International Journal for the Scholarship of Teaching and Learning

Volume 7 | Number 2 Article 19

7-2013

Using Performance Tasks to Improve Quantitative Reasoning in an Introductory Mathematics Course

Gerald Kruse
Juniata College, kruse@juniata.edu

David Drews
Juniata College, drews@juniata.edu

Recommended Citation
Kruse, Gerald and Drews, David (2013) "Using Performance Tasks to Improve Quantitative Reasoning in an Introductory Mathematics Course," *International Journal for the Scholarship of Teaching and Learning*: Vol. 7: No. 2, Article 19.
Available at: https://doi.org/10.20429/ijsotl.2013.070219

Using Performance Tasks to Improve Quantitative Reasoning in an Introductory Mathematics Course

Abstract
A full-cycle assessment of our efforts to improve quantitative reasoning in an introductory math course is described. Our initial iteration substituted more open-ended performance tasks for the active learning projects than had been used. Using a quasi-experimental design, we compared multiple sections of the same course and found non-significant gains on a pre/post, rubric-scored, measure of quantitative reasoning. Subsequent course modifications included more explicit emphasis on critical thinking as a course goal and extended experience with the rubric used to score the performance tasks. Results of the second iteration yielded stronger evidence for gains in quantitative reasoning and suggest that the impact of open-ended performance tasks is increased when supported by efforts that emphasize their importance.

Keywords
Quantitative reasoning, Critical thinking, Performance task, Assessment, Rubric

Using Performance Tasks to Improve Quantitative Reasoning in an Introductory Mathematics Course

Gerald Kruse
Juniata College
Huntingdon, Pennsylvania, USA
kruse@juniata.edu

David R. Drews
Juniata College Huntingdon,
Pennsylvania, USA
drews@juniata.edu

Abstract

A full-cycle assessment of our efforts to improve quantitative reasoning in an introductory math course is described. Our initial iteration substituted more open-ended performance tasks for the active learning projects than had been used. Using a quasi-experimental design, we compared multiple sections of the same course and found non-significant gains on a pre/post, rubric-scored, measure of quantitative reasoning. Subsequent course modifications included more explicit emphasis on critical thinking as a course goal and extended experience with the rubric used to score the performance tasks. Results of the second iteration yielded stronger evidence for gains in quantitative reasoning and suggest that the impact of open-ended performance tasks is increased when supported by efforts that emphasize their importance.

Keywords: quantitative reasoning, critical thinking, performance task, assessment, rubric

Introduction

There is wide agreement by many stakeholders that Critical Thinking (CT) and its disciplinary cousins such as quantitative reasoning (QR) should be one of the primary foci of an undergraduate education (Bok, 2006; Halpern, 1998; Jones, 1995). Pascarella & Terenzini (1991) emphasize CT skills as vital for students, especially as "factual knowledge becomes more obsolete," and they transition to a world where they are expected to change careers multiple times, and often into jobs that do not yet exist. Critical thinking is mentioned in a majority of college mission statements, and Juniata College is no exception (http://www.juniata.edu/about/mission.html).

Definitions of CT vary. Some involve formal reasoning (e.g., Ennis, 1987; Mulnix, 2010; Paul & Elder, 2009), while others are less wedded to formal logic. Halpern (1998), for example, argues that CT "is the kind of thinking involved in solving problems, formulating inferences, calculating likelihoods, and making decisions."

In addition to variations in the definition of CT, Williams, Oliver, & Stockdale (2004) distinguish between generic and discipline specific CT, and cite data which support this distinction. For example, although Dunwoody, Baney, & McKellop (2011) found significant correlations between generic measures of CT and a measure of critical thinking in psychology, the generic measures accounted for less than half of the variation in the

disciplinary measure. Similarly, Nelson, Golding, Drews, & Blazina (1995) found even weaker correlations between a generic measure of CT and a measure of problem solving in international relations. As efforts to improve CT most often take place in the context of particular disciplines, data such as these encouraged the development of CT definitions and measures that have a disciplinary focus. Our particular concern is with Quantitative Reasoning (QR), which also has many definitions (Garfunkel & Mumford, 2011, Madison, 2001, Shavelson, 2008). In the present study, we consider QR as CT in the mathematical domain, given the shared emphases in each on evaluating evidence, drawing conclusions, and problem solving. More precisely, our understanding of QR is operationalized in the scoring rubrics we use for assessment and training. These rubrics are extensions of the College Learning Assessment's (CLA) definition of CT ("Architecture of the CLA Tasks," n.d.). We also note that QR is sometimes also referred to as Quantitative Literacy or Numeracy (Madison & Steen, 2008; National Numeracy Network, 2011; Steen, 1997).

Many studies have focused on improving CT and, given the close relationship between CT and QR, advice given about improving one might plausibly apply to the other. At one extreme, some acknowledge (e.g., Halpern, 1998) that modest increases in critical thinking ability may occur by simply attending college, or having a disposition to CT. But many argue that a more focused approach is needed to produce significant improvement. Halpern (1998), for example, suggests a skills approach with focus on recognizing and applying individual skill components of her definition of CT. As a general approach that might help operationalize Halpern's advice, Broadbear (2003) argues for courses across the curriculum that focus on facing students with ill-structured problems, ones that cannot be described with a high degree of completeness or solved with certainty. These are problems where experts may disagree. They have better or worse solutions, but no single right answer. Along this line, Shavelson (2008) has argued for "performance tasks" (PT), which not only are ill-defined and lack conclusive solutions, but also face students with information of varying relevance and veracity. Often called "authentic assessments", these kinds of problems are one type of task in the CLA. Recently, Arum & Roksa (2011) have used PT to facilitate student learning. These arguments inspired us to use PT to improve students' QR in a non-majors introductory math class and ultimately led us into a full-cycle assessment resembling that done by Blue, Taylor & Yarrison-Rice (2008).

To provide context for our study, Juniata College is a selective private liberal arts college with a student population of approximately 1500-1600 students. The average ACT score is 23. A majority of students come from Pennsylvania and surrounding states, but there is a significant national draw and approximately 10% of the student body are international students. Men represent 45% of the student body and students of color represent 10%. Approximately 30% are first generation college students.

