Running Head: RULE FOLLOWING IN CAPSI

A Study of Rules Designed to Increase Peer-Review Accuracy In A Computer Aided Personalized System of Instruction Course

By

Kirsten Marianne Wirth

A Thesis

Submitted to the Faculty of Graduate Studies

In Partial Fulfillment of the Requirements for the Degree of

Master of Arts

Department of Psychology

University of Manitoba

Winnipeg, Manitoba

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION

A Study of Rules Designed to Increase Peer-Review Accuracy In A Computer Aided Personalized System of Instruction Course

BY

Kirsten Marianne Wirth

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

Of

MASTER OF ARTS

Table of Contents

Abstract

This study used archived data from a computer-aided personalized system of instruction (CAPSI) course that was taught at the University of Manitoba in 2001 and 2002. Students provided responses to unit test questions through the system, and these responses were archived through the computer program for future analyses. In the fall 2002 semester, the students were presented with a set of rules in the general course manual. A question about the rules was added to a study unit dealing with course procedures. Students were required to demonstrate mastery of this unit before proceeding in the course, increasing the likelihood that the rules were rehearsed. The set of rules specifically targeted peer-review accuracy and feedback, as well as test-writer responses such as appeals, with a mild contingency for not following the rules. The rules were also restated in an individualized email to each student half way through the course. Archived records of the students' responses were assessed by independent raters, and reliability checks were performed. The percentage of peer-review accuracy increased by 18% from the no-rule-semester (in 2001) to the rule-semester (in 2002), and the differences were statistically significant. The percentage of substantive feedback provided by peer-reviewers did not increase; however, in the rule-semester, substantive feedback occurred 23% more in the rule-semester when restudies were provided. Conversely, in the rule-semester, substantive feedback occurred 23% less in the rule-semester when passes were provided. There was a 39% increase in the percentage of restudies provided in the rule-semester compared to the no-rule semester. Neither the increase in substantive feedback or restudies assigned was statistically significant. It was hypothesized that as a function of increasing peer-review accuracy, peer-reviewers would score better on final examinations; however, there was no effect on final exam scores. This research has implications for developing ways to better train CAPSI peer-reviewers and test-writers, and for gaining rule-governed control over students' peer-reviewing behaviours.

A Study of Rules Designed to Increase Peer-Review Accuracy In A Computer Aided Personalized System of Instruction Course

Introduction

The Problem

The Computer-Aided Personalized System of Instruction (CAPSI) nicely amalgamates computer-based instruction (CBI) with the personalized system of instruction (PSI; Keller, 1968) as an important alternative to traditional teaching methods. The CAPSI system was developed and implemented at the University of Manitoba by Joseph Pear in 1983 (Kinsner & Pear, 1988). A feature unique to CAPSI is its archiving database. All transactions and correspondence within the system (e.g., emails, unit tests, and feedback) were archived and are accessible for analysis.

Students within a CAPSI-taught course engage in several functions. They construct answers to study questions, which are the learning objectives for the course. They serve as reviewers to other students once they have demonstrated mastery of a particular unit. Finally, they serve as research participants as their data-based responses are available for analysis at any time. There has previously been some analysis of peer-review accuracy (Martin, Pear, & Martin, 2002; Wirth, Gawryluk, Crone-Todd, & Pear, 2002) which showed that there was only a 5% difference between the accuracy of feedback statements given by instructors/teaching assistants or peer reviewers. However, accuracy of assigning passes or restudies by peer-reviewers on unit tests was not as high as would be desired.

There is a need to study methods to increase peer-review accuracy in terms of (a) assigning passes and restudies; and (b) giving accurate feedback statements provided to students by peer-reviewers. This thesis examined the use of rules to improve these two behaviours of students' when peer-reviewing.

Rules & Rule-Governed Behaviour

A rule is a statement that a specific behaviour that occurs in response to a specific antecedent or discriminative stimulus (S^D) will lead to a specific consequence. An S^D, simply put, is a cue that a response will lead to reinforcement (Martin & Pear, 2003). Rule-governed behaviour is defined as a response that is controlled by the statement of a three-term contingency: antecedent-behaviour-consequence (Martin & Pear, 2003; Malott, 1992). For example, suppose that a student has an examination scheduled for the next week (antecedent). The student could use a rule, "If I start studying now (behaviour), I will get an 'A' on the final exam (consequence)." The student then proceeds to review his or her notes and text in response to the statement of the rule, i.e., rule-governed behaviour has occurred. Skinner (1969) conjectured that the use of a rule intensifies the S^Ds and their corresponding reinforcing value. Rules can serve as warning stimuli and advice, and provide for us an "awareness" of the importance of performing a specific behaviour. An example of this is the student described previously who began to study in order to perform well, or get an 'A,' on the final exam. Formulating the rule drew attention to the date of the impending exam (S^D) and to the studying behaviour that must occur in order to do well (reinforcement). The student was "aware" that performing the behaviour was necessary for an 'A' on the exam. It also serves as a warning stimulus for the student in that if he or she did not study, he or she may do poorly on the exam.

Martin & Pear (2003) emphasized that rules are important when describing contingencies that involve delayed consequences. The student who is preparing for the exam is working for a consequence that will not occur for some time. He or she must study, write the exam, and then wait until it has been marked to receive the mark. The authors also indicated that rules should be used when a speedy behaviour change is needed. If the exam is going to occur in a week, in order for the student to perform well on the exam, it is likely that he or she will be required to change their behaviour immediately.

How does a rule successfully exert control over behaviour when the consequence is delayed? For example, in education a behaviour or series of behaviours are often required long before the reinforcement can be delivered. In order to receive a degree a student must take a specific number of credits and successfully complete each course over a period of years. There are three explanations as to why rule-governed behaviour occurs when reinforcement or punishment does not immediately follow the behaviour. First, over the course of any individual's life there have been many rules presented to them and many rules followed. Following rules has become a skill due to the individual's reinforcement history (Baum, 1992; Malott, 1989; Martin & Pear, 2003). Following rules has been reinforced in the past; whereas, not following rules has likely been punished. In this sense an individual can be said to be under control of the rule. Martin and Pear (2003) suggested that a history of punishment for not meeting deadlines (i.e., not following a rule) can intensify the aversiveness of the deadline as a warning stimulus. It is only after the rule is followed that the aversiveness of the warning stimulus is reduced; i.e., strengthening of rule following occurs through negative reinforcement or escape conditioning. Therefore, rules have become a generalized stimulus class (Malott, 1989) and rule following has become a robust generalized response class (Schmitt, 2001). That is, there is variability in the type of rules and in responding to various rules.

Another explanation for the success of rule control is that individuals may make positive self-statements immediately after following a rule (Martin & Pear, 2003). In contrast, individuals may make punishing statements to themselves that pertain to not following the rule. The aversiveness of the punishing self-statements could be relieved by complying with the rule, which is reinforcing (Malott, 1989). Consider again the rule, "If I start studying now, I will get an 'A' on the final exam." When the student begins to prepare their review notes, they may make statements to themselves such as "I feel so much better now that I have gotten started," or "Good for me! I did so much work today!" If they were not following the rule, they may make statements to themselves, such as "I am going to fail this exam if I don't start studying." This serves as a warning stimulus to begin studying immediately.

