have the additional benefit of increasing the amount of proctoring done in the course, which has pedagogical value in its own right, as shown by Johnson, Sulzer-Azaroff, and Maass (1976). Further, the individual proctor data (e.g. Table 4, column 7, rows 5, 13, 20, 23, 25, 30, and 33) show that at least some proctors consistently issue passes regardless of the quality of the answers they grade (and conversely, that some proctors are much too demanding; e.g. column 8, row 1). A possible improvement to the system would involve regular spot checks to identify such proctors and provide individualized feedback.

Regarding the three statistics that show a slight but consistent decline over the duration of the course, it seems reasonable to infer that one of them is a strong influence on the other two. Specifically, since wrong answers as a percentage of total answers decreased from month to month (Table 1, row 12), it should be expected that errors on wrong answers would make a smaller and smaller contribution to total errors. Hence, even though proctors were not more accurate at detecting wrong answers in the later months,

overall errors as a percentage of IOPs decreased from month to month (Table 1, row 3) as a result of a decreasing proportion of wrong answers being made.

Support for this interpretation can be found by comparing these three statistics for Samples One and Two separately. In Sample Two, wrong answers as a percentage of total answers dropped off somewhat from month to month (Table 3, row 12; 13.7 percentage points from September to November; χ^2 [2, $\underline{N}$ = 78] = 2.92, p > .05 , but to a lesser extent than for the combined sample (18 points), and to a much lesser extent than for Sample One (Table 2, row 12; 32.6 points; χ^2 [2, $\underline{N}$ = 26] = 22.97, p < .05). This relationship was mirrored in +errors as a percentage of total errors; in the combined sample this statistic dropped 8.7 points over 3 months, while in Sample One the drop was 23.8 points (χ^2 [2, $\underline{N}$ = 40] = 3.90, p > .05), and in Sample Two the statistic actually increased by 5 points: χ^2 [2, $\underline{N}$ = 88] = 0.18, p > .05. (Please see row 7, Tables 1, 2, and 3, respectively.) Similarly, total errors as a percentage of IOPs decreased by 7.8 points in the combined Sample, and by 25.9 points in Sample One (χ^2 [2, $\underline{N}$ = 45] = 16.876, p < .05), yet increased by 3 points in Sample Two: χ^2 (2, $\underline{N}$ = 115) = 0.16, p > .05. (Please see row 3, Tables 1, 2, and 3, respectively.)

A surprising result of analyzing the samples separately was that proctors in Sample Two were better at identifying wrong answers. +Errors as a percentage of IOPs on wrong answers was only 59.1% for Sample Two, versus 78.4% for Sample One (row 6, Tables 3 and 2, respectively). Since the same proctors were doing the grading in both samples, what might explain this difference of nearly 20 points? One possibility is that the greater frequency of errors on tests in Sample Two had a positive influence on the scrutiny the

proctors applied to grading them. That is, a kind of "halo effect" may have been operating on the tests in Sample One as compared to Sample Two. Perhaps when a test only has one wrong answer, proctors are more likely to overlook the deficiency if the remaining questions were answered correctly. However, a test with two or three wrong answers may prompt proctors to examine each answer more closely. Further research is required to isolate and confirm this relationship.

Factors not addressed in this study but which may be important determinants of proctor accuracy include question difficulty and proctor experience. Are there questions, or types of questions, for which proctors have more difficulty assessing whether they have been answered correctly? How does proctor accuracy change as a function of how many tests the proctor has graded? Hopefully, future analyses of proctor accuracy will attempt to answer both of these questions.

STUDY 2: EFFECT OF FEEDBACK ON SUBSEQUENT STUDENT RESPONSES

Method

The participants and materials employed in this study were identical to those employed in the previous study.

Procedure

The problem of feedback definition. Clearly, not all feedback is equal. Markers produce a wide range of comments in response to the answers they grade, and these comments vary in terms of their content, accuracy, importance, and any number of other properties. If the likelihood that a given instance of feedback will produce a desirable change in behavior depends at least in part on the properties it possesses, it is important to

understand this relationship. Investigating this issue is a prerequisite to addressing the practical question: what can be done to make feedback more likely to be followed? In other words, what properties should feedback have?

It is important to define and demarcate instances of feedback (IOFs) precisely because: (1) to apply any analysis of the properties of feedback (e.g. accuracy), it is necessary to know exactly what to apply it to, and (2) even if many IOFs seem easy to identify, the extent to which they were followed may not be so clear. The following sections describe the procedure used to identify and classify IOFs, assess their accuracy, and judge the extent to which they were followed.

Identifying and classifying IOFs. All IOFs were operationally identified by considering whether or not the student could have "followed" the comment; that is, by considering whether the comment modeled, stated, implied, or requested a behavior that the student could have (and presumably, should have) emitted in the context of writing a subsequent answer. However, a distinction was made between general IOFs, which can be applied to many different answers, and question-specific IOFs. Question-specific IOFs (the focus of this study) can only be used to guide changes in a student's answer to the particular question on which the feedback is given.

Determining precisely where a question-specific IOF began and ended was accomplished by identifying it as one of the five types described below.

Type 1. The marker provided a model of the correct answer (or some part of it) for the student. The criteria for judging membership in this type was whether the comment, or any part of it (even one word, if that word was clearly the key point that was

missing/wrong in the student's answer) could be "cut and pasted" directly into the student's answer. Discrete IOFs were defined for the sake of convenience as a word or contiguous series of words, up to and including a full sentence. Serializations were an exception: if a sentence provided an itemized list of things that needed to be included in the answer, each item in the list was a separate IOF. Similarly, within a single sentence any independent clauses separated by "and" or "but" were counted separately.

The models in Type 1 IOFs were often prefaced by some words like, "You need to mention...", making the comment resemble a Type 2 IOF. However, it did not matter that the comment could also be interpreted this way; as long as some part of the sentence could be transplanted more or less directly into the student's next answer, it was treated as a Type 1 IOF.

