Reciprocal Peer Tutoring: An Embedded Assessment Technique to Improve Student Learning and Achievement

William T. Mickelson
Georgette Yetter
Michael Lemberger
Scott Hovater
Rene Ayers

University of Nebraska - Lincoln

Presented at the annual meeting of the
American Education Research Association
Chicago, IL
April 2003

Direct all correspondence to:

William T. Mickelson, PhD
Educational Psychology Department
31 Teachers College Hall
University of Nebraska – Lincoln
Lincoln, NE 68588-0345

PH: (402) 472-1196
Email: wmickelson2@unl.edu

ABSTRACT

The importance of facilitating study and practice materials that are consistent with graded assessments and instructional objectives is well known, if not commonly used, in educational practice. Reciprocal Peer Tutoring (RPT) is a collaborative approach that embeds assessment in a formalized learning process to facilitate student involvement with course content and improve achievement. Students engaging in RPT are paired and given explicit instruction on how to construct multiple-choice questions for different types of statistical content knowledge, akin to Bloom's Taxonomy. During an RPT activity, each student of a dyad is independently responsible for synthesizing course content and constructing practice multiple-choice test questions, complete with answers, based on the course curriculum. Each dyad then administers practice tests to each other prior to formal class examinations. Upon completion of the practice exams, partners score each other's work and alternate roles as tutors and tutees to assess each other's performance, give feedback on missed items, and discuss individual questions and course content. In this dual role as tutor and tutee, students benefit through the preparation and instruction in which tutors engage, as well as from the instruction that tutees receive. This study examines the impact of reciprocal peer tutoring (RPT) on student achievement over six sections of an introductory statistics course. A comparison of RTP treatment relative to a control, accounting for instructor, showed an effect of RPT treatment at the time of the last examination of the semester. This finding is tempered by additional analyses into the effectiveness of the RPT treatment. Student achievement relative to increasing levels of cognitive complexity of exam items showed mixed results. Furthermore, a comprehensive analysis of the student work within RPT treatment revealed students having difficulties implementing the intervention.

INTRODUCTION

Peer collaboration includes a family of cooperative learning approaches that entail guided and formalized peer interactions to promote and facilitate academic achievement. Overall, such approaches have been effective across wide varieties of tasks and student populations (Cohen, 1997). Such activities as summarizing information, critiquing the work of peers, giving and receiving feedback, correcting errors, questioning thought processes and justifications, and explaining rationales have been especially beneficial in promoting academic achievement (Greenwood, Carter, & Kamps, 1990). Several models explain the effectiveness of these activities. For example, developmental researchers emphasize the facilitating effect of interactive processes such as exploring opposing ideas and mutual modeling between individuals at similar developmental levels in scaffolding the emergence of new understandings and cognitive growth (Damon, 1984; Vygotsky, 1978), while social-cognitive theories emphasize the beneficial effect of social interaction in stimulating active processing and reorganization of ideas (Slavin, 1992). This study examines the efficacy of embedding assessment in the ongoing discourse of an introductory statistics class through a technique called Reciprocal Peer Tutoring (RPT), and addresses student achievement, the role of cognitive complexity, and students' ability to engage in the process of RPT.

What is Reciprocal Peer Tutoring?

Peer tutoring is one collaborative approach where pairs of students interact to assist each other's academic achievement by one student adopting the role of tutor and the other the role of tutee. Peer tutoring has been well validated for promoting the development of low-level skills, such as spelling, math, and reading (e.g., Fuchs, Fuchs,

Phillips, Hamlett, & Karns, 1995; Greenwood, Delquadri, & Hall, 1989). This approach also has been used with college students to develop higher-order skills such as reading comprehension (Dansereau, 1987) and understanding of statistical concepts (Keeler & Steinhorst, 1994). Interestingly, students who provide the assistance seem to experience greater gains than those who receive the tutorial help (Webb, 1991; Webb, 1992; Yager, Johnson, & Johnson, 1985).

Recognizing the benefits gained by students from acting as tutors, Reciprocal Peer Tutoring (RPT), formalizes a process enabling both students in a peer tutoring pair to participate and experience the role of tutor (Pigott, Fantuzzo, & Clement, 1986; Wolfe, Fantuzzo, & Wolfe, 1986; Wolfe, Fantuzzo, & Wolter, 1984). In this dual role as both tutor and tutee, students benefit through the preparation and instruction in which tutors engage, as well as from the instruction that tutees receive. First developed for enhancing the academic achievement of elementary school children (Pigott, Fantuzzo, & Clement, 1986), RPT has been modified for college-age students (Fantuzzo, Dimeff, & Fox, 1989). Typically, college students engaging in RPT are paired, with each student independently responsible for constructing practice multiple-choice tests based on the course curriculum. Each dyad then administers these tests to each other prior to the formal class examinations. Upon completion of the practice exams, the partners score the practice exams and alternate roles as tutors and tutees, using the results of the practice tests as the context for providing explanatory feedback on missed items.