A third explanation involves social control. The individual who is expected to follow the rule may receive immediate consequences from others for following or not following the rule. Attention from these other individuals, such as parents, in themselves are likely generalized conditioned reinforcers or punishers. Their attention is paired with many backup reinforcers or punishers such as physical affection or punishment, praise or reprimands, and so on (Martin & Pear, 2003). Therefore, they provide contingencies for following or not following the rule. To this end, compliance or non-compliance with a rule becomes a conditioned reinforcer or punisher because it is paired with back-up reinforcers or punishers. This is consistent with Malott's (1989) proposal that the rule is an establishing operation because the statement of the rule momentarily increases the reinforcing or punishing value of the contingencies for following or not following the rule.

There are five general conditions that increase the effectiveness of a rule. Malott (1989) and Martin and Pear (2003) suggested that the first condition required to ensure a rule's effectiveness is to specify the behaviour so it may be easily measured. In the studying example, the behaviour is not explicitly defined. In order to make the rule more effective, "studying" could be changed to "answering a certain number of study questions relevant to the material." A second condition requires the antecedent to be described (Martin & Pear, 2003). In other words, the individual who is expected to follow the rule must know when to follow it.

The third and fourth conditions involve the description of the consequence. The consequence should be specific, but also should be probable and sizeable to the individual (Malott, 1989; Malott, 1992; Martin & Pear, 2003). A common example used to describe the importance of the probability of the consequence involves the use of a helmet when riding a bicycle. Many people neglect to wear a helmet because the odds of falling off the bike are slim even though the consequence could be quite sizeable. An example used to describe the importance of making the consequence sizeable could be using a swear jar. Each time an individual swears, he or she is required to put 25 cents into a jar. However, 25 cents is not a sizeable consequence and the behaviour is likely to continue to occur. It is therefore important that the consequence be both probable and sizeable in order to ensure rule-governance.

The fifth condition which increases the likelihood that a rule will be effective is providing a deadline for the behaviour to occur (Malott, 1992; Martin & Pear, 2003). For example, in a completely self-paced course, many students would never reach the final exam. Providing a deadline, i.e., an examination date; exerts control over the students to prepare for the exam for a specific date. Also, as illustrated earlier, a history of aversiveness for not meeting deadlines may elicit anxiety. The anxiety is only relieved as a function of following the rule, or meeting the deadline. This tends to render deadlines extremely effective.

Personalized System of Instruction

Personalized System of Instruction (PSI) was designed and first implemented by Fred S. Keller and his colleagues. PSI has a number of basic tenets. The course content is divided into small study units which require mastery. The student must master one unit before he or she can move onto the next. However, there is no penalty for repeating a unit, regardless of how many times it must be repeated. (Of course, if a student were repeating the unit an inordinate number of times, remedial action would be taken.) It is a self-pacing method. The student is encouraged to study and write unit tests or perform laboratory experiments at their own pace (Keller, 1968).

There are a number of assistants or helpers in the course. In addition to the instructor, there are teaching assistants and student reviewers. A student reviewer is one who has already mastered the material. As the title "student reviewer" indicates, he or she reviews certain work of other students who are currently taking the course. The student reviewer then assigns a pass or restudy on the unit. If a student feels his or her work was incorrectly assigned a restudy, he or she may appeal to a teaching assistant or instructor. This is all under the responsibility of the teaching assistants. The instructor is then free to work on the study material in terms of content, structure, and final assessment of the students' progress in the course. There is also a final examination in which all material covered in the course is represented (Keller, 1968; Martin & Pear, 2003).

Kulik, Kulik, and Bangert-Drowns (1990) performed a meta-analysis of studies that focused on two mastery-learning teaching methods: Learning for Mastery (LFM) and PSI, versus traditionally taught methods. Traditionally taught methods generally consisted of a series of lectures and a final examination. That is, all students would be taught in the same manner and receive the same lessons, regardless of aptitude. The authors found that 93% of the studies cited positive effects on the final examinations administered in the mastery learning courses. A further 70% of the studies indicated that there was a significant difference in the positive learning effects of the mastery versus traditionally taught students. There were also stronger effects on examinations that were locally developed versus standardized. Furthermore, in general, the mastery taught groups received more feedback than the traditional groups. When this occurred, there were larger differences between the learning effects than when less feedback was provided. The PSI average result was statistically significant versus the traditional method, and students' examination grades were increased by 0.48 standard deviations. When the variability in final examination grades was analyzed, 71% of the studies reported less variability in the mastery learning groups. In fact, there was 77% more variation in the scores of the traditionally taught groups.

Regarding students' attitudes towards the mastery learning methods, Kulik et al. (1990) report that 89% of the studies that measured this variable found that students had more positive attitudes towards the teaching method in the mastery learning groups. An additional 86% of the studies that measured the students' attitudes towards the subject matter reported more positive effects in the mastery learning versus traditional groups. An interesting finding is that although the mastery method effects are not consistent for all individual students, it was shown that low aptitude students could benefit more from the mastery methods than high aptitude students. Some studies incorporated follow up tests to their courses. In these cases the mastery taught students showed higher retention of the course material. In summary, Kulik et al. (1990) found that PSI methods were consistently more effective than traditional teaching methods.

Note, however, there have been problems associated with using PSI in the traditional classroom. In a tongue-in-cheek article, Keller (1985) feigned his regret in promoting the PSI method. It appeared to be too cumbersome to implement. It required a large number of people to administer and it wound up being very expensive. The question remained. How could PSI be utilized in the least expensive and least cumbersome method possible?

Computer-Based Instruction

With the rise in computer technology, many educators began to utilize computer-based instruction (CBI) methods. Kulik and Kulik (1991) performed a meta-analysis on CBI methods. Interestingly, the findings of this analysis were quite consistent with their findings from the meta-analysis on mastery learning methods. When comparing final examination grades, students' using the CBI methods had statistically significant scores 0.3 standard deviations higher than those in the traditional methods. Some studies reported students' attitudes, and generally the students' attitudes were positive towards the CBI methods. Most importantly, the CBI methods were found to considerably decrease the time required for teaching. I will now turn to a method for combining the effective components of PSI and CBI.

The Computer-Aided Personalized System of Instruction

The Computer-Aided Personalized System of Instruction (CAPSI) is an innovative teaching tool used since 1983 at the University of Manitoba to automate the tenets of PSI (Kinsner & Pear, 1988). The CAPSI system overcomes the cumbersome nature of PSI. The CAPSI email system allows students to write and submit tests to the instructor, marker, or peer-reviewers for marking. This function negates the necessity for a physical classroom which, in itself, overcomes part of the problem associated with implementing PSI in a classroom (Pear & Kinsner, 1988; Kinsner & Pear, 1988).

General Course Procedure. As explained in Kinsner & Pear (1988), students in the CAPSI courses receive a large pool of study questions derived from the relevant course textbook from which they can prepare answers. These study questions are the learning objectives for the particular course. A small sample from the pool is randomly assigned by CAPSI when a student requests a unit test. Once a student is able to correctly answer the sample of the learning objectives, he or she is said to have mastered that unit of the course. There is a unique feature with CAPSI that is generally not offered in CBI. Generally in CBI, mastery is shown by answering multiple-choice or fill-in-the-blank questions. Through CAPSI, mastery is reached through short, essay-style responses (Pear & Novak, 1996). For example, in a Behaviour Modification Principles course, a test question might be "What is a positive reinforcer? What is the principle of positive reinforcement? What is operant behaviour? In what way is positive reinforcement like gravity?" An answer would then include the definitions of positive reinforcement, a reinforcer, and operant behaviour; and also, a brief explanation of how reinforcement and gravity are alike.

