Journal of the Scholarship of Teaching and Learning, Vol. 13, No. 2, May 2013, pp. 23 – 30.

Interteaching: Discussion group size and course performance

Jacob C. Truelove, Bryan K. Saville, and Ryan Van Patten[1]

Abstract: Researchers have yet to examine whether discussion group size affects student performance in an interteaching-based course. In the current study, we addressed this question by manipulating discussion group size (smaller groups of 2 students vs. larger groups of 4 students) across 2 sections of an undergraduate psychology course. We found no significant differences between the sections on 6 unit exams, on a cumulative final exam, and in the total number of points earned across the semester.

Keywords: interteaching, group size, discussion, exam performance, behavior analysis

I. Introduction.

Interteaching is a relatively new method of classroom instruction that has its roots in behavior analysis (Boyce & Hineline, 2002). Like previous behavior-analytic teaching methods—which include precision teaching (Lindsley, 1964); programmed instruction (Skinner, 1968); direct instruction (Engelmann & Carnine, 1982); and, arguably the most well-known of these methods, Keller's (1968) personalized system of instruction (PSI)—interteaching attempts to improve student performance by identifying what behaviors students should emit to improve their course performance and then rearranging the reinforcement contingencies to produce those behaviors. Unlike previous behavior-analytic teaching methods, though, interteaching may be easier to implement in traditional classroom settings (Boyce & Hineline, 2002).

A typical interteaching session proceeds as follows (for more detail, see Boyce & Hineline, 2002; Saville, Lambert, & Robertson, 2011). Prior to class, students complete an instructor-created preparation (prep) guide that contains questions over a reading assignment. Each class typically begins with a lecture that lasts approximately one third of the class period and covers material from the previous class (see below). After the lecture, students divide into pairs and discuss the prep-guide questions they answered for class. During the discussions, the instructor moves around the classroom, answering questions and guiding discussion. After students finish their discussions, they complete a record sheet on which they note their partner's name, how well their discussion went (along with reasons why), and any questions they would like the instructor to review. The instructor then uses the information on the record sheets to prepare a lecture that begins the next class period and precedes discussion of the next prep guide.

Since Boyce and Hineline's (2002) introduction of interteaching, researchers have found that it typically produces better student-learning outcomes than lecture-based teaching methods (e.g., Saville, Zinn, & Elliott, 2005; Saville, Zinn, Neef, Van Norman, & Ferreri, 2006; Scoboria & Pascual-Leone, 2009; for a review, see Saville et al., 2011). Researchers have also conducted a small number of studies examining which components of interteaching contribute to its efficacy (Saville, Cox, O'Brien, & Vanderveldt, 2011; Saville & Zinn, 2009). To date, however, researchers have not studied the discussion component of interteaching.

[1] James Madison University, truelojc@dukes.jmu.edu

In their original description of interteaching, Boyce and Hineline (2002) suggested using pairs during the discussions to avoid social loafing (Latané, Williams, & Harkins, 1979). In contrast, Goto and Schneider (2010) reported that their students preferred working in larger groups of four students, which some researchers have suggested will provide superior outcomes in cooperative learning situations (e.g., Johnson & Johnson, 2009). Neither Boyce and Hineline (2002) nor Goto and Schneider (2010), however, reported any systematic performance data. Thus, the purpose of the present study was to examine discussion group size and student performance in an interteaching-based course. Specifically, we asked students to work in pair or in groups of four and then measured their performance on six unit exams and on a cumulative final exam; we also examined the total number of points students earned across the semester.

II. Method.

A. Participants.

Participants were 61 undergraduate students from James Madison University, a large, public university considered to be "more selective" by the Carnegie Foundation for the Advancement of Teaching (www.carnegiefoundation.org). The students in this study, the majority of whom were juniors (see Table 2), were enrolled in two sections of an undergraduate psychology of learning course taught by the second author. Section 1 contained 30 students (25 women, five men), and Section 2 contained 31 students (28 women, three men). Section 1 met on Tuesdays and Thursdays from 12:30-1:45 p.m., and Section 2 met on Tuesdays and Thursdays from 2:00-3:15 p.m.

B. Materials and Procedure.

The instructor assigned a prep guide for students to complete before each class. The prep guides usually covered 10 to 20 pages of textbook material and contained anywhere from eight to 12 items (each of which often contained multiple questions) that required students to define concepts, apply course material, and engage in higher-order thinking (see Appendix for a sample prep guide from the course). Once in class, students divided into groups (for more information, see below) and discussed their answers to the prep-guide questions. The instructor encouraged the students to choose different discussion partners each class period, but given the relatively small number of students in each section, it was not always possible for them to work with an entirely different set of partners each time. During the discussions, the instructor and a teaching assistant (TA) walked around the room and answered any questions that students had. After finishing their discussions, students completed a record sheet on which they listed their partner's name, how well their discussion went (along with reasons why it went well or poorly), and which material they wanted the instructor to clarify. Students who participated in the discussions and turned in a record sheet earned a small number of participation points that across the semester totaled 10% of their course grades (Boyce & Hineline, 2002). At the start of the next class period, the instructor lectured over material that the majority of students had listed on the record sheets. The class then got into groups and discussed the next prep guide.