As part of the college general education requirements, Quantitative Methods (QM) is an introductory course designed to serve students seeking to fulfill the college-wide quantitative literacy requirement. In an average year, five sections of the course serve a total of 120-150 primarily freshman and sophomores students. Approximately 35% of each graduating class takes this course. For many students, it is the only mathematics course they will take at Juniata. The main topics covered in the course include interpreting and creating graphs and statistics, personal finance, basic probability, sampling, and apportionment. Pedagogy focuses on active learning techniques and there is an emphasis on spreadsheet technology such as Microsoft Excel. The text used in all five sections, Quantitative Reasoning (Sevilla & Somers, 2007), aligns well with this pedagogy. Before beginning our experiment in fall 2009, all sections of the course employed daily labs to

supplement lectures as well as three more in-depth, open-ended projects. Analyzing student data for trends, weighing loan options and creating payment tables, and testing data for normality are examples of the content in these projects. On the first and last class days of the semester, attitudes toward mathematics, as well as quantitative skills were assessed. While students improved on these indices, the indices did not assess all facets of QR, which was part of the motivation for beginning this study.

Study 1

Method

The experimental design for Study 1 is presented in Table 1. Three sections of the course were involved. The first author taught the PT section, which employed performance tasks designed by that instructor. Each of the No PT sections was taught by a different instructor and both of those sections used the active learning tasks that had been traditional in QM. The only difference between the No PT sections was in the nature of the pretest. No PT/Quant Skills used the traditional pre test described above on the first day of classes. In contrast, the No PT/Pre Test and PT sections used a QR assessment modeled on the CLA In-the-Classroom tasks that use the "backward design" principles of Wiggins and McTighe (2005). Chun (2010) argues for this kind of active, "authentic" assessment task and

Table 1.
Design for Study 1

Instrument	**Section**		
	Performance Task (PT)	**No PT / Pre Test**	**No PT / Quant Skills**
Pre test	QR Assessment	QR Assessment	Traditional Quantitative Skills
Project 1 Project 2 Project 3	Performance Tasks, based on CLA	Traditional	Traditional
Post test	QR Assessment	QR Assessment	QR Assessment

describes them as requiring " . . . a complex, real-world challenge in which the scenario, role, process, and product are all authentic; they must then demonstrate that they have the skills and knowledge to complete the task." Madison (2006), Garfunkel & Mumford (2011), and Shavelson (2008) also argue for this type of real world approach to teaching and assessing QR.

The particular performance task (PT) we used for both the pre and post test presented students with a brief description of an election year controversy over the health consequences of legislation to regulate two different artificial sweeteners. Students were assigned the role of advisors to one of the candidates and were given a document library containing five documents to use as the basis for their recommendation. The information in the documents was presented in various forms, including graphs and tables, and it varied with respect to relevance and credibility. Students were given the entire hour class period to read these documents and write essays to the following prompts:

Prompt 1: Pat Sauer claims that "banning aspartame would improve the health of the state's citizens." What are the strengths and/or limitations of Pat Sauer's position on this matter? Based on the evidence, what conclusion should be drawn about Pat Sauer's claim? Why? What specific information in the documents led you to this conclusion?

Prompt 2: Pat Sauer claims that "aspartame should be banned and replaced with sucralose." What are the strengths and/or limitations of Pat Sauer's position on this matter? Based on the evidence, what conclusion should be drawn about Pat Sauer's claim? Why? Is there a better solution, and if so, what are its strengths and/or limitations? Be sure to cite the information in the documents as well as any other factors you considered (such as the quality of the research conducted on aspartame) that led you to this conclusion.

The instructions to students for the pre and post assessments were identical and were read to each class by the first author. As noted in Table 1, sections No PT/Pre Test and No PT/No Pre Test were taught as they had been before the start of the study. The fact that the only design difference between them is that No PT/Quant Skills section had the traditional pre test provides some information about the practice effect of doing the performance task used for the pre/post test twice. Table 1 also indicates that the projects used during the semester in the PT section were three CLA-type PTs. Each PT resembled the pre and post test in the sense that students were faced with multiple documents containing information presented in various forms, which varied in usefulness for the task at hand. During the semester though, these tasks were more directly focused on quantitative skills appropriate to that part of the course.

Student essays for pre and post tests were scored by means of rubrics such as that presented in appendix A, which operationalizes our definition of QR as critical thinking applied in the mathematical domain. The primary categories include: Evaluating evidence provided ("evaluating evidence"), Analysis/synthesis/conclusion ("conclusion"), Presenting/creating evidence ("create"), Acknowledging alternatives to their conclusion ("alternatives"), and Completeness ("completeness"). Score ranges on individual rubric dimensions varied with our ability to make distinctions in the quality of answers. While the rubric categories were stable across both studies 1 and 2, scoring details varied because of differences in the documents used in each study. The actual scoring of student essays was done with rubrics such as that presented in appendix A. The development of these rubrics was informed by the general advice of the Association of American Colleges and Universities (2009) and the CLA (2009) concerning the categories of critical thinking and by the recommendations of Boersma, Diefenderfer, Dingman & Madison (2011) and Madison & Dingman (2010) for making the transition from a holistic rubric to one suitable for scoring.

Students were told that the pre and post tests were course activities and that their participation was therefore required. However, they were also told that they could withhold their data from the analyses presented below and were asked to sign an informed consent statement that would allow the use of their data. Neither students nor faculty were randomly assigned to sections. A total of 61 students, all but one of those present on the day the pre test and quantitative skills test were administered agreed to participate; 7 men and 16 women in section PT, 8 men and 12 women in section No PT/Pre test, and 7 men and 11 women in section No PT/Quant skills. One student's data was dropped because they were under 18 years old.

Results and Discussion

Before the pre and post test responses were scored, a faculty colleague not connected to the project randomized them, removed student identification, and created a code so that section membership could be recovered and pre/post scores reunited after scoring was complete. Both authors independently read and scored essays from students who had missed either the pre or post test and used discussions of these to sharpen the scoring rubric. These essays, missing their pre or post mate, were used only for training purposes and the data from them are not included in the analyses that follow. Actual scoring began with the independent scoring of sets of 10-12 essays by both authors. After each set, we compared scores on each dimension to track scoring reliability, resolve differences, and make final adjustments to the rubric. A total of 22 essays were scored this way before dividing the rest of the set to be scored independently by each author. For question 1, possible total scores ranged from 0-25. For question 2, the range was 0-24. Total scores for each rater, summed across questions, were correlated as one measure of inter-rater reliability: $r = 0.873$, $n = 22$, $p<0.000$. Inter-rater reliability on individual dimensions was indexed by percent agreement. Across individual dimensions, scores were in complete agreement 63.6% of the time and within one 91.3% of the time. This approach to inter-rater reliability and our levels of agreement are consistent with a number of earlier studies using rubrics to assess student learning (e.g., Blue et al., 2008; Boersma et al., 2011; Stellmack, Konheim-Kalkstein, Manor, Massey, & Schmitz, 2009; Thaler, Kazemi, Huscher, 2009).