Kritch and Bostow (1998) conducted a study in which students were exposed to questions that required various responses. Students were split into groups where some were required to construct high-density responses during the presentation of the material, and some were required to emit low-density responses during the presentation of the material. A high-density response was defined as filling-in-the-blank after every frame presented, and a low-density response was defined as either filling-in-the-blank after every second frame presented or having no opportunity to respond. Students who were required to construct high-density responses scored much higher on a post-test than those who constructed low-density responses, regardless of aptitude. Furthermore, students who were required to construct high-density responses reported more positive attitudes towards the tasks than those who constructed low-density responses. This research suggests that essay style responses (i.e., high-density responses) would likely be followed by higher scores on post-tests. This provides support for the essay style responses required for mastery in the CAPSI courses, and indicates that the type of responses required may assist students to score higher on post-tests, such as final exams.

The CAPSI method incorporates peer-reviewers (Pear & Novak, 1996). Students who have mastered a unit may then make themselves available to review that particular unit for another student, and provide feedback on that student's answers. This process allows the instructor and teaching assistants to be more available to the students as resources when help is needed.

The Act of Peer-Reviewing. Peer-reviewers assess the students' answers and assign a pass or a restudy. A pass is awarded when the peer reviewer determines that the student had mastered the material, and should continue with the next unit. Conversely, a restudy is awarded when the peer-reviewer determines the student had not mastered the material, and should take some time to re-study the material (Pear & Crone-Todd, 1999). Peer-reviewers not only provide feedback to the students in terms of assigning a pass or restudy, but they also have space to provide details as to why a pass or restudy was granted. Peer-reviewers are required to provide this feedback to the students within 24 hours or lose points toward their final grade. In addition, the system sends out the same test to two peer reviewers as a quality control measure. In order for the student to pass the test, both peer reviewers must assign passes (Kinser & Pear, 1988).

As noted in Pear and Novak (1996), there has been concern expressed regarding allowing students to review other students' tests. There are, however, many advantages to this. First, students are able to interact with classmates regarding the course material and possibly receive novel feedback. Second, students acting as reviewers have likely shared the same experiences as the students writing the tests, and can relate to difficulties with specific sections of the material. Third, the peer-reviewers may learn more about the course material and may extrapolate ideas and writing techniques to incorporate into future tests, papers, and exams (Pear & Novak, 1996; Saunders, 1992). Pear and Crone-Todd (1999) ran an analysis of a questionnaire students had filled out regarding the CAPSI method. Seventy-one percent of the students felt that the peer-review system definitely helped them learn the course material, 24% were neutral about the peer-review system, and 5% disagreed. In other words, the majority of students thought that peer-reviewing was a valuable part of the learning process.

Peer-Review Accuracy & Feedback. Above it is stated that the benefits of peer-reviewing far outweigh the disadvantages. A disadvantage of the peer-review system could be that the peer-reviewers might not provide the same quality or quantity of feedback that a teaching assistant or the instructor would, which is why there are two peer-reviewers assigned to each unit test (Pear & Novak, 1996). There are two separate issues to consider in regards to the quality and quantity of feedback. These are difficult to present separately as they are very closely tied together. First, are peer-reviewers providing the correct feedback to students in terms of a pass or restudy? Second, are peer-reviewers providing the correct feedback in terms of statements made?

Martin, Pear, and Martin (2002a) examined a sample of unit tests written in an undergraduate psychology course at the University of Manitoba. Two "expert" raters independently assessed the unit test answers as correct or incorrect, and reached a final agreement of correct or incorrect for each answer. The raters' assessment of each question was then directly compared with the feedback the peer-reviewers had assigned the student (i.e., pass or restudy) on each question to establish peer-review accuracy. When answers were incorrect, peer-reviewers inaccurately marked them as correct 67% of the time (these were false negatives). When answers were correct, peer-reviewers inaccurately marked them as incorrect 7% of the time (these were false positives). Sixty-seven percent false negatives is high; however, when the second peer reviewer per test is taken into account, the percentage of false negatives decreased to 46%. Although the total percentage of incorrectly assigning a pass or restudy by peer-reviewers amounted to 27%;

conversely, the total percentage of correctly assigning a pass or restudy by peer-reviewers was 73%.

Martin, Pear, and Martin (2002b) investigated the type and accuracy of feedback provided on unit test questions using the same data set described above. There were five types of feedback statements assessed by two "expert" raters: model, suggestion, example, question, and page reference. Following this, there were two types of errors that could be made: (a) feedback was based on an inaccurate interpretation of the student's answer; or, (b) feedback was inconsistent with the course material. Martin et al. (2002b) reported that feedback provided by the instructor, teaching assistant, and peer-reviewers was mostly models and suggestions, and was 87% accurate overall. They also reported that the instructor and teaching assistant provided 31% of the feedback assessed, and their feedback accuracy alone was 91%. When taking the instructor and teaching assistant out of the equation, it is possible to calculate the percentage of accurate feedback provided by peer-reviewers, which was 85%.

Discrepancy of Peer-Review Accuracies. In Martin et al. (2002a) peer-review accuracy of pass/restudy designation was 73%. However, as shown in Martin et al. (2002b) using the same dataset as Martin et al. (2002a), peer-review accuracy of feedback statements was 85%. It can be inferred that peer-reviewers were correctly identifying something wrong in the answers, but inaccurately passing the student's unit test anyway.

This finding was replicated in Wirth, Gawryluk, Crone-Todd, and Pear (2002). Peer-review accuracy was obtained by the same procedure as in Martin et al. (2002a). Peer-review feedback was defined as substantive, which was inclusive of the categories

Martin et al. (2002b) used; and non-substantive, where a general comment was provided. The percentage of substantive feedback given by peer-reviewers was compared point-by-point with both passes and restudies that the peer-reviewers assigned, and the passes and restudies that the raters designated. The result was consistent with Martin et al. (2002a, b). The percentage of substantive feedback was higher when passes were given by peer-reviewers than when restudies were given. However, the percentage of substantive feedback was lower when passes were designated by raters than when restudies were given. This provides support to the theory that when peer reviewers identify something wrong in a student's answers; they tend to assign passes regardless. It appears that peer-reviewers are hesitant to provide restudies to their fellow students, but still provide feedback on errors.

According to Pear and Crone-Todd (1999), the 'worst thing' that would happen to a student once receiving a restudy is they would (a) wait an hour before they could re-write the unit; and (b) appeal to the instructor by providing a cogent argument as to why they truly deserved a pass. As pointed out in Plett (2003), re-writing a test or undergoing the appeal process probably enhances the students' learning.

Statement of the Problem

Information regarding peer-review accuracy is available in Martin et al. (2002) and Wirth et al. (2002). The present study focused on increasing peer-review accuracy and feedback. This research replicated procedures used previously to assess peer-review accuracy and feedback. Furthermore, this research assessed a specific rule to determine if the rule was consistent with the guidelines for effectiveness, would it exert control over students' peer-review behaviours.

Method

Participants

The participants were students who had taken a CAPSI course in Behaviour Modification Principles (course # 17.244) at the University of Manitoba in the fall sessions of 2001 (n = 16) and 2002 (n = 24), and peer-reviewed at least 4 unit tests above unit 1[1].