Previous Experience with RPT

Initial studies exploring RPT with students enrolled in undergraduate-level psychology courses found RPT to be a promising study strategy. Students using RPT to

prepare for course examinations demonstrated higher achievement compared with students who prepared and received practice test questions working independently (Fantuzzo, Dimeff, & Fox, 1989). The results of a follow-up study suggested that both peer interaction and the use of structured study materials, such as the use of multiple-choice practice questions, led to higher achievement (Fantuzzo, Riggio, Connelly, & Dimeff, 1989). A subsequent investigation where students prepared more frequently for examinations (four vs. three unit exams) found that while participants assigned to interact with peers demonstrated higher achievement than those working independently, students engaging in RPT (structured peer interactions) outperformed those in all the other conditions (Fantuzzo, Connelly, & Dimeff, 1991).

Recent studies, however, have not supported the early findings in favor of RPT with college students. In separate investigations, both Griffin and Griffin (1998) and Rittschof and Griffin (2001) found no achievement advantage in favor of RPT compared with a non-interaction control condition in either undergraduate-level psychology or graduate-level research methods courses. These investigators proposed that the inconsistent findings in the literature for college-level RPT might be related to differences in student populations and course content among the various studies. In a different investigation, Griffin and Griffin (1997) examined RPT using students enrolled in a graduate-level course in research methods. In this work, the investigators not only examined the effect of RPT on overall performance, but they also looked at the relationship between the effect of RPT and the cognitive complexity of the course content. Using Bloom's taxonomy (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956), Griffin and Griffin (1997) classified questions on the posttest into those requiring lower-

order cognitive skills and those requiring higher-order cognitive skills. They examined separately the effects of RPT on achievement for posttest items of different skill levels. Griffin and Griffin found no advantage for RPT compared with no-treatment controls, either in terms of overall achievement or for exam items of higher or lower complexity.

RPT Connected to Statistics Education

Building on the question of whether or not RPT is effective in the context of relatively complex cognitive tasks, this study extends the work of Griffin and Griffin (1997) to the area of introductory statistics for undergraduate students. David Moore (1997), past president of the American Statistical Association and a major proponent of statistics education reform, states "Statistics has its own substance, its own distinctive concepts and modes of reasoning. These should be the heart of the teaching of statistics at any level of mathematical sophistication." The substance, concepts, and modes of reasoning Moore refers to are more than mere formulas and computations. Rather, Moore points to the necessity of understanding the language of statistics, how the components of this language (i.e. concepts) are interrelated, and how an in depth conceptual understanding is important in the application of statistical reasoning. The challenges of teaching the abstract concepts, inferential reasoning, and art of data analysis typically associated with the introductory statistics class are well documented. Arguably, the introductory statistics context offers a rigorous setting for evaluating the effectiveness of RPT to enhance student achievement and assessing the difficulties of implementing RPT in cognitively complex situations.

METHOD

Participants consisted of 180 undergraduate students ranging in age from 18 to 52 years (mean age=21.01, sd=4.116) who were enrolled in one of six sections (approximately 27 students each) of an undergraduate introductory statistics course offered by an Educational Psychology Department at a large Midwestern, Research I University. Of those who indicated their gender, 50 were males and 118 were females. There were 16 freshman, 48 sophomores, 47 juniors, and 33 seniors with one student obtaining a second bachelor's degree. The vast majority, 88%, of participants indicated that they had no prior experiences with statistics or research methods.

At the beginning of the spring semester of 2001, students were told by a research assistant, independent of the class, that their section was to be involved with a statistics education research project focused on student achievement as measured by examination scores. This was not unexpected by students, as the Educational Psychology Department has a long-standing student research participation requirement for this introductory statistics class. Students were informed that different sections of this same class would receive different types of assignments and that all instructors for this course were involved with this research project. Furthermore, it was stated that different assignments in different sections were not unusual as instructors for this course are typically given autonomy over their sections of the course. Assignments were constructed such that the expected amount of outside of class work would be equivalent across sections. To participate in the study, students were simply asked to complete the course, finishing all assignments and examinations. Students agreeing to participate gave the researchers

permission to photocopy their work and use their examination scores. Participation in this study was voluntary.

Instructors

Three doctoral level graduate student instructors were each responsible for teaching two sections of the course with supervision from a faculty advisor. The instructors, two male and one female, had differing amounts of experience prior to teaching this course. Instructor A was a male in a Counseling Psychology program who had completed 3 years of coursework, including classes in Introductory Statistics, Experimental Design and ANOVA, Correlation and Regression, Measurement, and an Introduction to Education Research. In addition, instructor A also had one semester of teaching experience with the undergraduate introductory statistics course. Instructor B was female in a Research Methods program specializing in assessment. She had over 6 courses in advanced statistics and measurement, but was in her first semester of teaching the introductory course. Instructor C had recently started a doctoral program in Educational Administration, had previously completed courses in Introductory to Statistics and Education Research Methods, and was concurrently enrolled in an intermediate statistics course on Experimental Design/Analysis of variance. Instructor C started teaching two weeks into the semester, due to the illness of the originally assigned instructor. Finally, Instructor C had no prior experience teaching introductory statistics, but had several years of experience teaching writing/composition, communications, and debate/public speaking and received excellent recommendations from faculty.