Scores for each dimension of the rubric have been combined across both student prompts for ease of exposition. Initial analyses compared post scores across all three sections using a series of one way ANOVAs. These analyses failed to reveal any significant differences between sections, on total scores or scores on any individual rubric dimension, all Fs<1.00 except for "complete," $F(2,58) = 2.41$, $p = .098$. These results suggest that the performance tasks failed to have their intended impact on QR. The fact that No PT/Pre Test section was not better than No PT/Quant Skills section on any dimension also suggests the absence of a practice effect from taking the assessment twice.

The primary analyses used two way analyses of variance with repeated measure on one variable to compare pre/post changes for the PT and No PT/Pre Test sections. F values for these analyses are shown in Table 2.

Table 2.
Two Way ANOVA F values for Study 1

	Total	**Evaluating evidence**	**Conclusion**	**Create**	**Alternative**	**Complete-ness**
Pre/post X Section Interaction	1.132	4.730*	0.014	0.866	2.556	1.954
Pre/post X main effect	0.073	0.0264	0.014	0.867	1.839	0.598
Section main effect	0.378	0.897	0.808	1.713	0.009	0.670

Note. all df 1,41; * p < .05; all other ps > .17

The most compelling evidence for a PT effect would be in the form of interactions between section and pre/post that showed more improvement for the PT section. Although the pattern of results for total scores was consistent with this expectation, with scores in the PT section appearing to increase from pre (M = 15.1, SD 4.8 = , *n* = 23) to post (M = 16.4, SD = 6.7, *n* = 23) and scores in the No PT/Pre Test section appearing to decline from M = 17.0 (SD = 6.8, *n* = 20) to M = 16.3 (SD = 4.8, *n* = 20), the interaction failed to reach significance. In addition there was no overall pre/post difference or an overall difference between sections.

Analyses of individual rubric dimensions showed a similar pattern with the exception of "evaluating evidence". In this one case, there was an unexpected interaction, produced by greater pre/post improvement in the No PT/Pre Test section. Overall, however, there was no pre/post main effect, or main effect of section. For "conclusion", there was no interaction, no pre/post effect, and no difference between sections. Similarly, for "create" there was no interaction, no pre/post difference, and no between sections main effect. For "alternatives", the pattern was the same, with no interaction, no pre/post effect, and no difference between sections. For "completeness", there was no interaction, no pre/post effect, and no difference between sections.

Final course averages failed to correlate with total scores on the post test in all three sections (r_{PT} = .139, *n* = 23, *p* = 0.526, $r_{No\ PT/Pre\ Test}$ = .021, *n* = 20, *p* = 0.931, $r_{\text{No PT/Quant Skills}}$ = .187, *n* = 18, *p* = 0.458.). While differences in grading practices across sections preclude any single explanation, it is clear that there is no close connection between the goal of critical thinking and the reward of a good course grade.

The absence of significant interactions, differences between sections, or improvement from pre to post tests provided scant evidence for the impact of the PT manipulation. On the other hand, the fact that the pattern of cell means on each dimension other than "evaluating evidence" resembled that for total scores, with at least marginally greater improvement in the PT section than the No PT/Pre Test section encouraged us to examine and adjust our approach for another iteration of the study.

Study 2

The results of Study 1 suggested that performance tasks by themselves were insufficient to produce the QR gains we had hoped for. Others have produced such gains. Kaddoura (2011) found significant improvement on a global measure of CT in nursing students following practice with case based learning. But the amount of practice he provided was much greater, spread throughout the entire three year program. On the other hand, Blessing & Blessing (2010) produced modest gains in their disciplinary measure of CT in psychology through a much shorter series of "PsychBusters" exercises that required students to evaluate the status of psychological findings that might appear in news reports. Though the amount of practice on these exercises was similar to what we provided, they had a much larger sample size which conferred greater statistical power. As our opportunities were confined to a single course with fewer students, we sought other avenues to increase the impact of PT. Guided by Blue et al. (2008) and van Gelder (2005), we made several adjustments for the second iteration. In general, these involved more explicit and repetitive emphasis on QR as a course goal and explicit repeated experience with the general and scoring versions of the QR rubric.

Method

Two sections of MA 103 provided the students for Study 2. The experimental design for Study 2 omitted No PT/Quant Skills section, but was otherwise identical with Study 1. All students present on the day the pre test was administered signed consent forms allowing us to use their data. The PT section had 28 graded pre/post pairs (8 men, 20 women) while there were 25 graded pre/post pairs in the No PT section (7 men, 18 women). In addition, four students completed either the pre or post test, but not both. These protocols were used to train scorers and sharpen the scoring rubric. The No PT section was taught by a different instructor than in Study 1 but covered the same material, used the same textbook, used similar assignments, and was taught using the same active learning pedagogy as the No PT/Pre Test section in Study 1. The PT section was modified in several ways for Study 2:

- The course syllabus was modified with an increased emphasis placed on QR, in order to increase its legitimacy as a course goal. Lectures and course notes, while covering the same content as during Study 1, were occasionally modified to reinforce this emphasis.

- Prior to the first performance task during the semester, teams of students were asked to create a rubric for "chips," and then given a variety of snacks to score with their rubric. A discussion of rubrics ensued, concluding with a presentation of a generalized version of the QR rubric. The rubric was then discussed before each remaining performance task to promote an understanding of our definition of QR.

- During the semester, after each task was graded, a detailed scoring rubric and student responses were returned. Additional experience with the rubric was provided during discussions, which included examples of low and high scoring samples of student work for each rubric dimension.

- Modest revisions of the performance tasks used as projects, guided by the generalized rubric, helped sharpen the focus of the performance tasks on particular dimensions of the rubric.

As in Study 1, a pre/post test was administered by the first author on the first day of class and during the last week of class. In contrast to Study 1, the assigned role of the students was switched from an advisor for one of the candidates to an intern at a non-partisan foundation, to avoid the possibility of that their prior role constrained their responses. In addition, some simplifying wording changes were made, one document was adjusted to require more quantitative analysis, another document was dropped, and the two prompts were combined into one. As before, students were given the entire class period to write their responses.