Materials

The course materials consisted of the relevant text, a CAPSI general manual that outlined the system's procedures, and a course specific manual that contained the study questions or learning objectives for the course. In the fall of 2002, a rule was added to the general manual. The rule consisted of the following information: Researchers have been analyzing student data in terms of answers, peer-review feedback, and accuracy since CAPSI was first implemented. We have found many occasions on which a peer-reviewer assigned a pass to a student's test, even though an independent rater (i.e., researcher) determined the test should have received a restudy. In many of these instances the peer-reviewer gave substantive feedback, suggesting that he or she correctly identified something lacking in the answer. When peer-reviewing, if you feel that there is an area in which an answer is lacking, do not hesitate to assign a restudy, and you are encouraged to provide feedback to the student as to what is needed, or where he or she can find further information. Evidence suggests that peer-reviewing this way benefits you as well as the student, as you are reviewing the material as well as other students' answers. The instructor and teaching assistants periodically review instances of peer-reviewing, and

[1] Unit 1 tested students on the CAPSI system in order to familiarize them with the course method. Therefore, unit 1 tests were excluded as they had no relevance to course material.

peer-reviewers who regularly pass tests that show inadequate mastery will be asked to improve their level of scrutiny. (Martin, Crone-Todd, & Pear, 2002, p. 18).

A study question was added to the first unit in order to exert control over students' reading of the rule. The study question was as follows: Why is it important for peer-reviewers to carefully evaluate answers on unit tests and to assign a restudy if an answer doesn't demonstrate mastery? Furthermore, about half-way through the semester an e-mail was sent out to all students reiterating the rule. The e-mail was as follows: Hi (student's name), I'm writing to remind you of the guidelines for peer-reviewing, and to ask you to please: (a) assign restudies on any tests in which one or more answer(s) are not complete and correct, (b) explain as clearly as possible what is deficient in the incorrect answers, and (c) do so in a manner that is respectful and constructive. Peer-reviewing in this way provides the most benefit to you and the student writing the test. Keep in mind that speeding students through the unit test system when they have not displayed mastery does not help prepare them or you for the final exam, and it reduces the total number of peer-review points available in the course. The markers and I do review past peer-reviewing to confirm that peer-reviewers are providing an appropriate level of feedback. On the other hand, remember that answers are adequate if they satisfy the requirements of the question. If you think that you have been marked too strictly, please appeal the result by pressing the F1 key on the final comment screen of your test.

Thank you, (Instructor's name).

The rule was not presented in any shape or form in the fall semester of 2001. The rule specifically targeted peer-review accuracy and feedback; it encouraged students to assign restudies if the answer was incorrect, and to provide substantive feedback regarding the incorrectness of the answer (Martin, Crone-Todd, & Pear, 2002). The e-mail was sent to all students in the course. Each student received the same e-mail in terms of content, but the greeting was tailored to each individual (e.g., "Hi Susan").

The principle materials used for the data analysis were the data-based CAPSI files. The databases contained all information related to peer-review assessment of unit tests and feedback. As the comparison courses used different editions of the same text, the researchers used a 7th and 6th edition Behaviour Modification text (Martin & Pear, 2003; Martin & Pear, 1999) to analyze all student answers from the rule-semester and no-rule-semester, respectively. The computer programs Microsoft® Excel and SPSS were used to create graphs and perform all statistical analyses.

Procedure

Sample Selection. The students in the fall session of 2001 were not presented with a rule that targeted peer-review accuracy and feedback. This group is referred to as the no-rule-semester. The students in the fall session of 2002 were presented with the rule. This group is referred to as the rule-semester. A random sample of 25% of the overall questions peer-reviewed was taken from each peer-reviewer until there were 12 questions selected per peer-reviewer (i.e., 4 unit tests) for 2001 and 2002. The final sample for the current study consisted of 640 questions peer-reviewed; i.e., 640 instances of peer-reviewing.

Unavoidable Differences Between the Two Semesters. As indicated, the independent variable of interested with the presence or absence of the rule encouraging peer reviewers to provide restudies where appropriate. Unfortunately, there were two unavoidable differences between the two semesters: (1) different editions of the text were used; and (2) there were different numbers of study questions in the two semesters. As will be explained in the Discussion section, although undesirable, these unavoidable procedural differences probably do not seriously undermine the conclusions that can be drawn from this study.

Peer-Review Accuracy. In order to determine peer-review accuracy, it was necessary to assess whether the students provided complete and correct answers. Four graduate researchers and one undergraduate researcher (i.e., rater(s)) assessed each question in the sample marked by peer-reviewers. The answers were assessed as correct, mostly correct, and incorrect. If an answer was fully correct with respect to use of terminology and course content as per the course text and agreements between the researchers it was "correct". If one part of the answer was questionable in terms of an omission; however, fully illustrated with an example, it was rated as "mostly correct". For example, if the student provided a definition that was lacking, but illustrated each component of that definition in an example, the answer was rated as mostly correct. If the answer contained errors of omission or commission, used incorrect terminology, or did not rationally or cogently address the question it was "incorrect." Peer-review accuracy was ascertained by a point-to-point comparison between the peer-reviewer designation of pass or restudy with the rater designation.

Peer-Review Feedback. All feedback given by peer-reviewers was rated by two graduate researchers and one undergraduate researcher as substantive or non-substantive. Substantive feedback involved an explicit comment clearly related to the question, or the student's answer to the question, including the concepts or principles the question cued and textbook references. For example, if the question called for a definition and an example of positive reinforcement, and the peer-reviewer provided feedback including the term “positive reinforcement,” or “you can find the correct answer on page 29,” it was rated as substantive. In contrast, non-substantive feedback involved a general comment that had no specific relevance to the question, the student's answer to the question, the concepts or principles the question cued, or textbook references. For example, if the question called for a definition and an example of positive reinforcement, and the peer-reviewer provided feedback such as “good job,” or “nice answer,” the feedback was rated as non-substantive.

Exams. Correlations were performed between final exam scores and peer-review accuracy/feedback, and the no-rule-semester was compared to the rule-semester. All exam answers were independently rated by three undergraduate research assistants. Each exam question was broken down into its components, and rated out of a total score of 5 points. For example, “What is a positive reinforcer? What is the principle of positive reinforcement? What is operant behavior? In what way is positive reinforcement like gravity?” was one question on a CAPSI unit test or exam. This question was broken down into 4 components of 1.25 points each.

Inter-Rater Reliability. Inter-Rater Reliability (IRR, agreements / (agreements + disagreements) x 100%; Martin, Pear, & Martin, 2003) was calculated for a random selection of 25% of the data sample. IRR was calculated between raters assessing peer-review accuracy, substantive vs. non-substantive feedback, and final examinations. The procedure involved practice assessments for each type (i.e., peer-review accuracy, substantive vs. non-substantive feedback, and final examinations) until raters reached a minimum acceptable level of 80% agreement for 3 practice sets consecutively. During practice sessions, any IRR below 80% involved a discussion of what was to be required, and an agreement between raters as to what the correct assessment was. The mean IRR's for peer-review accuracy, substantive vs. non-substantive feedback, and final examinations were 81.67%, 98.33%, and 83.33%, respectively.

Comparisons of Peer-Review Accuracy and Feedback. The peer-review designation of pass or restudy was compared with the rater designation of pass or restudy per unit test question by calculating the percentage of true positives (both the rater and peer-reviewer assigned a restudy), false positives (the rater assigned a pass, and the peer-reviewer assigned a restudy), true negatives (both the rater and peer-reviewer assigned a pass), and false negatives (the rater assigned a restudy, and the peer-reviewer assigned a pass; see Table 1).