Measures and Data Sources

The primary dependent variable for this study was student achievement as defined performance on four in-class examinations. Each examination consisted of 25 multiple-choice questions with the specific test items selected from the instructor's manual that accompanies the introductory statistics text "Basic Practice of Statistics" by David Moore (1995). The items comprising each examination were decided upon by teacher consensus and selected to ensure content validity of the examinations relative to course objectives and content. The tests were secure and the same exam questions administered to both the treatment and control groups.

Additional data consisted of the multiple-choice items constructed by each student as part of the RPT treatment intervention and a end of semester survey on the effectiveness of the RPT treatment. The student developed items were collected at the end of each review session prior to the second, third and final examinations. The 3,180 items constructed by students became the primary data for a content analysis on student involvement with course content through item writing. Finally, during the last week of class, an end of semester class survey was administered to all the participants in the RPT treatment groups to obtain data on students' perceived value of RPT for an introductory statistics course. Students filled out the survey anonymously.

General Procedure

Instructors were randomly assigned to two sections. Each instructor taught one section using RPT and the other section without RPT as a control. Each instructor's section with the largest enrollment was assigned the RPT treatment. The sections each instructor taught were scheduled one immediately following the other. The assignment

of sections and treatment occurred three days prior to the beginning of the semester. Although it was not possible to randomly assign students to sections, no prior knowledge regarding instructor or teaching method was available to the students prior to the first day of class. All of the sections were equal in terms of % female, age, GPA, and class standing. Course instructors meet weekly with a faculty supervisor. The purpose of these meetings was to coordinate the course curriculum and to facilitate teaching and learning, both for the instructors and the undergraduate students. In addition, these sessions enabled the instructors to interact with each other and share ideas, experiences and obstacles, as well as plan examinations and coordinate the reciprocal peer tutoring treatment.

Experimental Design

In this quasi-experimental design with pre-test and multiple post-test measures, the experimental treatment factor (RPT versus control) was fully crossed with the three instructors participating in the study in order to examine the efficacy of RPT (Shadish, Cook, & Campbell, 2001). The instructor variable, while fully confounded with all aspects of a unique section of the introductory statistics course, is not of experimental interest and as such is considered a block (Kirk, 1995). Four examinations were given during the semester. Reciprocal peer tutoring intervention activities were not initiated prior to the first examination. Here, the first exam score acts as a pre-test baseline score. The experimental RPT intervention occurred prior to the second, third, and final examinations so that the scores from these three examinations are post treatment observations.

The sections assigned to RPT treatment conditions experienced the following intervention. Initially, students were introduced to the notion that they could construct multiple choice items as a means of studying for an examination. A short manual discussing how to construct multiple choice test questions was given to each student as a reference, and each RPT class took part in a discussion on item writing. Before each exam, except the first baseline exam, students in the RPT group were given an assignment to formulate 10 multiple choice questions on topics that would be covered in the upcoming examination and bring those questions, with answers, to class. During class time students were randomly paired and given time to complete the RPT activity. During this activity the pairs of students exchanged questions, answered their partners' questions, scored each other's paper as if they were a test or quiz, and had the opportunity to discuss questions that were answered incorrectly or that were questionably written. After discussion, students were allowed to ask questions of the instructor. The students' multiple choice questions were collected, graded as completed or not completed, and treated like any other class assignment. As such, failure to complete any of the RPT assignments did have a negative consequence on students' grades.

The control group did not receive the RPT item writing assignment or associated class time. The control group did receive a typical comprehensive review session prior to each examination that was led by the instructor. The review summarized the material recently covered and the instructor entertained students' questions. Homework assignments were augmented with additional problems so that the expected amount of time that students would work outside of class was equivalent.

Analysis of Student Developed Items

The content of the multiple choice questions constructed by the students in the RPT treatment group were examined and rated by the instructors. To develop classification categories and inter-rater guidelines for deciding between categories, the 100 items used in the four examinations were initially evaluated and categorized. Six categories were easily agreed upon by consensus. These categories, consistent with Bloom's taxonomy, were: basic definition, recognition, table application, calculation, conceptual and synthesis.

Students affiliated with the RPT treatment developed over 3000 multiple-choice items during this study. The items were divided into four roughly equal sets without regard to section or instructor. Each set of items was photocopied twice. Two different sets of items were assigned to each of the four instructor/raters, such that no two raters had the same two sets of items to rate. In other words, the first rater might have item sets 1 & 2, while the second rater had item sets 2 & 3. This resulted in each item being rated twice by separate raters, with each rater responsible for rating only half the items. The student developed items were rated relative to the scale previously developed using the 100 test items.