Type 1 feedback can be concerned with either behavioral excesses or deficits. Markers sometimes addressed an excess using Type 1 feedback by identifying exactly what the student said that needed to be removed from the answer. For example, one marker simultaneously identified a deficit and an excess by writing, "...in the next two parts of the question you want to say 'capability', not 'ability'."

Type 2. The marker provided a description of a change in the student's verbal behavior that would improve the answer (at least in the opinion of the marker, as best this could be inferred). The distinguishing feature of IOFs in this category (as compared to Type 1) was that the marker did not give the student a model of what needed to be done. For example, the comment, "You need to mention that extinction can be used instead of punishment when an undesirable behavior needs to be decreased" is a Type 1 IOF:

everything that appears after the words, "You need to mention that..." can be used by the student as a model for their next answer. However, the comment, "You need to mention an alternative to punishment" is Type 2: a deficiency is brought to the student's attention, but they must formulate an improved version of their answer on their own.

Like Type 1 feedback, Type 2 IOFs can identify an excess or a deficit in a student's answer. For example, a marker may write, "you overemphasized the respondent components of the emotional response".

Type 3. The marker gave one or more examples. There are some details of examples given by markers that students would not be expected to incorporate into their own answers (i.e. the particulars of the situation, behavior, etc.). However, a marker's example may have had one or more features required for a correct answer that were not present in the student's example(s). Each of these features was an IOF. This type of IOF can be difficult to demarcate, since the features may have to be abstracted from the particulars provided by the marker. When reading an example provided by a marker, the researchers asked, "what did the marker do in this example that the student didn't do in theirs?" The answer was the IOF(s).

Type 4. The marker asked a question whose answer would be a beneficial addition to the overall answer, or to draw attention to a mistake the student has made. Each question counted as a separate IOF.

Questions were occasionally asked in "multiple choice" fashion (e.g. "What defines 'indirect'? 1 s? 30 s? One hour?"). Obviously, not all of these questions would have to be answered in order to demonstrate compliance with the feedback. Accordingly, they were

counted as a single IOF which was followed if the correct option was selected.

Type 5. The marker gave one or more page numbers from the textbook, implying or asserting that the student should read those pages to find the information they needed to improve their answer.

Markers made many comments that were not counted as IOFs in this study, despite the fact that they were indeed feedback and might have had an effect on subsequent responding. Five sorts of comment that were excluded from consideration are described here.

1. Statements that identified only what the student did correctly, even if that statement was a prelude to identifying a problem with the answer, were not considered IOFs (e.g. "You are on the right track re: the notion of extinction. However...").

2. Statements that were offered as explanation for, or justification of, an IOF were not considered IOFs. For example, "You should also add that originally both men were on a continuous reinforcement schedule. If this [i.e. the way the student described it] were the way the study was actually carried out, one could simply say that the man on the intermittent schedule was more resistant to extinction for a variety of history reasons". In this example, the first sentence is clearly a Type 1 IOF. The 2nd sentence is a rationale for why the response indicated by the Type 1 IOF needs to be included in the answer.

3. Statements that asserted, "you don't have to do this, but you might", without indicating that it was necessary for mastery or even that it would improve the answer were not considered IOFs. A good example: "Note that behavior selected for baseline does not have to be one of social interaction, simply because you are trying to determine if 'social

attention' of adult is a reinforcer." The marker may have been trying to clear up some confusion on the part of the student, but the student's next answer could still have demonstrated mastery without following this comment.

4. Statements that recommended covert behavior (e.g. "you need to re-examine your understanding of this") were not considered IOFs.

5. Statements that were so ambiguous or poorly worded that it was impossible to tell what the marker meant were not considered IOFs.

IOF accuracy. IOFs were inaccurate in at least two important ways. Type A errors were made when a IOF was not based on an accurate reading of the answer. In other words, the marker asserted that the student needed to include something in their answer which was in fact already present (or in the case of Type 4 feedback, asked the student a question which they had already answered). Alternately, the marker may have asserted that the student needed to remove something from their answer which wasn't really there.

Type B errors were made when a IOF was not consistent with the information found in the text (or in the course lectures, where different). In other words, the marker told the student to do something that was wrong, or not to do something that was right.

It is possible for an IOF to be both an A and a B error, though no examples were found in this study. It would mean that the marker admonished the student to include X in their answer when X was already there and wrong (that is, inconsistent with the information in the text). An alternate sort of A/B error would involve a marker suggesting that the student remove Y from their answer when Y wasn't really there, but furthermore Y should be there (i.e. if it were there, removing it would have made the answer wrong).

However, this is such a strange tangle of mistakes that it is hard to imagine it ever actually occurring.

Type 1 IOFs can be quite detailed, so it is possible that only a part of the model may be inaccurate. When evaluating the accuracy of a Type 1 IOF, the observers identified all of the ways in which it strayed from 100% accuracy, and stated the type of error in each case. Three levels were used to assess accuracy: zero, partial, and total. This approach acknowledged the relative sophistication with which accuracy of Type 1 feedback may be defined, without becoming bogged down in a precise word-by-word analysis.

IOF importance. Often, markers suggest changes that would result in a superior answer, but that aren't really necessary for mastery. Students undoubtedly make judgements about the importance of the feedback they receive, and hence it is necessary to evaluate IOF importance as a possible determinant of IOF compliance. In other words, are students less likely to follow non-essential comments than essential ones?

In this study IOF importance was gauged in two ways. The first approach was to consider whether the test on which the IOF was presented received a pass or a restudy. It is reasonable to believe that markers who provided IOFs on tests to which they assigned a pass thought that compliance with those IOFs was not essential for mastery. It is also reasonable to suppose that students forming judgements about the importance of IOFs might arrive at this same conclusion. The second approach was to consider what the marker asserted or implied about the adequacy of the specific question on which the IOF was provided. This analysis takes into account cases in which an answer received one or

more IOFs, but it in fact it was the inadequacy of another answer on the test which resulted in a restudy result.