Table 1

Type of Agreement and Disagreement Between Peer-Reviewers and Raters

		Rater Decision	
		Incorrect	Correct
Peer-Review Decision	Incorrect	True Positives	False Positives
	Correct	False Negatives	True Negatives

The percentage of passes and restudies designated by both peer-reviewers and raters was examined. The rule-semester data was also broken down into two subsections: pre-email rule presentation, and post-email rule presentation, and the means were compared using a paired-samples *t*-test. A matrix correlation was performed between: (a) percentage of restudies assigned; (b) percentage of peer-review accuracy; (c) percentage of substantive feedback given by peer-reviewers; and, (d) final examination scores. These were calculated for both the rule-semester and the no-rule-semester. Means for restudies assigned, peer-review accuracy, substantive feedback given, and final examination scores were compared between the 2001 and 2002 groups using independent *t*-tests. To control for an effect due to the increase in number of units from the no-rule to the rule-semester, a correlation was performed between the total number of questions assessed and the percentage of peer-review accuracy. Furthermore, the total number of questions peer-reviewers assessed was compared between 2001 and 2002 groups using independent *t*-tests. Due to the differences in sample sizes, *t*-tests were used to compare the no-rule-semester to the rule-semester as it is robust even when the distributions are not normal. Individual cumulative records of restudies assigned, peer-review accuracy, and substantive feedback provided were graphed.

Results

Figure 1 shows that true positives (both the rater and peer-reviewer assigned a restudy) and true negatives (both the rater and peer-reviewer assigned a pass) increased from 45% in the no-rule-semester to 64% in the rule-semester. Figure 2 shows that false positives (the rater assigned a pass, and the peer-reviewer assigned a restudy) and false negatives (the rater assigned a restudy, and the peer-reviewer assigned a pass) decreased from 55% in the no-rule-semester to 36% in the rule-semester. The differences were statistically significant ($t = -3.258$, $p < .005$; $t = 3.258$, $p < .005$; respectively). A potential confound to the effect shown here might be that the number of unit tests increased from

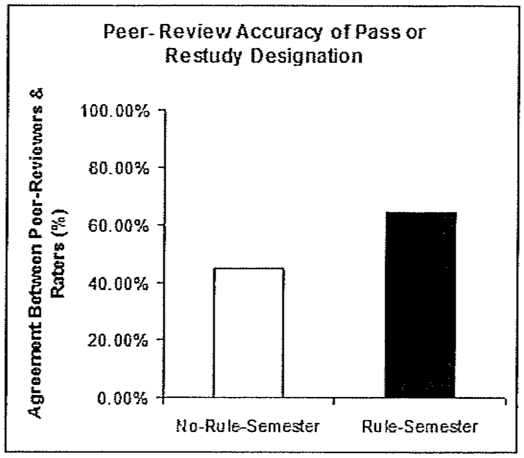


Figure 1. Percentage of true positives and and negatives (i.e., agreement between peer-reviewers and raters) in the no-rule-semester vs. the rule-semester.

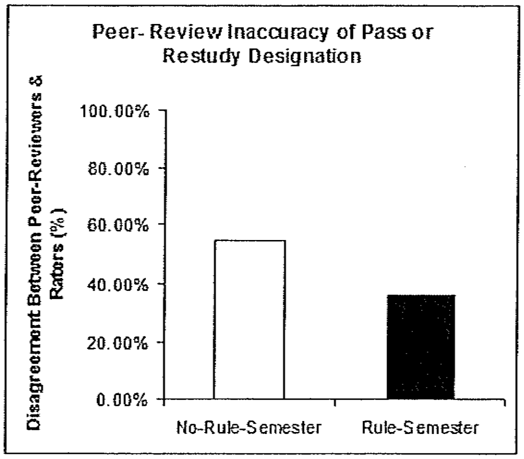


Figure 2. Percentage of false positives and negatives (i.e., disagreement between peer-reviewers and raters) in the no-rule-semester vs. the rule-semester.

10 in the rule-semester to 15 in the no-rule semester. However, if this were to be the case, there should be a significant, positive relationship between the percentage of peer-review accuracy and number of questions assessed in the rule-semester. The relationship was in fact, non-significant, small, and negative (r = -0.17, p > .05). Furthermore, if the increase in unit tests had an effect on peer-review behaviour, there should have been significant differences between the number of unit tests peer-reviewed between the no-rule-semester and the rule-semester. Again, there was no significant difference (t = -0.41, p > .05).

A point-by-point comparison of substantive feedback provided by peer-reviewers with passes and restudies assigned, versus a point-by-point comparison of substantive feedback provided by peer-reviewers with passes and restudies as designated by raters revealed a change from the no-rule-semester to the rule- semester (Figure 3). In the no-rule-semester, peer-reviewers provided more substantive feedback when they assigned passes than when they assigned restudies. However, the raters' designations of passes vs.

restudies suggest that peer-reviewers should have provided more substantive feedback with restudies than with passes.

In the rule-semester, there was more substantive feedback provided when peer-reviewers assigned restudies than when they assigned passes, as should have occurred according to the raters' designations. The peer-reviewers substantive feedback with restudies assigned increased from 29% in the no-rule-semester to 52% in the rule-semester, and the differences were not statistically significant (t = -0.92, p > .05). There was also no statistically significant change in substantive feedback provided from the no-rule-semester to the rule-semester (t = 0.03, p > .05). However, there was a large correlation between the percentage of substantive feedback given with the percentage of restudies assigned by peer-reviewers for both the no-rule and rule-semesters (r = 0.74, p < .001; r = .80, p <.001; respectively).

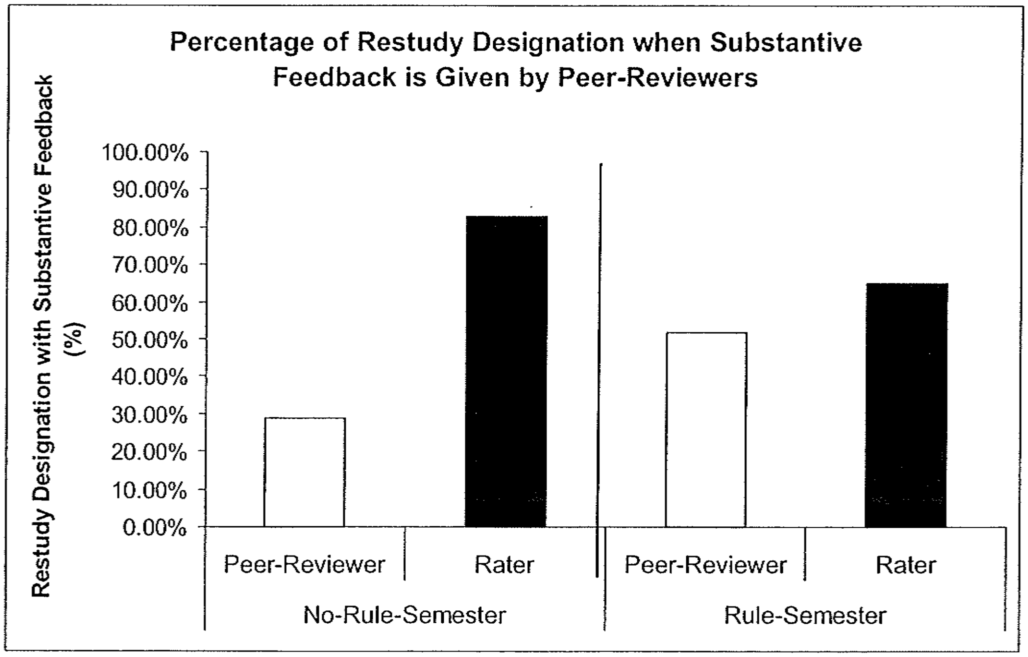


Figure 3. Point-by-point comparison of percentages of peer-review substantive feedback with peer-reviewer and rater assigned passes vs. restudies for the no-rule- and rule-semesters.