Inter-rater reliability was evaluated by percent of agreement between the four raters as well as Person Product Moment correlations. Discordant ratings were discussed among the raters until a consensus rating was determined. The student constructed items were then analyzed by rating category and compared to the ratings of the actual test items and student performance to gain a sense of the impact of the RPT intervention beyond overall test scores.

RESULTS

The first analysis component examines student achievement as measured by performance on multiple choice items administered during examinations common to all instructors and conditions. A subset of this analysis groups examination items by cognitive category to evaluate achievement relative to the complexity of the items. The next analysis component takes an in-depth look at the items developed by students during the RPT intervention to ascertain the effectiveness of the intervention. Finally, the results of students' perceptions of RPT through self-report are given in the last analysis component.

Student Achievement

Achievement was measured by the proportion of correct items out of the 25 items on an examination. Four examinations over the course of the semester and commonly administered across instructors and intervention type. Table 1 presents descriptive statistics detailing the number of students in each section, means, standard deviations, the lower (LB) and upper (UB) bounds of a 95% confidence intervals on mean achievement, and effect sizes (Cohen, 1988) between treatment and control groups for each teacher. The statistical conclusion validity of any hypothesis test conducted on this suspect given the unequal sample sizes, lack of random assignment of subjects to treatments, potential variance heterogeneity, and potential lack of independence due to the nature of the intervention (Shadish, Cook, & Campbell, 2001). Accordingly, and consistent with APA guidelines, the confidence intervals and effect sizes are used to interpret of the results. Effect sizes were individually calculated for each instructor and examination by subtracting the mean of the control group from the mean of the RPT treatment group,

then dividing by the pooled standard deviation (Lipsey, 1990, pp. 32). Effect sizes greater than 0.8 are considered large, effect sizes less than 0.2 are considered small, with effect sizes in the middle being considered moderate (Cohen, 1988).

Examination 1 was administered prior to any RPT intervention and is considered a baseline measure. The RPT intervention commenced prior to examination 2, as such exams 2, 3 and 4 are observations after treatment. Focusing attention on examination 1, the instructors have effect sizes of 0.52, 0.34, and 0.13, respectively. These small to moderate effect sizes at baseline suggest that there are differences between the RPT treatment and control groups attributable to non-treatment factors, such as student self selection for particular class times and prior exposure to basic statistical ideas. As the study progresses and we examine the effects of the RPT intervention over exams 2, 3, and 4, a common pattern emerges. Either the differences between the RPT and control groups diminish (Instructors A and C), or stay constant (Instructor B), across exams 2 and 3, relative to the baseline effect size. By the final test, however, there is a resurgence in the difference between mean test scores with the RPT treatment group outperforming the control group, with effect sizes increasing between 0.30 and 0.71 standard deviations. This pattern to the observed effect sizes is consistent with the notion that participants need some practice time, or experience, implementing RPT before the effect of treatment is observable.

In an attempt to understand the pattern of observed effect sizes beyond this initial interpretation, we turn our attention to the types of items that comprise the examinations. Table 2 presents the distribution items across the four examinations. The six categories (definition, recognition, application, calculation, conceptual and synthesis) were

developed and defined by consensus of all raters, who were also the instructors of this study. Each item from the four examinations was evaluated by all three raters/instructors and categorized by consensus. During the rating of the items, all three raters commented on the ease of categorizing the items. In Table 2, the categories are listed from left to right in order of cognitive difficulty and are loosely based on Bloom's Taxonomy. The distributional discrepancies between exams are due to construct validity issues in the development of the individual examinations. Unfortunately, no attempt was made to distribute the items evenly over the six categories for each test.

As can be seen in Table 2, there was a predominance of definition, calculation and conceptual items (84% overall) on all four exams with exams 1 and 4 having similar distributions (definition/conceptual) and exams two and three also having similar distributions (predominately calculation). As such, this leads to a second possible interpretation of the observed effect sizes, namely that the composition of the examinations is responsible for the observed pattern. This interpretation is intriguing and consistent with the nature of the course content in an undergraduate introduction to statistics course. Here, the initial units to the class cover basic ideas and formulas somewhat familiar, if not formally introduced, to most students. Difficult and foreign concepts, like sampling distributions, and logic, like hypothesis testing, fall in the middle toward the middle of the semester, with application concluding the semester.