Results

To establish inter-rater reliability, 30 protocols were selected at random in sets of 10 and independently scored by both authors. For total scores, which could range from 0-21, $r = .927$, $n = 30$, $p < .000$. Scores on individual dimensions of the rubric (see appendix A) also showed high levels of inter-rater reliability, with independent scores within one at least 89% of the time on each scale. After scores on each set of 10 protocols were compared, discussion resolved disagreements. The remaining protocols were scored by one or the other of the authors.

As in Study 1, the primary analyses used PT/No PT ***x*** pre/post two way ANOVAs with repeated measures on one factor. In addition, given the expectation of greater improvement in the PT section, planned pre/post comparisons were performed on each section.

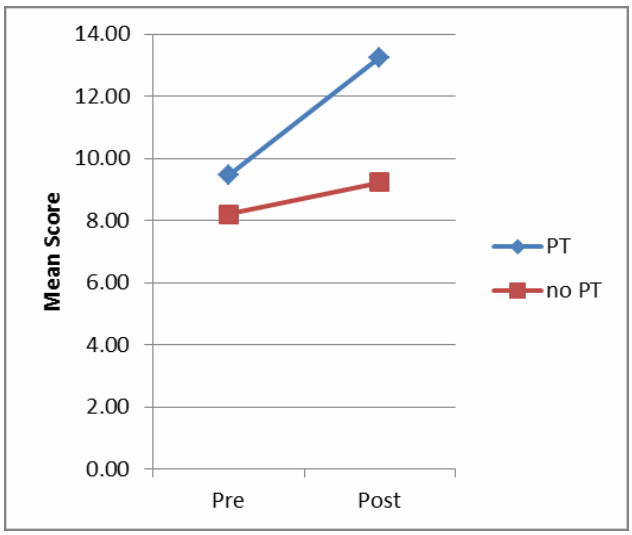

Figure 1: "Total Score" as a function of section

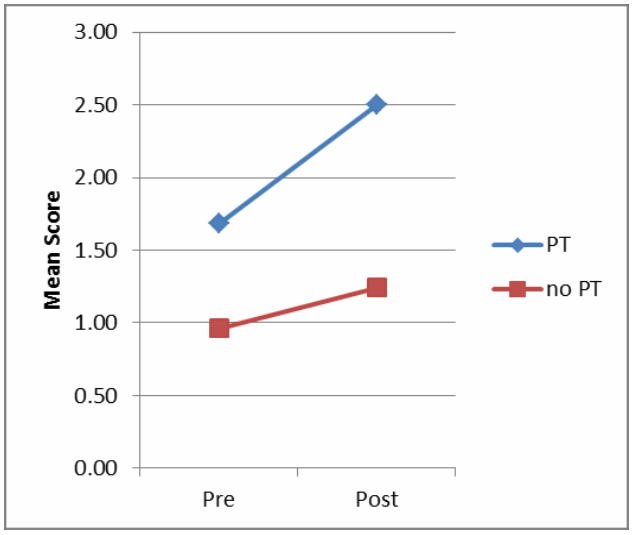

Figure 2**:** "Evidence" as a function of section

The interaction for "total scores" approached significance, $F(1,51) = 3.360$, $p = .073$. Cell means for this interaction are presented in Fig 1. Planned comparisons showed pre/post improvement in the PT section, $F(1,51) = 15.568$, $p = .000$, $d = .796$, but not in the No PT section, $F(1,51) = 1.455$, $p = .233$. Overall, there was a main effect reflecting improvement from pre to post, $F(1,51) = 12.864$, $p = .001$, and a main effect showing better overall performance in the PT section, $F(1,51) = 6.911$, $p = .011$.

For "evidence", although the interaction was not significant, $F(1,51) = 1.241$, $p = .271$, planned comparisons showed improvement in the PT section, $F(1,51) = 6.054$, $p = .017$, $d = .434$, but not in No PT section, $F(1,51) = 0.628$, $p = .432$. The cell means are shown in Fig 2. Overall, post scores were larger than pre scores, $F(1,51) = 5.134$, $p = .028$ and the PT section out-performed the No PT section, $F(1,51) = 6.307$, $p = .015$.

For "conclusion", the interaction was not significant, $F(1,51) = 0.748$, $p = .391$. Planned comparisons on the means presented in Fig 3 showed improvement in the PT section, $F(1,51) = 9.670$, $p = .003$, $d = .789$, and improvement that approached significance in No PT section, $F(1,51) = 3.057$, $p = .086$. There was overall improvement from pre to post, $F(1,51) = 11.605$, $p = .002$, but no overall difference between sections, $F(1,51) = 2.922$, $p = .092$.