Assessing compliance with IOFs. Evaluation of the extent to which an IOF was followed depended upon its type.

Type 1. Compliance could theoretically be evaluated by quantifying the degree of semantic equivalence between the IOF and what the student writes in their next answer. For example, when a marker says that "A baseline is the measurement that you take of your individually defined target behavior that you plan to change", the student's next answer to the same question might change in a way related to this IOF by stating (if it didn't already) that a baseline is:

(a) a measurement,

(b) taken of a behavior,

and that the behavior is:

i. a target behavior,

ii. individually defined,

iii. one that you plan to change.

The student's next answer may include none, some, or all of these facts, and thus, each fact potentially counts as a distinct IOF.

For practical purposes, however, analyzing a large amount of feedback in this way would be prohibitively time-consuming. A more manageable alternative, which this study used, was to identify each discrete instance of Type 1 feedback, and focus on the ways in which the student failed to produce a statement or statements having semantic equivalence

with the IOF. However, discrepancies were only recorded when they resulted in an inferior answer; the criteria here was not that the difference must be one which would prevent/permit mastery, but merely whether it contributed to a better answer at all, in the judgement of the observers. Based on this assessment, compliance was rated as either zero, partial, or full.

As previously noted, the statement being modeled was often prefaced by some words for which correspondence was not be determined (e.g. "It is important to mention that..."). The significance of these additional words might be addressed in a future study.

Type 2. Compliance with this kind of feedback cannot be evaluated by checking semantic equivalence with what the marker wrote, but must be operationalized in some other way. The researchers independently selected their own criteria for compliance with IOFs in this category, in much the same way that a marker decides whether a study question has been answered correctly or not.

Both Type 1 and Type 2 feedback were sometimes given in a conditional form, such as when mentioning an excess: "if you want to say X, you should say Y". If the antecedent X were not present in the student's next answer, it would be absurd to expect the student to include the consequent Y. If the antecedent were present but the consequent was not, the conditional counted as an IOF that did not produce compliance. If neither the antecedent nor the consequent was present, it counted as an IOF that did produce compliance.

Type 3. These IOFs were complied with if the example(s) in the student's next answer included those features which were in the marker's examples.

Type 4. Obviously, when markers ask questions, semantic correspondence can't be expected of subsequent student responses. Compliance was judged according to whether the student answered the question, in a manner similar to that described for Type 2 IOFs.

For questions not meant to evoke an answer, but to draw attention to a mistake (e.g. "Is complaining a desirable behavior that is unfortunately extinguished?"), elimination of the mistake constituted compliance.

Questions asked in "multiple choice" fashion were counted as a single IOF which was followed if the correct option was selected.

Type 5. One way of analyzing this feedback would be to say that it is a sub-category of Type 1, and that the IOFs are those elements from the quoted textbook page(s) which were missing from the student's answer. Analysis could theoretically proceed in a way similar to what was described (but ultimately rejected) for Type 1, with the semantic breakdown being performed not on what the marker said, but on the information from the textbook which the researchers judge to be relevant to improving the answer. However, apart from being prohibitively laborious, this approach makes the assumption that the students were able to find the appropriate information in the textbook, and that they could discriminate what was relevant from what wasn't. Thus, compliance with this kind of IOF was evaluated simply by judging whether the student gave the correct answer on their next encounter with the question.

The problem of redundant IOFs. Sometimes two proctors make virtually the same comment in response to an answer. Should these comments count as a single IOF, or have two IOFs been given? Similarly, a single marker may make one comment, and then repeat

it a slightly different way. How many IOFs have been given in this case?

One way to address this problem would be to count every IOF, even if redundant. Using this method would produce data that accurately reflected the frequency of the different types of IOF. It would also prevent the strange situations created by counting two or more comments as a single IOF (such as an IOF that is simultaneously two different types). A serious disadvantage of this approach is that a single behavior from a student could count as compliance with a number of IOFs. This would adversely affect the data by inflating the proportion of IOFs that produced compliance.

Another solution would be to devise a scheme for analyzing IOFs that eliminates redundancy. The scheme would provide guidelines for deciding which IOFs should be counted, and which should not, on the grounds that they are either equivalent to or entailed by another IOF. One disadvantage of this approach has already been mentioned: strange situations develop that can only be eliminated by creating guidelines that are essentially arbitrary. For example, it is possible for a Type 2 IOF and a Type 4 IOF to have exactly the same practical meaning. These should presumably be counted as a single IOF, but of what type? Another disadvantage of this approach is that potentially valuable information is obscured when some IOFs are ignored (for example, the relative frequency of the various IOF types).

The approach taken in this study was to combine these two solutions. All IOFs were identified, categorized, and assessed for accuracy. After all the data produced by this procedure were collected, a scheme for eliminating redundant IOFs was applied, in order to accurately depict the extent to which IOFs were followed.

Guidelines for eliminating redundant IOFs. Two relationships commonly exist between IOFs: equivalence and implication. Equivalence exists between IOFs that have the same operational requirement for compliance. For example, "Your answer re: rules is too vague. What specifically is meant?" The student must do one thing to comply with both of these IOFs: be more specific about what they said regarding rules. Therefore, they counted as a single IOF when analyzing compliance.

Equivalence can exist between statements which are not only of different types, but which differ in number. In each case, the objective was not to let a single behavior emitted by the student count as compliance for more than one IOF. For example, suppose a marker gave the Type 2 IOF, "you need to mention the effects of an FR schedule" and then stated the three characteristic effects (which would be Type 1 feedback). If the student identified all three effects in their subsequent answer, then the "you need to mention" statement would not count as a separate IOF. Similarly, however, the student would not be "penalized" for non-compliance more than once, even if a single behavior would satisfy more than one IOF. In the previous example, suppose the student identified only two of the characteristic effects; it would make little sense for both the third effect and the "you need to mention..." comment to count as IOFs not followed. The rule for such cases was this: when one IOF had the same practical meaning as the conjunction of two or more other IOFs, the lone IOF would not be counted.