The analysis now turns to student achievement performance on groups of examination items by cognitive category to evaluate achievement relative to the cognitive complexity of the items. The outcome measure again is proportion of items correct separated by category of test item. Table 3 presents descriptive statistics and effect size

results for the categories of Definition, Calculation, and Conceptual, all other categories not having sufficient numbers of items, across all examinations. What is evident in the patterns of these effect sizes is that there are not consistent patterns across teachers, nor are there consistent patterns from test to test. The majority of the effect sizes would be considered small with many of them close to zero or negative. Only instructor A has a large effect size for definitional items in exam 4. Only instructor B has moderate effect sizes for calculation items in exams 2 and 3, but in exam 4 the effect size for calculation items drops to 0.06. For the conceptual items, instructor B has effect sizes that consistently stay in the moderate range indicating no effect of treatment relative to the baseline measure, while instructor A has a moderately large effect size for exam 4. To address and explain the variability in the effect size results across categories, the analysis turns to the student generated multiple choice items

Analysis of Student Generated Multiple Choice Items During RPT Intervention

During the RPT intervention students generated 3,062 multiple-choice practice test items. Four raters read the student generated items. The items were divided equally among the raters such that exactly two independent raters evaluated each item. The same rubric developed from the examination items was used in rating the student items. In addition to the sheer magnitude of the rating task, the effort was also hampered by the quality of the students' generated items. The simple percent of agreement on the ratings of the items was 65.7%. Table 4 presents initial inter-rater reliability statistics, as measured by Pearson Product Moment Correlations. As can be seen, the reliability statistics are all in the range of 0.50 and attest to the difficulty of evaluating the student generated items. All four raters, during project meetings, commented on the difference

between rating the exam items and the student generated items, and the difficulty associated with the later.

There were a number of difficulties associated with the student generated items. First, 506 of the student written items ended up being non-valid items either because they were not multiple choice items or the items did not make sense relative to course content. An informal qualitative analysis of the student generated items after the initial reliability analysis pointed to student difficulty developing items for the definition and conceptual categories. These types of items lead to the majority of discordant ratings. Out of the 823 student items ultimately rated as conceptual, 452 of them initially were discordant. The primary explanation unanimously put forth by the raters was that the difficulty in rating the student developed items stemmed from "lack of clarity." Lack of clarity was evident across all item categories, but was most evident in the definition/conceptual items. Here, the typical problem stemmed from students attempting to write definitional questions, ostensibly because definition is cognitively easier, when the actual topic of the item was clearly conceptual.

For the purposes of continuing this assessment, it was to come to agreement on all of the items developed by students. The discordant items that exhibited the definition/conceptual problem discussed above were categorized as conceptual. This decision places the students' items in the best possible light. With the merging of the definition/conceptual discordant items, the percent of agreement among the four raters improved to 80.5%. The remaining discordant items were then discussed and evaluated one by one and rated by consensus into one of the six categories.

The consensus rating on the student developed items permit a comparison of student work and perspective on preparation for examinations relative to the cognitive difficulty of the exams themselves. Table 5 presents this comparison in terms of percentages of items within each category to see if the students were writing items as similar to those found on the exams. First, students wrote fewer calculation, conceptual and synthesis items across the three examinations associated with RPT treatment intervention. At the same time, students wrote more definition, recognition, and non-valid items. The decision to place all of the discordant definition/conceptual items into the conceptual category, however, may dramatically distort this result in this category. Therefore, the actual number of conceptual items may be much lower and the number of definitional items may be higher. Comparing the two distributions, it is clear that the students as a whole were not writing similar items as given on the exams. The students tended to write items of lower cognitive skill level than the actual exam items. This fact may hinder the overall effect of RPT on test achievement. A further hindrance to finding an effect due to RPT treatment is the number of non-valid items.

The Effectiveness of RPT from Student Self Reports

During the last day of class, all students in each of the three treatment groups were administered a six item survey. This anonymous, self-report survey sought data to evaluate: a) students' satisfaction with the RPT treatment intervention; b) whether or not students preferred the RPT intervention over other experiences they have had in other classes; and, c) how much time students spent on actually writing the multiple-choice items as part of the intervention. Over 60% of the students reported that they had a better understanding of the material after writing the items and felt the peer review sessions

were beneficial to their learning. While a majority of students felt satisfied with the RPT intervention, only 35% preferred the RPT intervention to other possible experiences, with the most common alternative cited as lecture by the instructor only. This may suggest that students may prefer an authority figure presenting factual information instead of having to sufficiently internalize the statistical content to be able to write multiple-choice test items themselves.

Finally, the last survey item asked how much time they actually spent on writing the multiple-choice items. Given four choices of (1) less than one hour, (2) 1-2 hours, (3) 2-4 hours, and (4) over four hours, almost all (96.9%) claimed to have only spent less than two hours and 52.3% of those claimed to have used less than one hour. This lack of time spent on writing the multiple-choice items may have weakened the effectiveness of the RPT intervention.

DISCUSSION

The importance of using study and practice materials relevant to instructional objectives and assessments is well known. Collaboration between student peers through well-structured activities and study strategies like Reciprocal Peer Tutoring explicitly incorporate this notion into instructional activities by providing students with practice opportunities similar to the demands placed on them by actual course examinations. Because of this direct connection between student work and assessment, the collaborative processing of course material in such a context endows RPT with the potential for being a beneficial teaching and learning tool for college instructors and students. Explicitly organizing students' study material, as was done in this study, by giving the students guidance in generating practice questions that matched the format of actual exam

questions, provided potentially valuable cues that facilitated the formation of associations within the students' knowledge network (McKee & Witt, 1990).