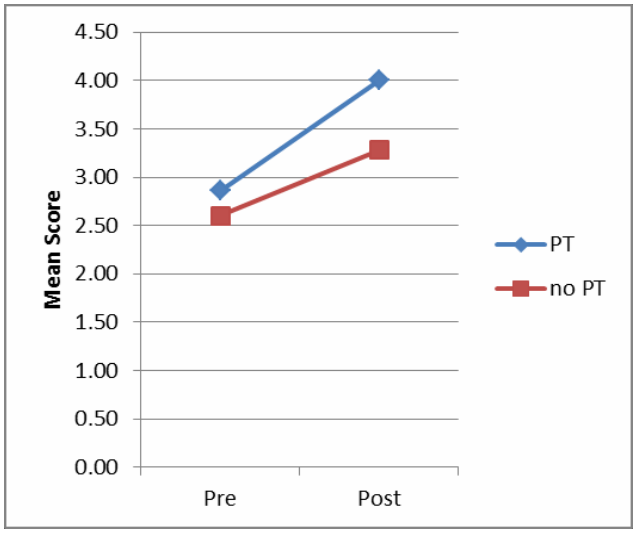

Figure 3: "Conclusion" as a function of section

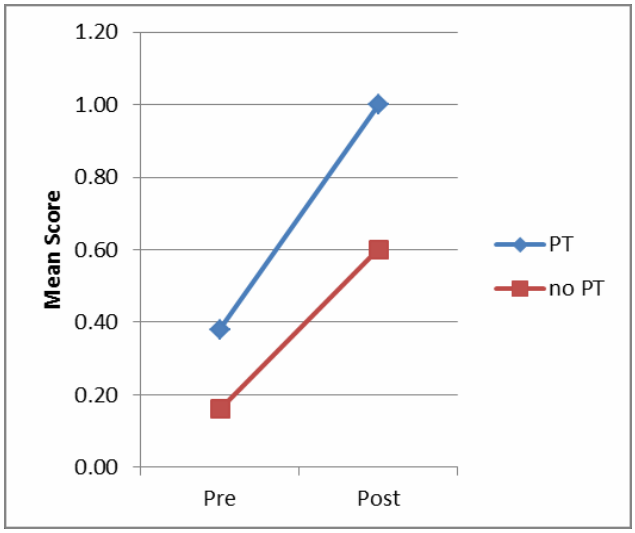

Figure 4: "Create" Evidence as a function of section

The interaction for "create" also failed to reach significance, $F(1,51) = 0.484$, $p = .489$. Planned comparisons on the means shown in Fig 4 indicate significant progress in both the PT section, $F(1,51) = 13.553$, $p = .001$, $d = .693$, and the No PT section, $F(1,51) = 6.355$, $p = .015$, $d = .732$. Consistent with this, there was an overall main effect for pre/post, $F(1,51) = 19.017$, $p = .000$. The main effect for section approached significance, $F(1,51) = 3.493$, $p = .067$ with somewhat higher scores in the PT section.

The analysis for "alternatives" indicated an interaction that approached significance, $F(1,51) = 3.341$, $p = .073$. Despite the appearance of the means in Fig. 5, planned comparisons show no significant change for either the PT section, $F(1,51) = 1.190$, $p = .280$, or the No PT section, $F(1,51) = 2.202$, $p = .144$. In addition, there was no overall pre to post improvement, $F(1,51) = 0.109$, $p = .743$ and no difference between sections, $F(1,51) = 0.962$, $p = .331$.

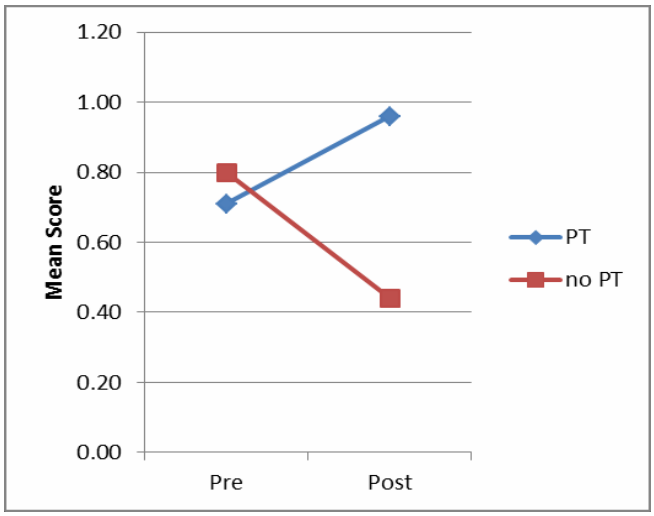

Figure 5: "Alternatives" as a function of section

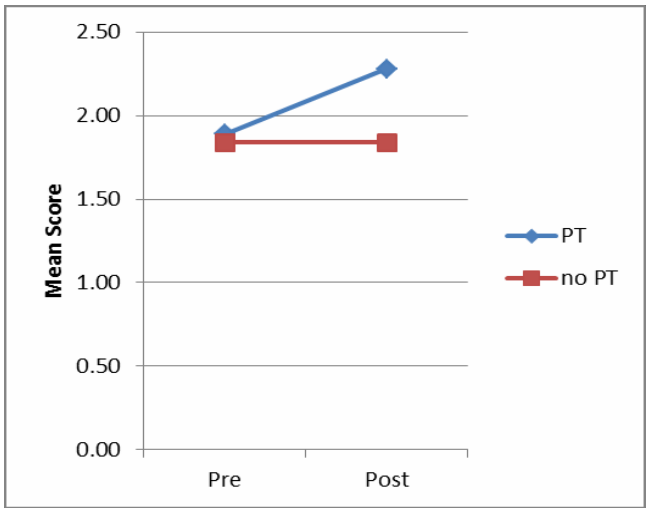

Figure 6: "Complete" as a function of section

The cell means for "complete" are shown in Fig 6. The interaction for these means did not reach significance, $F(1,51) = 2.834$, $p = .098$, but planned comparisons showed improvement in the PT section, $F(1,51) = 6.009$, $p = .018$, $d = .584$, and none in No PT section, $F(1,51) = 0.000$, $p = 1.000$. Overall, pre/post improvement approached significance, $F(1,51) = 2.834$, $p = .098$. The performance of the PT section was marginally better than that of No PT section, $F(1,51) = 2.910$, $p = .094$.

Looking across these analyses, the pattern of greater improvement in the PT section than No PT section is present on total scores and on three individual dimensions ("evidence", "conclusion" and "complete"). On "create" that pattern fails to obtain only because both sections show improvement. It is worth noting that "create" is the only dimension that shows pre/post change in the No PT section. Given the emphasis on learning basic manipulations of numerical data in all the versions of this course, improvement on this dimension is not surprising. More generally, these data are consistent with the notion that performance tasks, combined with the other course changes made to the PT section in Study 2, can produce significant improvement in QR. The only rubric dimension that failed to show significant improvement in the PT section was "alternatives", which, on reflection, received relatively little emphasis during the course.

As in Study 1, final course averages were correlated with total post QR scores. In section NO PT, $r = .102$, $n = 25$, $p = 0.628$. In section PT however, course averages did correlate with total QR scores, $r = 0.455$, $n = 28$, $p = .015$. This suggests the possibility that the incentive of course grades might have helped facilitate QR in Study 2 more than they did in Study 1. The only conscious change in grading in the PT section was that in Study 1, two of the PTs were team projects with shared grades while in Study 2 just one of the PTs was a team project.

In Study 2, since we recorded how much time students took to finish both the pre and post tests in both sections, we correlated time spent with scores on each of the rubric dimensions. In general, as one would expect, scores improved as students spent more time on either test. On the pre test, time spent was positively related to the total score, $r = .311$, $n = 53$, $p = .023$, "conclusion", $r = .333$, $n = 53$, $p = .015$, "create", $r = .334$, $n = 53$, $p = .014$, and "completeness", $r = .311$, $n = 53$, $p = .023$. Time spent was not related to "evidence", $r = .182$, $n = 53$, $p = .191$, or "alternatives", $r = -.055$, $n = 53$, $p = .003$. On the post test, time spent was positively related to total score, $r = .398$, $n = 53$, $p = .003$, "evidence", r = .396, n = 53, p = .003, "conclusion", r = .399, $n = 53$, $p = .003$, "alternatives", $r = .329$, $n = 53$, $p = .016$, and "completeness", $r = .426$, $n = 53$, $p = .001$. Time spent was less strongly related to "create" $r = .252$, $n = 53$, $p = .068$. Not surprisingly, these correlations suggest that student motivation to engage the task posed by the pre and post test contributes to their QR scores.