There were six 45-point exams during the semester, each of which followed discussion of three or four prep guides. Each exam consisted of approximately 20 items, most of which were short-answer questions along with a few multiple-choice and fill-in-the-blank questions. The

questions were based on, but were not identical to, items from the prep guides and typically required students to solve problems, apply information, and show higher-level comprehension. For example, a sample prep-guide question was "Discuss the one-process and two-process theories of avoidance." whereas two related short-answer exam questions were "How would a one-process theory of avoidance explain a fearful person's tendency to avoid dogs?" and "How would two-process theory explain a person's fear of heights?" At the end of the semester, students took a 90-point cumulative final exam that covered all of the prep guides and contained short-answer, multiple-choice, and fill-in-the-blank questions.

To measure the impact of group size on student performance, we had Section 1 (Large Group) discuss the prep guides in groups of four students (cf. Goto & Schneider, 2010) and Section 2 (Small Group) discuss the prep guides in pairs (cf. Boyce & Hineline, 2002).[2] Because we could not randomly assign participants to the conditions, we took two steps to ensure that the groups were relatively equal prior to manipulating group size. First, at the beginning of the semester, we collected the following demographic data: gender, age, current year in school, cumulative GPA, number of psychology classes taken so far, number of credits taken during the semester, and employment status. Second, prior to the first exam, we had both sections complete their discussions in pairs (which is the way interteaching was originally described by Boyce & Hineline, 2002). These measures provided a baseline against which we could compare the sections after our manipulation.

C. Interobserver Agreement.

For each exam, one TA graded all 61 exams, while a second TA graded a subset of 15 exams. To determine interobserver reliability (IOR), we divided the number of questions on which the TAs gave the same number of points by the total number of questions on the exam and multiplied by 100. The average IOR across the six exams was 87% (range = 83-92%). When the TAs disagreed on a question, they discussed the item and came to agreement on the final score.

III. Results and Discussion.

We first examined students' demographic information. One student in the Large-Group section (Section 1) only provided her gender on the demographic questionnaire. In the Small-Group section (Section 2), one student provided no information other than gender, another student did not report the number of credits she was taking, and a third student did not report her age and GPA. Our demographic analyses are thus based on the remaining data that participants provided. In sum, we found no significant differences between the sections on any of the demographic measures (all ps > .30, Table 1).

We next examined students' performances on the first exam (which, along with the demographic information, served as a baseline). One student in the Large-Group section did not take Exam 1. Thus the following analysis is based on the scores of 29 students in the Large-Group section and 31 students in the Small-Group section. We found no significant difference between sections on Exam 1, $t(58) = 0.50$, $p = .62$ (Large Group mean = 81%, Small Group mean = 80%). Together with the demographic information, this finding suggests that the sections were relatively similar prior to our manipulation.

[2] In Section 1 (Large Groups), depending on attendance, we sometimes had to let one or more groups have five students. Similarly, in Section 2 (Small Groups), we sometimes had to let one group have three students.

Table 1. Demographic information for the Large-Group (Section 1) and Small-Group (Section 2) sections.

	Section 1 (Large Group)	Section 2 (Small Group)
Gender		
Male	5	3
Female	25	28
Age (in years)	*M* = 20.90 (SD=0.94)	*M* = 21.03 (SD=0.73)
Year in School		
Junior	27	26
Senior	3	4
GPA (out of 4.00)	*M* = 3.37 (SD=0.41)	*M* = 3.40 (*SD* = 0.33)
Psychology Courses Taken	*M* = 10.86 (*SD* = 4.00)	*M* = 9.79 (*SD* = 3.83)
Semester Credits	*M* = 14.66 (*SD* = 2.50)	*M* = 14.69 (*SD* = 1.92)
Employed		
Yes	14	14
No	15	16

For Exams 2 through 6 (which students took after we manipulated group size), we conducted a 2 (group size) x 5 (exam) mixed ANOVA. We found a main effect for exam, $F(4, 236) = 2.61$, $p = .04$, $\eta_p^2 = .04$ (which, for the purposes of this study, is of little importance), but no main effect for group size, $F(1, 59) = 0.01$, $p = .92$. The average score across all of the unit exams was approximately 83% for both sections. In addition, we found no interaction between group size and exam, $F(4, 236) = 0.66$, $p = .62$ (see Figure 1 for all exam scores). Finally, we did not find a significant difference between sections on the cumulative final exam, $t(59) = 1.22$, $p = .23$ (Large Group mean = 84%, Small Group mean = 86%) or in the total number of points (unit exams plus cumulative final exam) earned across the semester, $t(59) = .527$, $p = .60$ (Large Group mean = 296 of 360 possible points, Small Group mean = 301 of 360 possible points).