In addition to evaluating overall statistics achievement as a function of an RPT intervention, the present inquiry expanded on previous investigations of college-level RPT studies attempting to 'look inside the black box' of reciprocal peer tutoring in two ways. First, it examined student achievement at different levels of cognitive complexity; and second, it comprehensively performed a detailed evaluation of student-generated multiple choice practice items. These findings regarding the difficulty of implementing RPT for cognitively complex tasks concur with prior research (Griffin & Griffin, 1997), and elaborate on possible reasons for this difficulty, offering insight into possible improvements of the RPT approach.

In this study, students in the RPT treatment condition were instructed on how to construct multiple choice items consistent with the classroom assessments. Through the use of effect sizes, the RPT groups were compared to control groups, and achievement differences were examined across four examinations. Higher achievement of the RPT treatment compared with no treatment was displayed on the fourth examination. Lack of difference between RPT and control in the first three exams appears to indicate that in complex content areas, on average, students needed time to develop skills and practice with the intervention before the effect of treatment is observed.

Looking at the practice items generated by students, however, it is clear that the RPT experience is highly variable among the students. Whether motivation, statistical content knowledge, mechanics of writing items, or understanding of a taxonomy of knowledge is the causal agent involved, a large proportion of students in this study had

difficulty writing challenging practice items, much less valid ones. Our conjecture is that in cognitively complex situations, the students' ability and motivation to write good practice items has a direct bearing on the effectiveness of the RPT treatment. Furthermore, this is the primary reason why little difference is observed between the RTP treatment and control groups in terms of performance on test items of higher cognitive complexity.

IMPLICATIONS

Due to design limitations, this study cannot offer a definitive statement about the effectiveness, or lack thereof, of RPT. Still, these findings indicate ways in which the RPT approach can be improved for college students in introductory statistics courses. In cognitively complex domains, like statistics, it is not sufficient to give basic instruction in item writing, even with examples, and then to ask students to prepare multiple choice questions and act as both tutors and tutees prior to each examination. For RPT to be more effective with students in statistics, the RPT intervention needs to be modified from the form originally proposed by Fantuzzo et al (1989). For example, the act of item writing could be embedded in a classes' weekly or day to day activities for very short periods of time to give students additional experience writing items. This will allow students to see the process of item-writing and peer collaboration as an ongoing and important aspect of their learning. In addition, it is critical to make explicit knowledge taxonomies, such as Bloom's Taxonomy or a similar type of taxonomy of statistical knowledge, and give detailed instruction on how to develop items to assess higher order thinking and statistical reasoning skills. The students in this study clearly struggled with developing practice items targeted toward higher order statistical thinking and reasoning.

Making this aspect of testing and assessment clear has the potential to increase student awareness of what multiple choice tests attempt to do, to engage students in the substantive content of the course in a deeper manner, and to improve student achievement and engagement with the content of statistics.

REFERENCES

Bandura, A. (1989). Human agency in social cognitive theory. *American Psychologist,* (44), 1175-1184.

Bloom. B. S., Englehart, M., Furst, E., Hill, W., & Krathwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1, Cognitive domain*. New York: McKay.

Cohen, E. G. (1997). Restructuring the classroom: Conditions for productive small groups. In E. Dubinsky, D. Mathews, & B. E. Reynolds (Eds.), *Readings in cooperative learning for undergraduate mathematics* (pp. 135-156). Washington, DC: Mathematics Association of America.

Cohen, J., (1988). *Statistical Power Analysis for the Behavior Sciences, 2nd Ed.* Hillsdale: Earlbaum.

Damon, W. (1984). Peer education: The untapped potential. *Journal of Applied Developmental Psychology,* (5)*,* 331-343.

Dansereau, D. F. (1987). Transfer from cooperative to individual studying. *Journal of Reading,* (), 614-619.

Dansereau, D. F. (1988). Cooperative learning strategies. In C. E. Weinstein, E. T. Goetz, & P. A. Alexander (Eds.), *Learning and study strategies: Issues in assessment, instruction, and evaluation* (pp.103-120). San Diego: Academic Press.

Fantuzzo, J. W., Dimeff, L. A., & Fox, S. L. (1989). Reciprocal peer tutoring: A multimodal assessment of effectiveness with college students. *Teaching of Psychology,*(16), 133-135.

Fantuzzo, J. W., Riggio, R. E., Connelly, S., & Dimeff, L. A. (1989). Effects of reciprocal peer tutoring on academic achievement and psychological adjustment: A component analysis. *Journal of Educational Psychology,* (81), 173-177.

Fuchs L. S., Fuchs, D., Phillips, N. B., Hamlett, C. L., & Karns, K. (1995). Acquisition and transfer effects of classwide peer-assisted learning strategies in mathematics for students with varying learning histories. *School Psychology Review,* (24), 604-620.