Summary and Concluding Discussion

We believe that we have produced significant improvements in QR using a combination of performance tasks, emphasis on QR as a course goal, and explicit attention to the definition of that goal via repeated experience with the rubric that operationalizes it. The most compelling evidence of our success is the planned comparisons from Study 2 that show greater pre/post gains in the PT section than in the No PT section.

There are potential alternative explanations for these QR gains. One might be that the instructor of the QR section was simply more successful. Though we cannot rule out such an explanation, other successful studies with manipulations similar to ours (e.g. Blessing & Blessing, 2010), especially ones that use the same instructor across groups (Kaddoura, 2011) argue for the importance of method over instructor differences. Another is the possibility that, since students were not assigned at random, the QR section was graced with better students. This possibility is weakened by the pre/post planned comparisons. These analyses ignore possible student differences at the beginning of the term and focus on changes that took place within students during the course. A third possibility is that students in the PT section were more motivated than those in the non PT section. Time spent on task is a plausible index of motivation, one that is positively correlated with QR scores. And students in the PT section did spend more time than those in No PT section. These time differences between sections might explain QR differences between sections, but they cannot explain the QR gains in the PT section or the lack of same in No PT section. The reason for this is that students within each section spent the same amount of time on the pre test as they did on the post test.

A final issue is whether our success in Study 2 might be understood as simply "teaching to the test." While we believe this could be a fair criticism of content based assessments, where teaching to the test might mean providing the content answers ahead of time, we do not believe it to be applies well here. First, we would point out that although the rubrics used during the course had the same general categories as the one used to score the post test, each had details appropriate to the particular PT. Similarly, the PT documents used during the course differed from those in the pre/post test. Thus training and test were not identical and at least some generalization of training was needed to produce our scoring gains. More generally, we do not think “teaching to the test” is a fair criticism of any skills based assessment. Consider a physical skill such as shooting a jump shot in basketball or hitting a forehand in tennis. The most obvious way to produce such a skill is to give people extended practice under a variety of conditions, including conditions that resemble game conditions as much as possible. We do not think that skill development resulting from this training is less geniuine because the coaching techniques are "teaching to the test". Likewise, we believe that QR reasoning is a skill that is most likely to show measurable improvements after similarly pointed practice.

If one accepts the proposition that we have produced real gains in QR, it is difficult to be very precise about the causes of those gains. There are good reasons to believe that the practice at QR provided by the performance tasks is very important (Blessing & Blessing, 2010; Kaddoura, 2011; Mayes, Bonilla, & Peterson, Wiggins, 2001). That said, the results from Study 1 do not provide convincing evidence that, by themselves, performance tasks are enough, at least within the bounds of the practice we provided. There are also reasons to believe that introducing students to the rubric and giving them practice using it made an important contribution (e.g., Blue et al., 2008). Although we have argued that the motivational effects reflected in time spent on the pre/post task can’t explain the gains in QR during the semester, we do believe that instructor enthusiasm and efforts to highlight the importance and legitimacy of QR as a course goal are not irrelevant to student motivation and to our success. Finally, we would note that course grades might provide a powerful way to legitimize QR or any other course goal. Telling students that QR is important may be more believable if improvements in it lead to better grades. The fact that course averages were not correlated with QR scores in Study 1 but positively related in Study 2 may have contributed to our greater success in Study 2.

Our results agree with Blue et al. (2008) and Blessing and Blessing (2010) that gains in CT/QR are achievable within the confines of a semester-long course. However, our experience in Study 1 also argues that Tsui (1999) and Pascarella & Terrenzini (2005) are correct when they suggest that such gains are not easy to achieve. As suggested by van Gelder (2005) and Broadbear (2003), the more explicitly one pursues such a goal and the more course elements are directed at that goal, the greater the reward is likely to be.

Although our results are consistent with a number of earlier efforts and generally act to confirm advice previously given, we believe that our quasi-experimental design is methodologically more convincing than successful demonstrations based on single-group pre/post comparisons. In particular, showing pre/post gains in the PT section and lack of same in a substantially parallel No PT section, especially one with some emphasis on active learning experiences, is more convincing than the pre/post gains of the PT section by themselves.

And finally, evidence of domain specificity in critical thinking (Dunwoody, et al., 2011, Nelson, et al., 1995), argues for circumspection with regard to our conclusions, which apply most clearly in the context of reasoning in a quantitative domain. At the same time, the fact that the pedagogical advice that emerges from our study is so similar to that from quite different domains (Blue et al., 2008, Ennis, 1993, Halpern, 1998) holds some promise that it may be more generally useful.

Acknowledgements
We are indebted to Dr. David Widman for his help with our statistical analyses. We also owe much to the faculty and administrators of Juniata's vibrant SoTL community. In 2011, Juniata's SoTL center was officially renamed the James J. Lakso Center for the Scholarship of Teaching and Learning, in recognition of Provost Lakso's support of teaching excellence. The SoTL Center promotes scholarly teaching through bi-weekly Brown Bag lunches, Learning Communities, and summer research grants, and this project was a beneficiary of each type of support. It was initially presented at a Brown Bag lunch in a mentoring session that provided valuable suggestions. Progress reports were presented several times in that forum as our research evolved. The project was also awarded two summer research grants from the SoTL center, for which we are particularly grateful. Finally, we are grateful for the advice of the IJSoTL reviewers.

References

Arum, R., Roksa J. (2011), *Academically Adrift: Limited Learning on College Campuses*. Chicago: The University of Chicago Press.

Association of American Colleges and Universities VALUE Rubrics. (2009). Retrieved from http://www.aacu.org/value/rubrics/.

Blessing, S. B. & Blessing, J. S. (2010). PsychBusters: A means of fostering critical thinking in the introductory course. *Teaching of Psychology*, *37*, 178-182.

Blue, J., Tayor, B., & Yarrison-Rice, J. (2008). Full-Cycle Assessment of Critical Thinking in an Ethics and Science Course, *International Journal for the Scholarship of Teaching and Learning*, *2*(1). http://academics.georgiasouthern.edu/ijsotl/.