Implication exists between IOFs A and B if compliance with A entails compliance with B, and the reverse is not true. For example, on one unit test a proctor made the general comment: "you also have to be a little bit more specific." The proctor then

elaborated by saying, "you have to identify why the response is being reinforced for occurring at a low rate." Compliance with the second comment entails compliance with the first: if the student were to identify why the response is being reinforced for occurring at a low rate, it would follow that they were being more specific. However, the student could have produced a more specific answer without complying with the second comment.

Both comments in the previous example are Type 2, but the proctor elaborated even further by saying, "In short, you have to mention... that the response has to be an undesirable one or desirable but preferred to be occurring at lower rates." Here the IOF is of Type 1. If this last IOF produced compliance, the previous two comments would not be counted as distinct IOFs (since they are automatically followed if the last one is). Consistent with this, if the second comment were followed, the first comment would not count as an IOF, but the third comment would (albeit one that didn't produce compliance). Thus, the rule of economy which guided the decision in the example about FR schedules must be supplemented by a guideline to select the most explicit IOF that was followed, when a relationship of implication exists among two or more IOFs. If it happens that none of the comments were complied with, only the most explicit comment should be counted as an IOF not followed.

The procedure outlined in the preceding sections was performed by the researcher on all unit test questions from Samples One and Two that were asked again on a subsequent exam or unit test, whether graded by student proctors or by the instructor or teaching assistant. All IOFs given to the answers to these questions were identified and rated for accuracy. Compliance was then assessed for all non-redundant IOFs.

The procedure was also performed by second researcher with expert knowledge of the course material (the same assistant described in Study 1) on 36 unit tests taken from Samples One and Two. An IOR was calculated on four separate tasks: IOF identification, determination of IOF type, assessment of IOF accuracy, and assessment of compliance with each IOF. To perform these tasks, the researchers applied criteria from an early version of the descriptions of IOF types, accuracy, and standards for compliance presented in the preceding sections. Three practice sets of unit tests, consisting of 4, 11, and 6 tests respectively, were assessed by both researchers. When each set was completed, the researchers discussed and refined the criteria for any task on which agreement was less than 80%. When agreement on all tasks was at least 80%, the final version of the descriptions and examples were written into the introduction. This occurred on the third set, which was subsequently expanded to 21 tests; the IORs were calculated on the evaluations of these 21 unit tests.

On the first task, the researchers agreed on 77 comments being IOFs, and disagreed on 19, resulting in an IOR of 80.2% (77/(77+19) x 100%). On the second task, the researchers agreed about the type of 63 of the 77 mutually acknowledged IOFs, resulting in an IOR of 81.8% (63/77 x 100%). On the third task, the researchers agreed about the accuracy of 75 of the 77 mutually acknowledged IOFs, resulting in an IOR of 97.4% (75/77 x 100%). On the fourth task, the researchers agreed about compliance with 69 of the 77 mutually acknowledged IOFs, resulting in an IOR of 89.6% (69/77 x 100%). The 77 agreed-upon IOFs represent a 35.5% sample of the 217 IOFs ultimately identified in the feedback given in Samples One and Two.

In the case of disagreements on any of these four tasks, the researchers discussed the disputed comment in an attempt to reach a consensus. A consensus was ultimately reached on all comments and all applicable tasks.

Results

Markers provided 217 instances of feedback (including redundant IOFs) to 26 students on 77 questions. Of the 217 IOFs, 167 were found to be non-redundant according to the guidelines described in the procedure section. Table 5 provides a comprehensive summary of the relative frequency of the five types of IOF, as well as IOF accuracy as a function of type. Types 1 and 2 accounted for the great majority (nearly 90%) of IOFs provided by markers (column 3). IOF accuracy was high: 88% for all IOFs, and 87.4% for non-redundant IOFs (column 5). Accuracy was slightly lower for Type 1 and Type 4 feedback than for the other types: for the 3 Types that were represented by 5 or more IOFs, χ^2 (2, $\underline{N}$ = 140) = 0.639, $p > .05$. In only one case (noted separately in Table 5) did equivalence exist between two IOFs of different types. In only eight cases was the same IOF provided by both markers.

Compliance was analyzed for all non-redundant IOFs; the results are summarized in Table 6. For 12 IOFs, compliance was not applicable, either because the IOF committed type A error, or because the student's subsequent answer was sufficiently changed that the IOF became irrelevant. Eighty-five (54.8%) of the remaining 155 IOFs were complied with fully (row 7, column 5). An additional 10 IOFs produced partial compliance, so that 61.3% of IOFs were complied with to at least some degree.

Twenty-five IOFs in the sample were applicable to a total of two answers. That is,

a student was given feedback on a particular answer, and was asked that same question twice more, either on two subsequent unit tests, or on a unit test and an exam. Of the 25 IOFs, 15 were non-redundant, and for 1 of the 15, compliance was not applicable.

For four of these IOFs, compliance was zero on the first recurrence of the relevant question. On the second recurrence of the question, compliance was zero for 2 of the 4; for the third, compliance was partial, and for the last, compliance was full.

For ten of these IOFs, compliance was full on the first recurrence of the relevant question. On the second recurrence of the question, compliance was zero for 4 of the 10. However, it should be noted that all four of these were directed at a single student, and that the relevant question recurred on a mid-term exam, not on a unit test. Compliance was full for the remaining six IOFs.

Seven IOFs in the sample were applicable to a total of three answers; five of these were non-redundant, and compliance was applicable to all five. One IOF failed to produce compliance on any subsequent answer. Two produced partial compliance on each subsequent answer. The remaining two IOFs produced full compliance on each subsequent answer.