Greenwood, C. R., Delquadri, J. C., & Hall, R. V. (1989). Longitudinal effects of classwide peer tutoring. *Journal of Educational Psychology,* (81), 371-383.

Griffin, B. W., & Griffin, M. M. (1997). The effects of reciprocal peer tutoring on graduate students' achievement, test anxiety, and academic self-efficacy. *The Journal of Experimental Education,* (65), 197-209.

Griffin, M. M., & Griffin, B. W. (1998). An investigation of the effects of reciprocal peer

tutoring on achievement, self-efficacy, and test anxiety. *Contemporary Educational Psychology,* (23)*,* 298-311.

Keeler, C. M., & Steinhorst, R. K. (1994). Cooperative learning in statistics. *Teaching Statistics,* (16), 81-84.

Kirk, R.E. (1995). *Experimental Design: Procedures for the Behavioral Sciences.* Pacific Grove: Brooks/Cole

Koch, L. C. (1992). Revisiting mathematics. *Journal of Developmental Education,* 16(1), 12-18.

Lipsey, M.W., (1990). *Design Sensitivity: Statistical Power for Experimental Research.* Newbury Park: Sage.

McKee, W. T., & Witt, J. C. (1990). Effective teaching: A review of instructional, and environmental variables. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology, 2nd ed.* (pp. 821-846). New York: Wiley.

Moore, D.S., (1997). New pedagogy and new content: The case of statistics. *International Statistical Review,* 65(2), 123-165.

O'Donnell, A. M., Dansereau, D. F., Hythecker, V. I., Larson, C. O., Rocklin, T., Lambiotte, J. G., & Young, M. D. (1986). The effects of monitoring on cooperative learning. *Journal of Experimental Education,* (54), 169-173.

O'Donnell, A. M., Dansereau, D. F., Rocklin, T., Hythecker, V. I., Young, M. D., Hall, R.
H., Skaggs, L. P., & Lambiotte, J. G. (1988). Promoting functional literacy through cooperative learning. *Journal of Reading Behavior,* (20), 339-356.

Pigott, H. E., Fantuzzo, J. W., & Clement, P. W. (1986). The effects of reciprocal peer tutoring and group contingencies on the academic performance of elementary school children. *Journal of Applied Behavior Analysis,* (19), 93-98.

Riggio, R. E., Fantuzzo, J. W., Connelly, S., & Dimeff, L. A. (1991). Reciprocal peer tutoring: A classroom strategy for promoting academic and social integration in undergraduate students. *Journal of Social Behavior and Personality,* (1), 387-396.

Riggio, R. E., Fantuzzo, J. W., Connelly, S., & Dimeff, L. A. (1991). Reciprocal peer tutoring: A classroom strategy for promoting academic and social integration in undergraduate students. *Journal of Social Behavior and Personality,* (6), 387-396.

Rittschof, K. A., & Griffin, B. W. (2001). Reciprocal peer tutoring: Re-examining the value of a cooperative learning technique to college students and instructors. *Educational Psychology,* (21), 313-331.

Shadish W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* New York: Houghton-Mifflin.

Slavin, R. E. (1992). When and why does cooperative learning increase achievement? Theoretical and empirical perspectives. In R. Hertz-Lazarowitz and N. Miller (Eds.), *Interation in cooperative groups: The theoretical anatomy of group learning* (pp. 145-173). New York: Cambridge University Press.

Webb, N. M. (1991). Task-related verbal interaction and mathematics learning in small groups. *Journal for Research in Mathematics Education,* (5), 366-389.

Webb, N. M. (1992). Testing a theoretical model of student interaction and learning in small groups. In R. Hertz-Lazarowitz and N. Miller (Eds.), *Interaction in cooperative groups: The theoretical anatomy of group learning* (pp. 102-119). New York: Cambridge University Press.

Yager, S., Johnson, D. W., & Johnson, R. T. (1985). Oral discussion, group-to-individual transfer, and achievement in cooperative learning groups. *Journal of Educational Psychology,* (77), 60-66.

		Instructor A		Instructor B		Instructor C	
		RPT	Control	RPT	Control	RPT	Control
EXAM	n	35	21	41	18	30	23
	Mean	0.879	0.827	0.873	0.836	0.852	0.835
Exam 1	Std.	0.091	0.112	0.107	0.092	0.098	0.121
(baseline)	LB	0.844	0.782	0.841	0.787	0.815	0.792
	UB	0.913	0.871	0.905	0.884	0.889	0.877
	d	0.52		0.34		0.13	
	Mean	0.830	0.803	0.782	0.713	0.833	0.854
Exam 2	Std.	0.130	0.147	0.139	0.108	0.110	0.121
	LB	0.787	0.747	0.743	.0654	0.787	0.801
	UB	0.872	0.857	0.822	0.773	0.880	0.907
	d	0.22		0.46		-0.22	
	Mean	0.789	0.792	0.763	0.702	0.804	0.805
Exam 3	Std.	0.146	0.128	0.185	0.171	0.109	0.140
	LB	0.738	0.727	0.716	0.632	0.750	0.743
	UB	0.839	0.857	0.809	0.772	0.858	0.867
	D	-0.03		0.35		-0.01	
	Mean	0.846	0.787	0.795	0.711	0.851	0.819
Exam 4	Std.	0.093	0.108	0.139	0.164	0.086	0.123
	LB	0.806	0.735	0.758	0.656	0.808	0.770
	UB	0.885	0.838	0.832	0.767	0.894	0.868
	d	0.68		0.68		0.36	