Although the differences were not statistically significant, there was a small, negative correlation between the percentage of substantive feedback provided and peer-review review accuracy in the no-rule-semester (r = -0.21, p > .05), vs. a moderate, positive correlation between the percentage of substantive feedback provided and peer-review accuracy in the rule-semester (r = 0.37, p < .10).

Figure 4 shows that the percentage of restudies assigned by peer-reviewers increased from 6% in the no-rule-semester to 45% in the rule-semester, and the inset shows an increase in restudies assigned pre- and post-email rule presentation from 45% to 55% in the rule-semester. Neither difference was statistically significant (t = -1.21, p > .05; t = -0.20, p > .05; respectively).

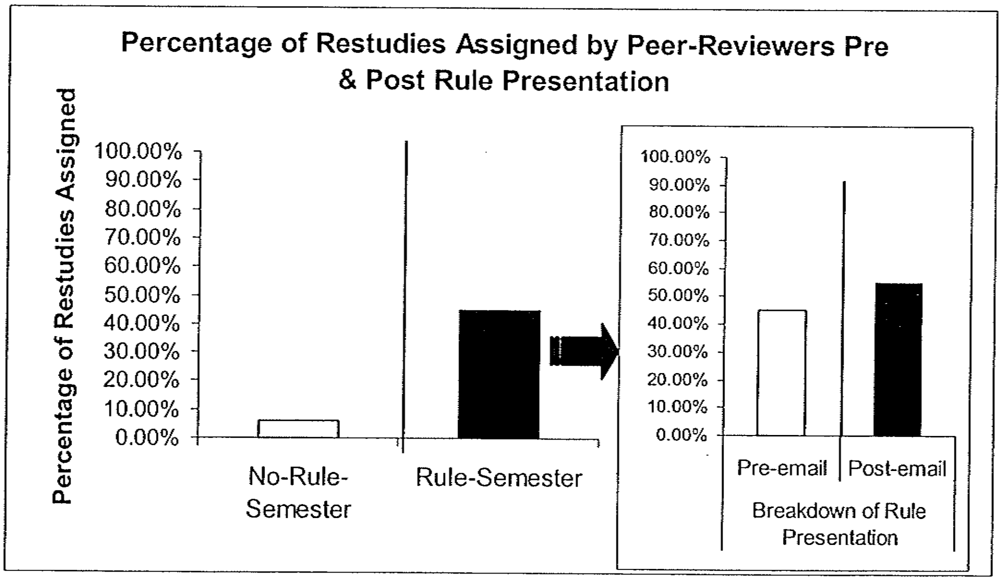


Figure 4. Percentage of restudies assigned by peer-reviewers in the no-rule- semester vs. the rule-semester, and a breakdown of restudies assigned by peer-reviewers in the rule-semester before and after the rule was presented in an e-mail.

The correlation between the percentage of peer-review accuracy and restudies assigned was moderate to small for both the no-rule- and the rule-semester (r = .37, p >

.05; r = .17, p >.05; respectively). The correlation between the percentage of substantive feedback provided and final examination scores was moderate for both the no-rule- and rule-semester (r = .41, p > .05, r = .32, p > .05; respectively). The correlation between final examination scores and percentage of restudies were small for both the no-rule- and rule- semester (r = .172, p > .05; r = .09, p > .05; respectively), and the correlation between final examination scores and percentage of peer-review accuracy were also small for both the no-rule- and rule- semester (r = -0.30, p > .05; r = .14, p > .05; respectively). There was no statistically significant difference between final examination scores in the no-rule- and rule-semester (t = -0.40, p > .05). Cumulative records of peer-reviewing behaviours illustrate the variability in peer-reviewing per individual for the no-rule and rule-semesters (Appendices A & B).

Discussion

The purpose of this study was to determine whether the presentation of a rule could improve peer-reviewer behaviour in a CAPSI-taught course. The rule appears to have been very successful in increasing true positives and true negatives (i.e., the accurate detection of incorrect and correct answers, respectively) by 19%, and in turn, decreasing false positives and false negatives (i.e., the inaccurate detection of incorrect and correct answers, respectively) by 19% as shown in Figures 1 and 2.

A potential confound to the effect shown here might be that the number of unit tests increased from 10 in the no-rule-semester to 15 in the rule-semester. The students in the rule-semester may have had to mark more tests, and therefore accuracy may have improved as a function of experience. However, if this were to be the case, there should have been a significant, positive relationship between the percentage of peer-review accuracy and number of questions assessed in the rule-semester. The relationship was non-significant, small, and negative. This means that the more questions peer-reviewers assessed, the less accurate they were. Based on anecdotal evidence, it is a more likely hypothesis that the students in the rule-semester may have had to mark more tests, and therefore did not improve as much as they could have because they may have become 'lazy' in their peer-reviewing as the semester progressed. Furthermore, if the increase in unit tests had an effect on peer-review behaviour, there should have been significant differences between the number of unit tests peer-reviewed between the no-rule-semester and rule-semester. Again, there was no significant difference between the number of questions assessed from the no-rule to the rule-semester.

A second possible confounding variable could have been that a new edition of the course text was used in the rule-semester. The content changed as a function of presenting updated research in some sections, and the order that the material was covered was changed. However, it is doubtful that this produced any effect on this study. Few study questions pertaining to the material changed as a result of the new edition, and as the courses were self-paced, peer-reviewers were subject to marking any of the unit tests at any particular time.

While there were no statistically significant differences in the amount of substantive feedback provided from the no-rule-semester to the rule-semester, it is evident from Figure 3 that there was a large change from the no-rule to rule-semester in the designation of restudies with substantive feedback given (i.e., 23%). It is plausible that statistical differences were difficult to extract from the data since the overall percentage of substantive feedback provided did not increase, but its 'placement' in terms of passes vs. restudies changed. In support of this hypothesis, the small correlation between the percentage of substantive feedback provided and peer-review accuracy changed from a negative relationship in the no-rule-semester (i.e., the more accurate the peer-reviewer was, the less substantive feedback he or she provided) to a small, positive relationship in the rule-semester (i.e., the more accurate the peer-reviewer was, the more substantive feedback he or she provided).

It is plausible that the significance test would neglect a small change because *t*-tests are generally used to detect large changes in a sample, whereas, if a small change did occur here, it may not be detectable. In other words, although the results do not appear to be statistically significant, Figure 3 illustrates that the rule may have been successful in reducing the discrepancy between the raters' assessment of passes or restudies with substantive feedback, and peer-reviewers' assessment of passes or restudies with substantive feedback. It was expected that as a result of presentation of the rule, there should have been a larger percentage of restudies with substantive feedback given than passes with substantive feedback to make the peer-review results more consistent with the rater results. This is apparent in the rule-semester as compared to the no-rule-semester in figure 3.