Tables 5 and 6 characterize the IOFs provided by all markers in the course: the proctors, the instructor, and the teaching assistant. Considered separately, the instructor and TA provided 68 IOFs, 31.3% of the total 217 provided in the sample; data on accuracy and compliance for these IOFs are summarized in Tables 7 and 8, respectively. Feedback provided by these individuals was highly accurate: 92.6% for redundant and 91.1% for non-redundant IOFs, figures quite close to those for the overall sample (column

5). A greater proportion of IOFs provided by the instructor and TA were non-redundant: 56 out of 68, or 82.4%, versus 77.0% for the entire sample. Full compliance with non-redundant IOFs was also somewhat higher: 63.3% for the instructor and TA versus 54.8% for proctors only. However, the inclusion of IOFs producing partial compliance virtually erases this difference (61.3% for all markers versus 63.3% for instructor/TA). Finally, salient differences in the distribution of these 68 IOFs among the 5 types (detailed fully in Table 7) include a total absence of both Type 3 and Type 5 feedback, and greater proportions of Type 2 and Type 4 feedback. In fact, all 10 non-redundant Type 4 IOFs found in the overall sample were provided by either the instructor or teaching assistant.

In the first importance analysis (performed on a test-by-test basis) the 155 non-redundant IOFs for which compliance was applicable were considered. Fifteen of those IOFs were given on tests which received a pass, while 140 were given on tests which received a restudy. For the 15 IOFs on passed tests, compliance was 60%: 9 were complied with (either fully or partially), while 6 were not. For the 140 IOFs on "restudy" tests, compliance was 61.4%: 86 were complied with (either fully or partially), while 54 were not.

The second importance analysis (performed on a question-by-question basis) considered the same 155 IOFs. There were 61 IOFs provided on answers that all markers indicated or implied were inadequate for a pass. For these IOFs, compliance was 63.9%: 39 were complied with, while 22 were not.

There were 57 IOFs provided on answers that one marker indicated or implied was inadequate for a pass, while the other marker indicated the opposite. For these IOFs,

compliance was 61.4%: 35 were complied with, while 22 were not.

There were 37 IOFs provided on answers that all markers indicated or implied were adequate for a pass. For these IOFs, compliance was 54.1%: 20 were complied with, while 17 were not.

Another potential determinant of IOF compliance is the amount of time between the presentation of the IOF and the recurrence of the relevant answer. To evaluate this, the 155 IOFs for which compliance was applicable were examined. For the 95 IOFs which produced either full or partial compliance, mean time between the answers to which the IOFs applied was 6.28 days. For the 60 IOFs which did not produce compliance, mean time between the answers to which the IOFs applied was 4.87 days.

Discussion

IOF accuracy was impressively high. However, the fact that only 8 out of 217 IOFs were given by both markers of a single answer highlights the problem revealed by Study 1: proctors are too likely to miss or ignore deficiencies in answers.

The evidence did not indicate that an IOF's type is a useful predictor of compliance with the IOF. Although full compliance was only 45.9% for Type 1 versus 60.3% for Type 2 (for the 3 Types represented by 5 or more complied-with IOFs, χ^2 [2, N = 81] = 2.44, $p > .05$), this 15 point difference vanishes if both full and partial compliance are considered for Type 1 (Table 6, column 5). However, more data are needed to obtain a clear picture of the interaction between IOF type and compliance, especially with respect to Types 3, 4, and 5. Additional data would also help to determine whether certain students respond more favorably to IOFs of a particular type.

Overall compliance with non-redundant IOFs was rather low (about 55% full compliance, or 61% for both full and partial), and improved only slightly (to about 63%) for IOFs which were given by the instructor and TA. An interesting implication of this result is that it may be difficult to train proctors sufficiently to produce a marked improvement in the proportion of their IOFs which produce compliance. After all, both the instructor and TA in this course had greater knowledge of the subject matter and more experience providing feedback than could realistically be imparted to a student in the context of a single university course, and yet were unable to provide more effective feedback. Hopefully, techniques for improving feedback can be developed that would benefit instructors and TAs as well as proctors.

Compliance also changed little as a function of IOF importance, at least as importance was defined in this study. Whether the feedback was given along with a pass or a restudy result, and whatever the marker(s) indicated about the adequacy of the individual question to which the feedback pertained, compliance remained close to 60%.

Finally, it does not appear that compliance with IOFs is less likely when longer time intervals pass between presentation of the feedback and the next occurrence of the relevant question.

It is important to note that due to a temporary difficulty with the computer files, it was not possible to obtain data on any appeals which students made about the grade they received on unit tests. Clearly, in every case that a student had to write two or more tests on the same unit, it's safe to say that if there was an appeal, it was denied (otherwise, the student would not have had to repeat the unit test). Nevertheless, there is no accounting

for the feedback that the instructor might have given to the student at the time that they made their appeal. This feedback might have increased the student's understanding of the relevant course topic, thereby increasing the likelihood that they would comply with the original IOF. Fortunately, this circumstance was probably only present in a fraction of the cases examined in this study, and even if the percentage of compliance was inflated as a result, that percentage was still low enough to illustrate a legitimate concern with the CAPSI proctoring system.

It may be objected that the "compliance" data does not actually demonstrate a clear relationship between the presentation of feedback and improvements in subsequent answers. Without a control condition, we do not know the frequency or extent to which students improve their answers in the absence of marker feedback. Some confidence that the degree of compliance shown in Study 2 was indeed a function of marker feedback can be based on the specificity of Type 1 feedback. By definition, compliance with Type 1 IOFs entailed extremely close correspondence (almost word-for-word) between the IOF and the subsequent answer. It is therefore not hard to believe that the emission of the latter was a function of the presentation of the former, at least in the 60.3% of such cases for which at least partial compliance was shown.

Nevertheless, data should be obtained from a control condition. To accomplish this, a search was made for questions which evoked wrong answers and were repeated on later unit tests/exams, but which received no marker feedback (other than, "good answer"). Unfortunately, while it was not uncommon for wrong answers to go unnoticed by both proctors, there were few instances in which this happened on a question that was

repeated on a later test. A total of eight examples were found; in seven of these, the student did not make the necessary change to demonstrate mastery on the relevant question. Future research should also explore how compliance varies across students; again, this was not feasible in the context of the single class that was used in this study, due to the relatively small sample size.