Table 1: Descriptive Statistics, Confidence Intervals and Cohen Effect Sizes Across Examinations by Instructor and Treatment

	Definition	Recognition	Application	Calculation	Conceptual	Synthesis	Total
BASELINE	8 (32%)	3 (12%)	0	3 (12%)	11 (44%)	0	25 (100%)
EXAM 2	2 (8%)	0	2 (8%)	10 (40%)	6 (24%)	5 (20%)	25 (100%)
EXAM 3	4 (16%)	0	2 (8%)	10 (40%)	7 (28%)	2 (8%)	25 (100%)
EXAM 4	7 (28%)	1 (4%)	0	6 (24%)	10 (40%)	1 (4%)	25 (100%)
TOTAL	21 (21%)	4 (4%)	4 (4%)	29 (29%)	34 (34%)	8 (8%)	100(100%)

Table 2: Distribution of Test Items by Category for Each Exam

Test			Exam 1 (Baseline)				Exam 2			Exam 3			Exam 4		
Items	**Instr**	**Exp.**	**N**	**Mean**	**S**	**d**	**Mean**	**S**	**d**	**Mean**	**S**	**d**	**Mean**	**S**	**d**
	A	**RPT**	35	.896	.123	0.33	.771	.253	0.03	.764	.191	0.25	.808	.155	0.22
D		**Control**	21	.851	.156		.762	.301		.714	.213		.776	.124	
E	**B**	**RPT**	41	.857	.138	0.13	.781	.336	-0.08	.719	.269	0.24	.777	.150	0.86
F		**Control**	18	.840	.112		.806	.304		.653	.273		.635	.197	
N	**C**	**RPT**	30	.854	.119	-0.15	.750	.286	-0.11	.783	.205	0.15	.881	.136	0.34
		Control	23	.875	.164		.783	.295		.750	.238		.832	.153	
	A	**RPT**	35	.829	.234	0.47	.880	.126	0.29	.806	.201	-0.15	.938	.108	0.35
C		**Control**	21	.714	.264		.843	.125		.833	.146		.897	.134	
A	**B**	**RPT**	41	.805	.223	0.04	.802	.172	0.53	.812	.206	0.50	.882	.187	0.06
L		**Control**	18	.796	.259		.706	.204		.706	.226		.870	.177	
C	**C**	**RPT**	30	.767	.234	-0.00	.837	.119	-0.02	.870	.112	0.10	.906	.113	0.13
		Control	23	.768	.255		.839	.131		.856	.167		.891	.117	
C	**A**	**RPT**	35	.847	.107	0.25	.767	.222	0.25	.788	.231	-0.13	.789	.143	0.64
O		**Control**	21	.818	.135		.706	.268		.816	.170		.695	.153	
N	**B**	**RPT**	41	.905	.128	0.49	.720	.205	0.50	.735	.232	0.37	.732	.194	0.51
C		**Control**	18	.833	.185		.620	.188		.651	.215		.628	.224	
E	**C**	**RPT**	30	.846	.131	0.30	.839	.167	0.07	.771	.178	-0.06	.777	.155	0.15
P		**Control**	23	.802	.161		.826	.191		.783	.206		.752	.168	
T															

Table 3: Descriptive Statistics and Cohen Effect Sizes for Categories of Test Items Across Exams by Instructor and Treatment

	Rater #1	Rater #2	Rater #3	Rater #4
Rater #1	1.0 (n=1532)	0.59 (n=779)	NA	0.53 (n=753)
Rater #2	0.59 (n=779)	1.0 (n=1559)	0.59 (n=780)	NA
Rater #3	NA	0.59 (n=780)	1.0 (n=1563)	0.58 (n=773)
Rater #4	0.53 (n=753)	NA	0.58 (n=773)	1.0 (n=1536)

Table 4: Kendall's Tau-Beta Correlations for Initial Inter-Rater Reliability

	Definition	Recognition	Application	Calculation	Conceptual	Synthesis	Not Valid
From Students	634 (20.7%)	370 (12.1%)	74 (2.4%)	543 (17.7%)	823 (26.9%)	112 (3.6%)	506 (16.5%)
Across Exams	13 (17.3%)	1 (1.3%)	4 (5.3%)	26 (34.7%)	23 (30.7%)	8 (10.7%)	0 (0%)

Table 5: Distribution of Student Written Multiple-Choice Items Relative to Actual Exam Items during RPT Treatment