It was expected that peer-reviewers would have assigned more restudies than passes in the rule-semester, as a result of the rule targeting peer-review accuracy, as was illustrated in Figure 4. However, due to a lot of variability; e.g., just over a third of the peer-reviewers in the no-rule-semester never assigned restudies (Appendix A), and almost half of the peer-reviewers in the rule-semester never assigned restudies (Appendix B); there was no statistically significant difference between semesters. The case was the same for restudies provided in the rule-semester before and after the e-mail re-stating the rule was sent, as almost half of the peer-reviewers never assigned any restudies. It is likely that for the same reason that there were no statistically significant differences in restudies assigned per group, that there was a very small relationship between peer-review accuracy and restudies assigned.

It was hypothesized that students who provided more substantive feedback and who were more accurate in peer-reviewing would score higher on the final exams than those who were not. In addition, it was thought that the rule-semester students would score higher on the final exams than the students in the no-rule-semester. Although the rule did not target final examination behaviour, it was expected that as a function of peer-reviewing more accurately, and providing more substantive feedback that peer-reviewers would perform better on the final examinations. However, this was not the case. There was no effect on the final examination scores.

There are some aspects of the course procedure that produce variation in peer-review responding. Students obtained points towards their final grade in the course by peer-reviewing. That is, the more tests they peer-reviewed, the more points they got. Therefore, it is possible that some students did not provide substantive feedback because they did not fully read or understand the question, or the answers that other students provided. This could have increased the probability that they would pass students on tests where a restudy would have been beneficial. However, peer-reviewers received points for marking tests in both groups, so this would not have had a systematic effect on the rule.

It is important to note that a truly random sample was not used. Since it is the students who enroll in the course, they are in essence selecting the sample available for study. Students taking CAPSI courses may select them because they prefer to work more independently than other students.

Implications & Future Research

The evidence suggests that the rule was successful in terms of increasing peer-review accuracy. This indicates that there is a reliable way to effectively train students to more accurately assign passes or restudies when peer-reviewing. This illustrates how a rule can be used to gain effective control over students' peer-review behaviour. In addition, this study adds to the literature by demonstrating how providing a specific rule can increase peer-review accuracy in a computer-mediated PSI course. However, the rule was not as effective as anticipated. There was no statistically significant mean difference between the amount of substantive feedback provided in the no-rule-semester and the rule-semester. Future research should manipulate the rule provided in order to more specifically target peer-reviewer feedback as well as accuracy. Furthermore, future research could investigate using rules to improve students' final examination behaviour.

References

Baum, W.M. (1992). For parsimony's sake: Comments on Malott's "A theory of rule-governed behavior and organizational management." *Journal of Organizational Behavior Management, 12*(2), 81-84.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

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

Keller, F.S. (1985). Lightning strikes twice. *Teaching of Psychology, 12*(1), 4-8.

Kinsner, W., & Pear, J.J. (1988). Computer-Aided Personalized System of Instruction for the Virtual Classroom. *Canadian Journal of Educational Communication, 17*(1), 21-36.

Kritch, K.M., & Bostow, D.E. (1998). Degree of constructed-response interaction in computer-based programmed instruction. *Journal of Applied Behavior Analysis, 31*(3), 387-398.

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.

Malott, R.W. (1989). The achievement of evasive goals: Control by rules describing indirect-acting contingencies. In S.C. Hayes (Ed.), *Rule-governed behaviour:*

Cognition, contingencies, and instructional control (pp. 269-322), New York: Plenum.

Malott, R.W. (1992). A theory of rule-governed behaviour and organizational behaviour management. *Journal of Organizational Behavior Management, 12*(2), 45-65.

Martin, G. L., & Pear, J. J. (1999). *Behaviour Modification: What It Is and How to Do It* (6th ed.). Upper Saddle River, NJ: Prentice-Hall.

Martin, G. L., & Pear, J. J. (2003). *Behaviour Modification: What It Is and How to Do It* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.

Martin, T.L., Crone-Todd, D.E., & Pear, J.J. (2002). *General manual: CAPSI-taught courses.* Winnipeg: University of Manitoba Bookstore.

Martin, T.L., Pear, J.J., & Martin, G.L. (2002a). Analysis of proctor marking accuracy in a personalized system of instruction course. *Journal of Applied Behavior Analysis, 35*(3), 309-312.

Martin, T.L., Pear, J.J., & Martin, G.L. (2002b). Feedback and its effectiveness in a computer-aided personalized system of instruction course. *Journal of Applied Behavior Analysis, 35*(3), 427-430.

Pear, J.J., & Crone-Todd, D.E. (1999). Personalized system of instruction in cyberspace. *Journal of Applied Behavior Analysis, 32*(2), 205-209.

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. *Computers in Teaching, 23*(2), 119-123.

Plett, R.C.S. (2003). *Rule governed behaviour and higher-order thinking.* Unpublished honours thesis, University of Manitoba, Manitoba.

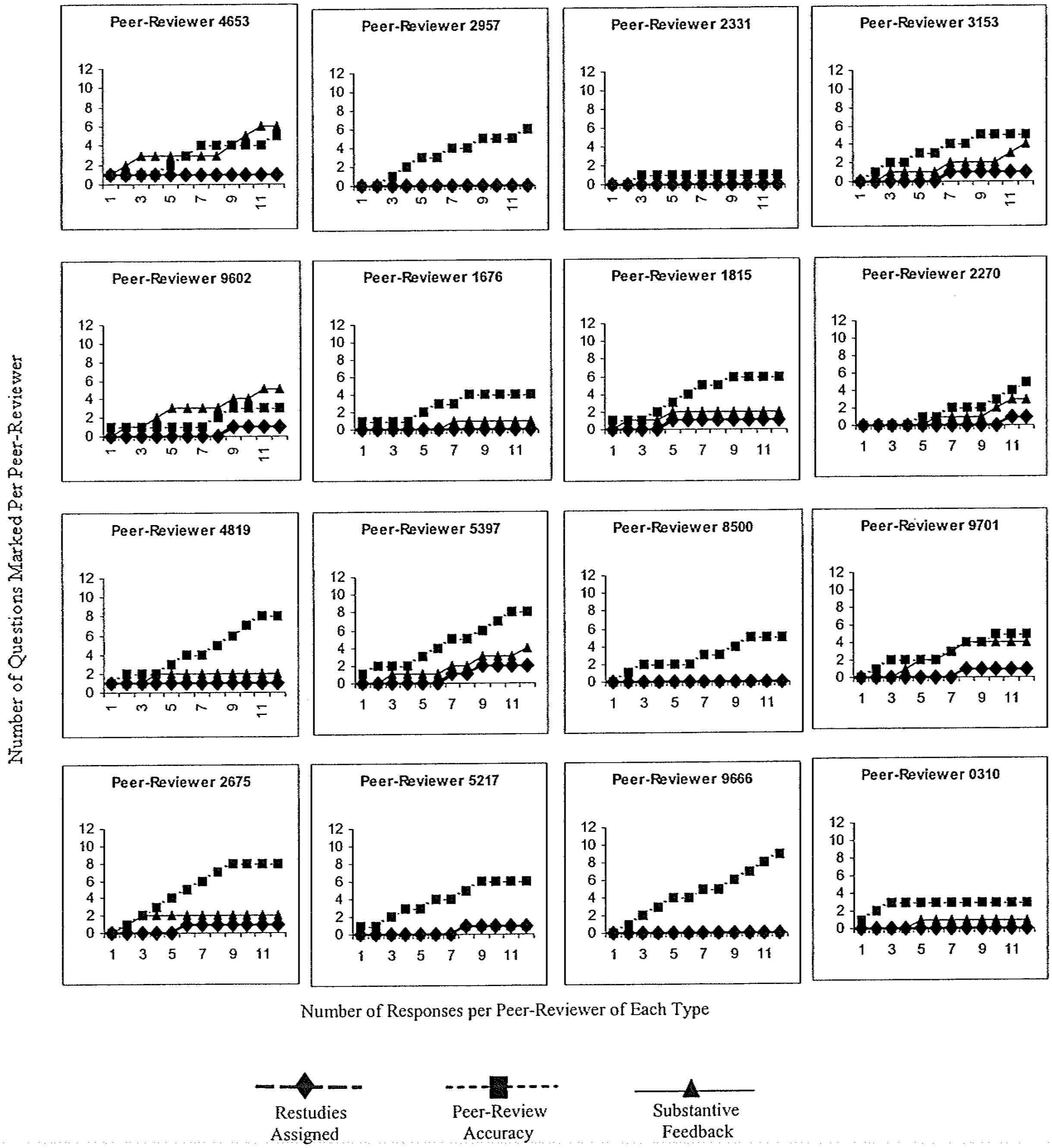
Schmitt, D.R. (2001). Delayed Rule Following. *The Behavior Analyst, 24*(2), 181-189.