General Discussion

Study 1 provides a comprehensive picture of proctor grading accuracy; the results suggest that steps should be taken to increase the likelihood that inadequate answers are detected by CAPSI proctors. Study 2 provides information on student compliance with feedback, and considers a number of factors in relation to that statistic. Compliance in the CAPSI course examined was lower than might be hoped (though higher than might be feared), and was not found to vary systematically with any of the other measures that were considered.

In light of these results, the single change to the CAPSI proctoring system that would provide the most benefit to students would be to make it more likely that wrong answers are detected. When students submit inadequate answers on unit tests yet still receive a pass, the mastery requirement that has been experimentally demonstrated to be a keystone of PSI's success is compromised.

Study 2 also provides a method of analyzing feedback that is comprehensive and yet simple enough to permit high agreement among multiple observers. Hopefully future studies will attempt to further define the relationship between feedback type and compliance, especially for those categories of IOF which were not well-represented in

a keystone of PSI's success is compromised.

Study 2 also provides a method of analyzing feedback that is comprehensive and yet simple enough to permit high agreement among multiple observers. Hopefully future studies will attempt to further define the relationship between feedback type and compliance, especially for those categories of IOF which were not well-represented in Study 2. It would also be interesting to examine whether different students respond differently to different types of IOF, and whether it would be effective to fade feedback from the more explicit types (e.g. Type 1) to the less explicit types (e.g. Types 4 or 5) as the course progresses. Future studies should also take advantage of the opportunity the CAPSI program provides to study in detail the effects of prompting, modelling, and rule-giving on the verbal behaviour of students.

References

Bitgood, S. C., & Segrave, K. (1975). A comparison of graduated and fixed point systems of contingency managed instruction. In J. M. Johnston (Ed.), Behaviour research and technology in higher education. Springfield, IL: Charles C. Thomas.

Blackburn, T., Semb, G. B., & Hopkins, B. L. (1975). The comparative effects of self-grading versus proctor grading on class efficiency and student performance. In J. M. Johnston (Ed.), Behaviour research and technology in higher education. Springfield, IL: Charles C. Thomas.

Bloom, B. S. (1981). Taxonomy of educational objectives : the classification of educational goals. New York, NY: Longman.

Bono, S. F., & McAvoy, R. M. (1977). Student preference and performance under a self-grading and a proctor-grading procedure in a personalized system of instruction approach. Journal of Personalized Instruction, 2(1), 28-34.

Calhoun, J. F. (1976). The combination of elements in the personalized system of instruction. Teaching of Psychology, 3, 73-76.

Cooper, J. L., & Greiner, J. M. (1971). Contingency management in an introductory psychology course produces better retention. Psychological Record, 21, 391-400.

Crosbie, J., & Kelly, G.(1993). A computer-based Personalized System of Instruction course in applied behaviour analysis. Behaviour Research Methods, Instruments, & Computers, 25(3), 366-370.

DuNann, D. H., & Weber, S. J. (1974, August/September). A two-year follow-up

study of the effects of individualized instruction. Paper presented at the annual meeting of the American Psychological Association, New Orleans, LA.

Halcomb, C. G., Chatfield, D. C., Stewert, B. E., Stokes, M. T., Cruse, B. H., & Weimer, J. (1989). A computer-based instructional management system for general psychology. Teaching of Psychology, 16, 148-151.

Harasim, L. M. (1989). Online education: A new domain. In R. Mason & A. Kaye (Eds.), Mindweave: Communications, computers, and distance education. New York, NY: Pergamon Press.

Hiltz, S. R. (1986). The 'virtual classroom': using computer-mediated communication for university teaching. Journal of Communication, 36(2), 95-104.

Heward, W. L., & Dunne, J. D. (1993) For students of behaviour analysis: A teleconference with Professor Fred S. Keller. The Behaviour Analyst, 16(2), 341-345.

Hindman, C. D. (1974). Evaluation of three programming techniques in introductory psychology courses. In R. S. Ruskin & S. F. Bono (Eds.), Personalized instruction in higher education: Proceedings of the first national conference, 38-42. Washington, D.C.: Georgetown University.

Johnson, K. R. (1977). Proctor training for natural control. Journal of Personalized Instruction, 2(4), 230-237.

Johnson, K. R., Sulzer-Azaroff, B., & Maass, C. A. (1976). The effects of internal proctoring upon examination performance in a personalized instruction course. Journal of Personalized Instruction, 1(2), 113-117.

Keller, F. S. (1968). Good-bye, teacher.... Journal of Applied Behaviour Analysis,

1, 79-89.

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

Kinsner, W., & Pear, J. J. (1988). Computer-aided personalized system of instruction for the virtual classroom. Canadian Journal of Educational Communication, 17, 21-36.

Kulik, C. L., & Kulik, J. A. (1991). Effectiveness of computer-based instruction: An updated analysis. Computers in Human Behaviour, 7, 75-94.

Kulik, C. L., & 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., Bangert, R. L., & Williams, W. (1983). Effects of computer-based teaching on secondary school students. Journal of Educational Psychology, 75, 19-26.

Kulik, J. A., Jaksa, P., & Kulik, C. C. (1978). Research on component features of Keller's personalized system of instruction. Journal of Personalized Instruction, 3(1), 2-14.

Martin, G. L., & Pear, J. J. (1996) Behaviour Modification: What It Is and How to Do It (5th ed.). Upper Saddle River, NJ: Prentice-Hall

McComb, M. (1994). Benefits of computer-mediated communication in college courses. Communication Education, 43, 159-169.

Michael, J. (1991). A behavioural perspective on college teaching. The Behaviour Analyst, 14(2), 229-239.