Boersma, S., Diefenderfer, C., Dingman, S. W., & Madison, B. L. (2011). Quantitative Reasoning in the Contemporary World, 3: Assessing Student Learning. *Numeracy*, *4*. doi:10.5038/1936-4660.4.2.8

Bok, D. (2006), *Our Underachieving Colleges: A Candid Look at How Much Students Learn and Why They Should Be Learning More*, Princeton: Princeton University Press.

Broadbear, J. T. (2003). Essential elements of lessons designed to promote critical thinking. *The Journal of Scholarship of Teaching and Learning*, 3(3). http://www.iupui.edu/~josotl/VOL_3/NO_3/broadbear_vol_3_no_3.htm.

Chun, M. (2010, Mar-Apr). Taking Teaching to (Performance) Task: Linking Pedagogical and Assessment Practices. *Change Magazine*, *42*(2), 22-29.

Council for Aid to Education, (n.d.). Architecture of the CLA Tasks. Retrieved from http://www.collegiatelearningassessment.org/files/Architecture_of_the_CLA_Tasks.pdf.

Dunwoody, P. T., Baney, J., & McKellop, J. M. (2011, October). Linking departmental and institutional assessment of critical thinking. Poster presented at the *International Society for the Scholarship of Teaching & Learning: Transforming the Academy through the Theory and Practice of SoTL*, Milwaukee, WI.

Ennis, R. H. (1987). A Taxonomy of Critical Thinking Skills and Dispositions. In Baron, J. B. & Sternberg, R. J., (Eds.), *Teaching Thinking Skills: Theory and Practice* (pp. 9-26). New York: W. H. Freeman.

Ennis, R. H. (1993). Critical Thinking Assessment. *Theory into Practice*, *32* (3), 179-186.

Garfunkel , S. & Mumford, D. (2011, August 24). How to Fix Our Math Education. *New York Times*. Retrieved from: http://www.nytimes.com/2011/08/25/opinion/how-to-fix-our-math-education.html.

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American Psychologist*, *53*(4), 449-455. doi:10.1037/0003-066X.53.4.449

Jones, E. A. (1995). (Ed.), National Assessment of College Student Learning: Identifying College Graduates' Essential Skills in Writing, Speech and Listening, and Critical Thinking. Final Project Report.

Kaddoura, M. A. (2011). Critical Thinking Skills of Nursing Students in Lecture-Based Teaching and Case-Based Learning. *International Journal for the Scholarship of Teaching and Learning*. *5*(2). http://academics.georgiasouthern.edu/ijsotl/.

Madison, B. L. (2001). Quantitative Literacy: Everybody's Orphan, *Focus, 6*.

Madison, B. L. (2006). Assessment of Student Learning in College Mathematics: Towards Improved Programs and Courses. Association for Institutional Research.

Madison, B. L. & Dingman, S. W. (2010). Quantitative Reasoning in the Contemporary World, 2: Focus Questions for the Numeracy Community, *Numeracy*, *3*. doi:10.5038/1936-4660.3.2.5

Madison, B. L. & Steen, L. A. (2008). Evolution of Numeracy and the National Numeracy Network, *Numeracy*, *1*. doi:10.5038/1936-4660.1.1.2

Mayes, R., Bonilla, R., & Peterson, F. (n.d.). Quantitative Reasoning: Current State of Understanding. Retrieved from http://coe.georgiasouthern.edu/QR/QR%20Overview%20-%20Mayes,%20Peterson,%20Bonilla.pdf.

Mulnix, J. W. (2010). Thinking critically about critical thinking. *Educational Philosophy and Theory*. 44(5), 464-479. doi:10.1111/j.1469-5812.2010.00673.x

National Numeracy Network (NNN). (2011). What is numeracy/QL/QR? Retrieved from http://serc.carleton.edu/nnn/resources/index.html.

Nelson, L., Golding, N. L., Drews, D. R., & Blazina, M. K. (1995). Teaching and assessing problem solving for international conflict resolution. *Peace and Conflict: Journal of Peace Psychology*, 1(4), 399-416.

Pascarella, E. T. & Terenzini, P. T. (1991). *How College Affects Students: Findings and Insights from Twenty Years of Research*. San Francisco: Jossey-Bass Inc.

Pascarella, E. T. & Terenzini, P. T. (2005). *How College Affects Students: A Third Decade of Research* (2nd ed.). San Francisco: Jossey-Bass Inc.

Paul, R. & Elder, L. (2009). *Miniature Guide to Critical Thinking-Concepts and Tools*. Foundation for Critical Thinking.

Sevilla, A. & Somers, K. (2007). *Quantitative Reasoning: Tools for Today's Informed Citizen*. Emeryville, CA: Key College Publishing.

Shavelson, R. J. (2008). Reflections on quantitative reasoning: An assessment perspective. In Madison, B. L. & Steen, L. A. (Eds.), *Calculation vs Context: Quantitative Literacy and its Implications for Teacher Education*, Mathematical Association of America, Washington, DC.

Steen, L. A. (1997). The New Literacy. In L. A. Steen (Ed.), *Why Numbers Count: Quantitative Literacy for Tomorrow's America* (pp. xvi-xxviii). New York, NY: The College Board.

Stellmack, M. A., Konheim-Kalkstein, Y. L., Manor, J. E., Massey, A. R., & Schmitz, J. A. P. (2009). An Assessment of Reliability and Validity of a Rubric for Grading APA-Style Introductions. *Teaching of Psychology*, *36(*2), 102-107. doi:10.1080/00986280902739776

Thaler, N., Kazemi, E., & Huscher, C. (2009). Developing a Rubric to Assess Student Learning Outcomes Using a Class Assignment. *Teaching of Psychology*, *36*(2), 113-116. doi:10.1080/00986280902739305

Tsui, L. (1999). Courses and Instruction Affecting Critical Thinking. *Research in Higher Education*, *40*(2), 185-200.

van Gelder, T. J. (2005). Teaching Critical Thinking: Some Lessons from Cognitive Science. *College Teaching*, *53*(1), 41-46.

Wiggins, G. (2001, Dec). "Get Real!" Assessing for Quantitative Literacy. In Madison, B. L. & Steen, L. A. (eds.) *Quantitative Literacy: Why Numeracy Matters for Schools and Colleges* Proceedings of the National Forum on Quantitative Literacy held at the National Academy of Sciences in Washington, D.C. Retrieved from http://www.maa.org/ql/pgs121_143.pdf.

Wiggins, G. P., McTighe, J. (2005). *Understanding By Design*. Alexandria, VA: Association for Supervision & Curriculum Development.