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

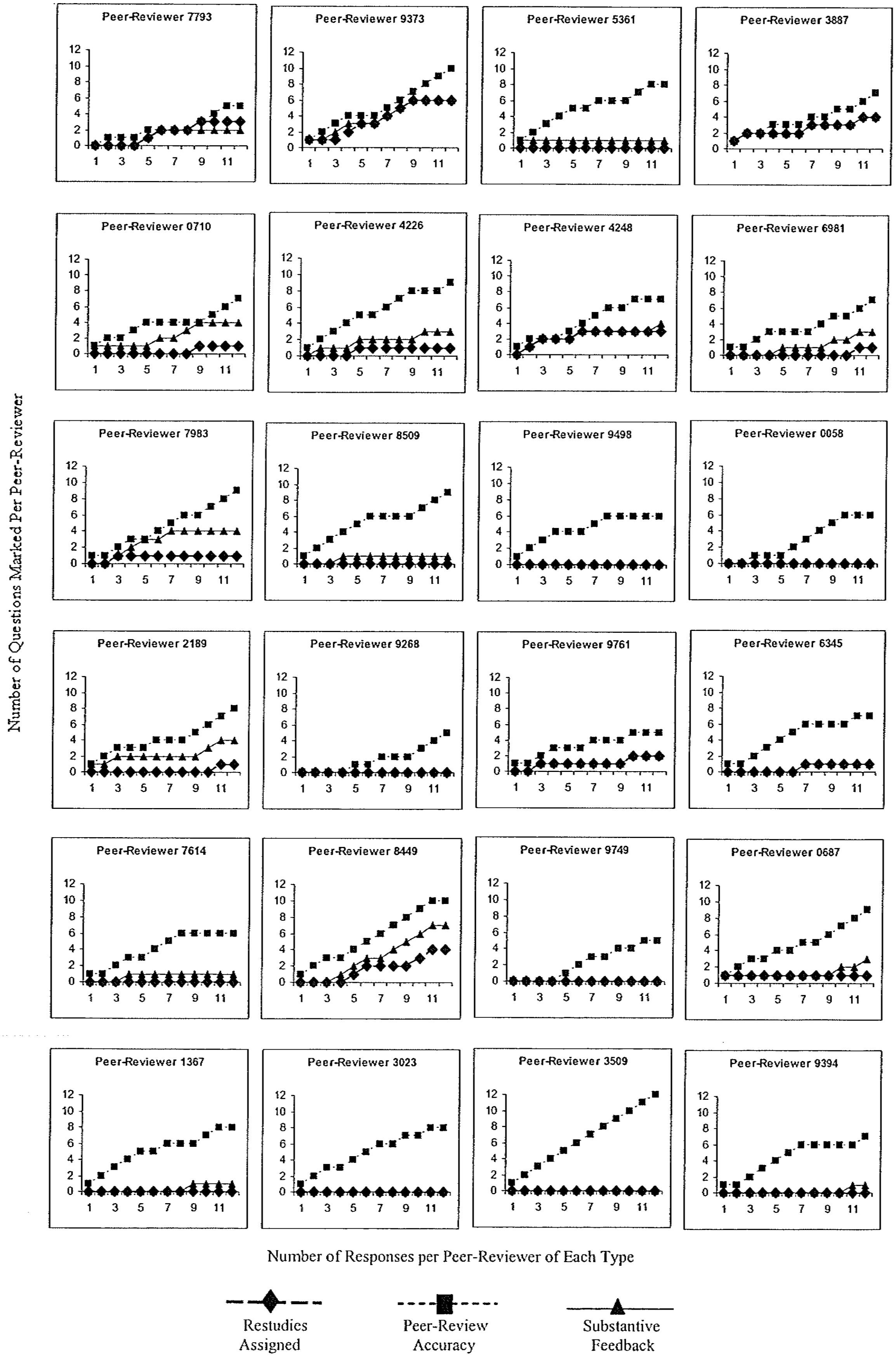
Saunders, D. (1992). Peer tutoring in higher education. Studies in Higher Education, 17(2), 211-219.

Wirth, K.M., Gawryluk, J., Crone-Todd, D.E., & Pear, J.J. (2002). *The Relationship Between Substantive Feedback and Final Examination Performance in CAPSI-Taught Undergraduate Courses.* Poster presented at the 28th annual international conference for the Association for Behavior Analysis, Toronto, ON, May, 2002.

Appendix A – Individual Cumulative Record of Responses (Restudies assigned, peer-review accuracy, & substantive feedback) for the No-Rule-Semester.



Appendix B – Individual Cumulative Record of Responses (Restudies assigned, peer-review accuracy, & substantive feedback) for the Rule-Semester.



Author Note

Kirsten M. Wirth, Department of Psychology, University of Manitoba.

This manuscript is submitted in partial fulfillment for the requirements of the degree of Master of Arts in Psychology under the supervision of Dr. Pear. I thank Joseph J. Pear for acting as my advisor and giving me the freedom to research various questions I have had regarding the CAPSI procedure and to work independently.

I thank Drs. Kathryn Young and Garry Martin for participating on my advisory committee, and Toby L. Martin for his advice and assistance on this project. I would also like to thank Maribel Abrencia for assistance with training for reliability checks, and Kerri L. Walters and Heather Simister-Firth for assessing data for reliability checks. Thanks to Gabriel Schnerch for his valuable contribution and assistance on this project regarding training and assessing data for reliability checks, and data analysis. Finally, thanks to Steven Wirth for his patience and support throughout this project, and for assistance with data analysis.

Correspondence concerning this manuscript should be addressed to Kirsten M. Wirth, Department of Psychology, University of Manitoba, Winnipeg, Manitoba, R3T 2N2. E-mail: umwirthk@cc.umanitoba.ca.