Miles, D. T., Kibler, R. J, & Pettigrew, L. E. (1967). The effects of study questions on college students' test performances. Psychology in the Schools, 32, 25-26.

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 of Psychology, 23(2), 119-123.

Quigley, P. A. (1975). An analysis of student manager-student interactions during performance sessions. In J. M Johnston (Ed.). Behaviour research and technology in higher education. Springfield, IL: Charles C Thomas.

Rea, C. P., & Modigliani, V. (1988). Educational implications of the spacing effect. In M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), Practical aspects of memory: Current research and issues. Chichester, England: John Wiley.

Robin, A. L. and Cook, D. A. (1978). Training proctors for personalized instruction. Teaching of Psychology, 5(1), 9-13.

Robin, A.L. and Heselton, P. (1977). Proctor training: the effects of a manual versus direct training. Journal of Personalized Instruction, 2(1), 19-24.

Ross, L. L., & McBean, D. (1995). A comparison of pacing contingencies in classes using a personalized system of instruction. Journal of Applied Behaviour Analysis, 28(1), 87-88.

Ryan, B. A. (1974). PSI Keller's personalized system of instruction: An appraisal. Washington, D.C.:American Psychological Association, Inc.

Sanchez-Sosa, J. J., Semb, G. B., & Spencer, R. (1975, September). Using study guides to promote generalization performance in university instruction. Paper presented at the meeting of the American Psychological Association, Chicago, IL.

Semb, G. B., (1974). Personalized instruction: The effects of grading criteria and assignment length on college student test performance. Journal of Applied Behaviour Analysis, 7, 61-69.

Sherman, J. G., Ruskin, R. S., & Semb, G. B. (1982). The personalized system of instruction: 48 seminal papers. Lawrence, KS: TRI Publications.

Sulzer-Azaroff, B., Johnson, K. R., Dean, M. R., and Freyman, D. (1977). An experimental analysis of proctor quiz-scoring accuracy in personalized instruction courses. Journal of Personalized Instruction, 2(3), 143-149.

Terenzini, P. T., & Pascarella, E. T. (1994, January/February). Living with myths: Undergraduate education in America. Change, pp. 28-32.

Tudor, R. M. (1995). Isolating the effects of active responding in computer-based instruction. Journal of Applied Behaviour Analysis, 28(Fall), 343-344.

Tyree, A. (1995?). The Keller Plan at Law School. [On-line article]. Available WWW: URL: http://www.law.usyd.edu.au/~alant/j_leged.html

Weaver, F. H., & Miller, L. K. (1975). The effect of a proctor training package on university student' proctoring behaviours in a personalized system of instruction setting. In J. M Johnston (Ed.). Behaviour research and technology in higher education, pp. 168-182. Springfield, IL: Charles C Thomas.

Table 1

Proctor Accuracy Summary for Samples One and Two Combined (101 Tests)

Samples One and Two	All Months	Sept.	Oct.	Nov.
Instances of Proctoring	583	109	264	210
Total Errors	160	36	71	53
Total Errors as Percentage of IOPs	27.4	33	26.9	25.2
IOPs on Wrong Answers	200	50	94	56
+Errors (marking a wrong answer as correct)	128	31	56	41
+Errors as Percentage of IOPs on Wrong answers	64	62	59.6	73.2
+Errors as Percentage of Total Errors	80	86.1	78.9	77.4
-Errors (marking a correct answer as wrong)	32	5	15	12
-Errors as Percentage of Total Errors	20	13.9	21.1	22.6
Total # of Answers	299	56	141	102
Total # of Wrong Answers	104	26	49	29
Wrong Answers as Percentage of Total Answers	34.8	46.4	34.8	28.4
# of Wrong Answers Detected by at Least One Proctor	59	14	34	11
Percentage of Wrong Answers Detected by at Least One Proctor	56.7	53.8	69.4	37.9

Note. In the context of the table, "wrong answers" refers to answers which the researchers agreed were wrong.

Total Errors: Instances in which a proctor's judgement about the correctness of an answer differed from the judgement of the researchers.

+Errors: Instances in which the proctor judged an answer to be correct while the researchers judged it to be incorrect.

-Errors: Instances in which the proctor judged an answer to be incorrect while the researchers judged it to be correct.

Table 2

Proctor Accuracy Summary for Sample One (38 Tests)

Sample One	All Months	Sept.	Oct.	Nov.
Instances of Proctoring	216	58	90	68
Total Errors	45	21	17	7
Total Errors as Percentage of IOPs	20.8	36.2	18.9	10.3
IOPs on Wrong Answers	51	24	21	6
+Errors (marking a wrong answer as correct)	40	20	15	5
+Errors as Percentage of IOPs on Wrong answers	78.4	83.3	71.4	83.3
+Errors as Percentage of Total Errors	88.9	95.2	88.2	71.4
-Errors (marking a correct answer as wrong)	5	1	2	2
-Errors as Percentage of Total Errors	11.1	4.8	11.8	28.6
Total # of Answers	111	29	48	34
Total # of Wrong Answers	26	12	11	3
Wrong Answers as Percentage of Total Answers	23.4	41.4	22.9	8.8
# of Wrong Answers Detected by at Least One Proctor	10	3	6	1
Percentage of Wrong Answers Detected by at Least One Proctor	38.5	25	54.5	33.3

Table 3

Proctor Accuracy Data for Sample Two (63 Tests)

Sample Two	All Months	Sept.	Oct.	Nov.
Instances of Proctoring	367	51	174	142
Total Errors	115	15	54	46
Total Errors as Percentage of IOPs	31.3	29.4	31	32.4
IOPs on Wrong Answers	149	26	73	50
+Errors (marking a wrong answer as correct)	88	11	41	36
+Errors as Percentage of IOPs on Wrong answers	59.1	42.3	56.2	72
+Errors as Percentage of Total Errors	76.5	73.3	75.9	78.3
-Errors (marking a correct answer as wrong)	27	4	13	10
-Errors as Percentage of Total Errors	23.5	26.7	24.1	21.7
Total # of Answers	188	27	93	68
Total # of Wrong Answers	78	14	38	26
Wrong Answers as Percentage of Total Answers	41.4	51.9	40.9	38.2
# of Wrong Answers Detected by at Least One Proctor	49	11	28	10
Percentage of Wrong Answers Detected by at Least One Proctor	62.8	78.6	73.7	38.5