Williams, R. L., Oliver, R. & Stockdale, S. (2004). Psychological Versus Generic Critical Thinking as Predictors and Outcome Measures in a Large Undergraduate Human Development Course. *The Journal of General Education*, *53*(1), 37-58.

Study 2

QuantitttlthlRtcuoninQ Assssm nt Rubrk

,.,__onJN st•	HoM / EIMIJI/ng	C.	Maerlttg
tti!lal!!ll ll!!!illllf<>'lidj!!jl ....tit evil- ..... — - ----- la. KHp taNn bw forbit mlwndlnlondwhDt says/mil S0!< Docwt!e'>l AM , -___Oftd/l'Offl M Oco.lm"'' IJ Is on«z/otol Docwt!e'>l C.bi«H Oft(/"""' ' Ocla.l!nl'll D • and from "oeuptobl.sOUIC•	E!a!ll2f !foam.nt d; ....,,.,.. sll) lfchcu .16xvn1ntaA 1111dQ, i1nd flnomi Bilnd C,dleR II 1011'11 ra!Ofldl.tthltt llllllta — 11111'AiltAl} !!!!!! 2f,7 toam.ntl!!!i !5V'ttt!n.!t · · 11;21	h:t!ll2n l!!!$0111DIW l!!!l.t: !!"' Qble!IMMiooJill h:t!ll2n mw : eeoc! tvt!ut!iorM 00CVT'IInt A II r•'o'¥1t1111clfrom iln iiQQIP.tlblt $(JUI'W DOCV!!lnt a b1111eccbUI Oocur.nt Cb bAled 1111dlnot RIIYilnt 00CVT'IInt ;0 Is 11111v¥Jt1111d from 1111 iiQQIP.tlblt aourc.	El:t!lalt!l 2f 4 doamll!!!1J :tt!:l!!: -RbluMMiooJ 1$} tB!Ialt!l 2f sfocunll!!!1J:tt!:l!!: &Wdtvt!ll siGJ •0ocv1l'1111t A ii 11111'Ailtlllld tTclm 1111 iiCQIIptlblt $oU'c:t •Oocv1!'1111tiUai1ntcciCJtl1 •OocvIl'lll1t C i lblilled'•nd not 11111v¥1t •0ocvlll'lll1t0 ii RIIYilnli1nd flom1111 iiCQIIptlblt — —
61J8ikiWI1JIJS1il lg;nl — 1. Doc-tol!. CIMtiN MJiln, lflwr!COIICt.llll ... n0 C!CfNtt b lb.__tiM<iiiCIIMttd IIW'lit, Sxnlas•,orfll..t>oo:n< 111'11an 2. Doc-to has <>0I'IdlioNI prd>ahllity t.otio...t.ct. s hows fytMdadoas •/'d A>porume L Cmcblll>n • .....tmdeano ton «J/1 ajti.d ls...,..,.thtxl gnor ng lb.d«untiH!il tlflrihH	o. dlrectlr .,correc:tll •peo• willl S.U..-, or CX111clusionls uncle.-: 'Ooidl ''"'Wfence(1) or !!Ome <Md.,ce, I'l0t just *ilfe'*'r/"'<> 1 SMJer(2 If n.tyzed Doc:ument A (CMJdler!) or Document D e -dlestnd as lndePI!ndentl wel but no c:andullotl 12)	1...,..or dudfy dislp11es willl S.U..-, a. CXll1ducies "'more resellrdl 11eededl: ilniilyllrc Doame11t A (Ollt5en101t oocunent o fild'adleotnd MPI!Ume -.dq.eooerot}, oot illlilysis is In CaJme/tlnd- ccm te (3) M uses A AND D lilt !Niint811n!tsooeof dleml3) << •netr.Js of A 01t OIs ccod (4)-	cw dlrectlr d.a.....,es Willi S.uer, cw ca cludes "more res-- rct.11eeded = illlify!q OxtJment A j<Mdien) AND 'Doc:ument D llee•ct.esend asPttttme lindepe11diimt}. bit illlifrllis is: ""ct..rcw.,CDIT"fllel>!(5] Orlt c-!(6
fiiiiDS!!.Cmil!l•rddBSI Owq......, • ...,...an:!dan.. proboMiit(	Tiles to reCifllnille infomwtian to nfNI fomwt lilt d-5C) Poortv (1}.	:MIA:eosome r-Oflltie e«ort to reotp lre Infomwdon lilt aMJid do mcn,.tletter, h I'!C (21	ReclrpriZeoJ infomwiion to Cl'fllte ctmPI! _ ..,.!{tlbtesa.CCMidJiiCMlll p-cWIIfity, RiAl (3}
.old>1!lf!!lldli!!JII!rtmrifta UUl!l!::CCHtduslcH! lfdlupeewfS.ow.•rd11otbul A>port.me,o m.iodp flit A>piiW!!le c...a.tbo bod, ..._.,.u:•looe ls rot .,oltft'nllil<'e	AdBiwltdPI poaitilityd•lttmiltMII) todlelr ca clu!iion, IMJt does 11at ellbonte or 'Ooidl jiO<>r reil!ai(s}(l).	D•CII!ci ilnllttmiiMI poss ,;- tolfls coodusiofl, aa:epUiile ratiOnale M iUAiOft c-	•Diaibci llll llttmritt /Witspre"Qiiootodis c:anl!lu!lion,'Ooidlattejblti!riltiacille o. :iUpp;rt(3
Cm\doJonct' Ufiiii"M llifdtSialuorDIJIII!formore,..•dt	Otff mll:e! c:atdu boor Seuer's c:llkn. lilt uses no«iniiJII'()IIIIMe Ndlua (1)'-	oeels wldl :S.Uer'sl*lNimll •tts tian toslnlnllflt nd weeblt!leoof •'l'llmentlll	ouh ..tdl :S.Uer'sdlpb ccod MttrdOfl to m.rcltll nd wellllmsesof lfT'Jmenl m
cqn•mcu A(pft w.tlt	Otff mite! c:atdu!lOll boor seuer's ctwn.. llltusesnoo.ineNdenoe (1)-	oeehwllfl seuer'sdlim tndll'o'fOI •l*n Olt ...rcdls tncl wellllleo sesof lfT'Jmtnt m	ouh w.iltl seuer'sdltnd"""u .,..., AND m.nllflttnd welllleo sesof ilriUmlnl c/1 llll n131

•• the c:ategcries for hl!tler crder -sa the same asIn 1tudy1, but the S(l)ring dec:alls were changed to It the 1.4sed doo.mer(s