Table 4

Accuracy Data for Individual Proctors (101 tests)

Proctor	IOPs	Total Errors	%	Wrong	+Errors	%	-Errors
50717	42	12	28.6	16	2	12.5	10
30305	17	2	11.7	3	2	66.7	0
40237	12	3	25	3	2	66.7	1
50138	6	3	50	4	3	50	0
50187	51	13	25.5	13	13	100	0
60075	12	3	25	4	1	25	2
60078	17	4	23.5	4	3	75	1
60122	19	8	38.1	7	5	71.4	3
60362	21	7	33.3	8	7	87.5	0
60394	3	0	0	0	0	0	0
60466	18	2	11.1	5	2	40	0
60518	12	5	41.7	5	4	80	1
60519	15	5	33.3	4	4	100	1
60755	6	0	0	0	0	0	0
61068	35	10	28.6	10	9	90	1
70058	12	2	16.7	3	2	66.7	0
70076	6	2	33.3	0	0	0	2
70090	18	2	11.1	8	1	12.5	1
70198	15	4	26.6	5	3	60	1
70233	6	3	50	3	3	100	0
70333	15	6	40	6	5	83.3	1

Proctor	IOPs	Total Errors	%	Wrong	+Errors	%	-Errors
70371	47	10	20.8	18	9	50	1
70483	23	9	37.5	8	8	100	1
70557	36	10	27.7	15	9	60	1
70795	6	2	33.3	2	2	100	0
70823	29	7	23.3	12	5	41.7	2
71009	9	4	44.4	5	4	80	0
71458	12	3	25	6	2	33.3	1
71489	6	2	33.3	5	2	40	0
71738	18	7	38.9	7	7	100	0
71874	21	4	19	5	4	80	0
72650	15	5	33.3	5	4	80	1
72846	3	1	33.3	1	1	100	0

Note. To conceal the identities of the proctors, the first and last digits of their student numbers have been deleted. The first column identified with a "%" gives the values for total errors as a percentage of total instances of proctoring. The column labelled "Wrong" gives the values for total instances of proctoring on answers which the researchers judged to be incorrect. The second "%" column gives the values for +errors as a percentage of instances of proctoring on wrong answers.

Table 5

IOF Accuracy as a Function of Type

	All IOFs						
	Total	% of Total IOFs	Accurate	% Accurate	Inaccurate	A	B
Type 1	94	41.9	78	85.7	13	9	4
Type 2	104	47.9	94	90.3	10	10	0
Type 3	3	1.4	3	100	0	0	0
Type 4	16	7.3	13	81.3	3	3	0
Type 5	3	1.4	3	100	0	0	0
All	217	-	191	88	26	22	4
	Non-Redundant IOFs						
	Total	% of Total IOFs	Accurate	% Accurate	Inaccurate	A	B
Type 1	78	46.7	66	84.6	12	8	4
Type 2	73	43.7	66	90.4	7	7	0
Type 3	3	1.7	3	100	0	0	0
Type 4	10	6	8	80	2	2	0
Type 5	2	1.1	2	100	0	0	0
Type 2/4	1	0.1	1	100	0	0	0
All	167	-	146	87.4	21	17	4

Note. In the bottom half of the table, the full title of the third column is "% of Total Non-Redundant IOFs".

Table 6

Compliance with Non-Redundant IOFs as a Function of IOF Type (for IOFs Provided by All Markers)

	Total	N/A	Full	% Full	Partial	None
Type 1	78	4	34	45.9	10	30
Type 2	73	5	41	60.3	0	27
Type 3	3	0	2	66.7	0	1
Type 4	10	3	6	60	0	1
Type 5	2	0	1	50	0	1
Type 2/4	1	0	1	100	0	0
All	167	12	85	54.8	10	60

Note. Values in the fifth column indicate percentage of IOFs which produced full compliance, after discounting any N/A (not applicable) IOFs.

Table 7

Accuracy as a Function of Type (for IOFs Provided by Instructor or TA)

	All IOFs						
	Total	% of Total IOFs	Accurate	% Accurate	Inaccurate	A	B
Type 1	15	22.1	15	100	0	0	0
Type 2	39	57.4	36	92.3	3	3	0
Type 3	0	0	0	0	0	0	0
Type 4	14	20.6	12	85.7	2	2	0
Type 5	0	0	0	0	0	0	0
All	68	-	63	92.6	5	5	0
	Non-Redundant IOFs						
	Total	% of Total IOFs	Accurate	% Accurate	Inaccurate	A	B
Type 1	14	25	14	100	0	0	0
Type 2	31	55.4	28	90.4	3	3	0
Type 3	0	0	0	0	0	0	0
Type 4	10	17.9	8	80	2	2	0
Type 5	0	0	0	0	0	0	0
Type 2/4	1	1.8	1	100	0	0	0
All	56	-	51	91.1	5	5	0

Note. In the bottom half of the table, the full title of the third column is "% of Total Non-Redundant IOFs".

Table 8

Compliance with Non-Redundant IOFs Provided by Instructor or TA

	Total	N/A	Full	% Full	Partial	None
Type 1	14	1	8	61.5	0	5
Type 2	31	3	16	57.1	0	12
Type 3	0	0	0	0	0	0
Type 4	10	3	6	60	0	1
Type 5	0	0	0	0	0	0
Type 2/4	1	0	1	100	0	0
All	56	7	31	63.3	0	18

Note. Values in the fifth column indicate percentage of IOFs which produced full compliance, after discounting any N/A (not applicable) IOFs.