In sum, there were no significant differences between the Large-Group section (groups of four students) and the Small-Group section (pairs) on any of the unit exams, on the cumulative final exam, and in the cumulative number of exam points earned across the semester. In their original recommendations on how to implement interteaching, Boyce and Hineline (2002) suggested using pair discussions to minimize social loafing. Goto and Schneider (2010), however, reported that students in their interteaching-based course preferred larger groups of four students. Neither Boyce and Hineline nor Goto and Schneider provided any systematic data, however, to show whether discussion group size affected performance. The present results suggest that when teaching an interteaching-based course, using smaller groups of two (or three) students or larger groups of four (or five) students may not result in differential course performance, at least as measured by exam performance and the related measure of cumulative exam points.

There are at least two possible reasons why we did not find significant differences in the present study. First, as Williams, Harkins, and Latané (1981) demonstrated, identifiability of an individual's contribution can help deter social loafing in larger groups. With interteaching, one's contribution to the discussion is often recorded by other group members on their record sheets and may also be apparent to the instructor as he or she roams the classroom answering questions. These publicly viewable events may function to increase participation (i.e., eliminate social loafing), even in larger groups of four or five students. If social loafing was reduced in the

Large-Group condition, then it may not be surprising that exam performance between the two conditions was similar in the present study. Given, however, that identifiability often becomes more difficult with increased group size (see Guerin, 1994), there may be a point where performance begins to deteriorate. Future research may thus wish to examine what happens to performance in an interteaching-based course when group size increases beyond four or five members. Specifically, researchers could replicate the present study but include a greater number of students (e.g., six to eight or more) in the Large-Group condition. This would provide more information on whether there is a point at which a "large" group becomes too large.

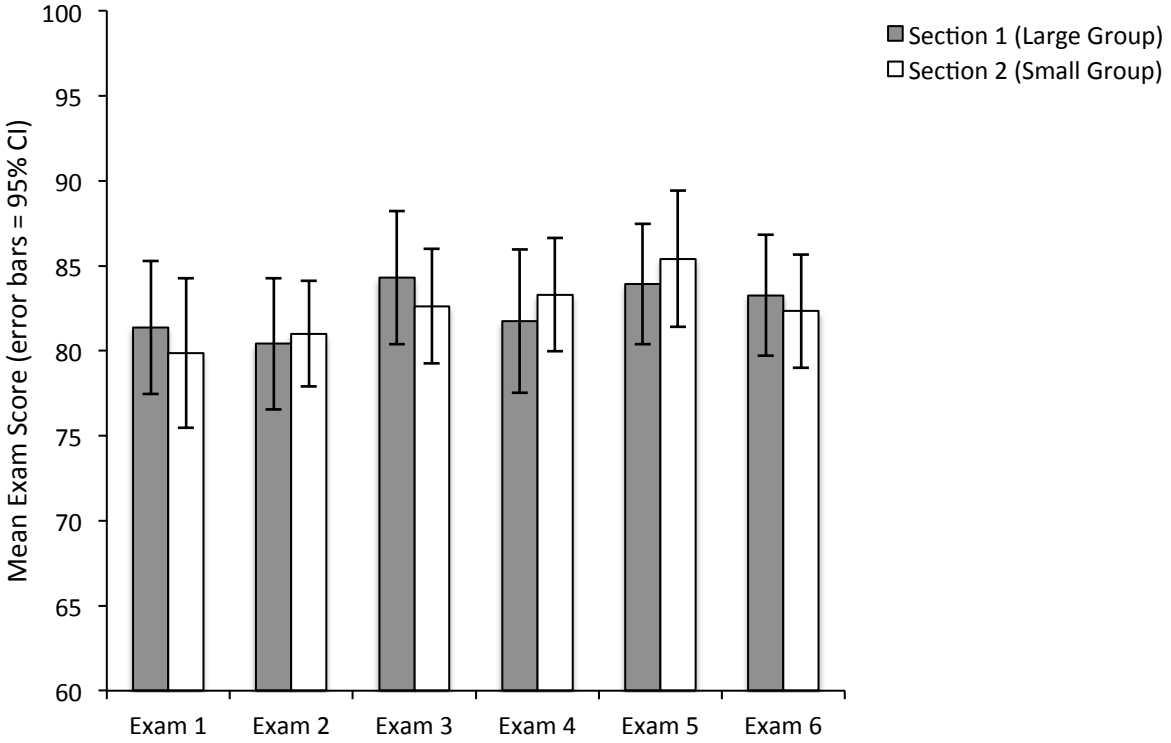

Figure 1. Shows the mean exam scores and 95% confidence intervals for students in the Large-Group (Section 1) and Small-Group (Section 2) sections. Students in both sections worked in pairs prior to Exam 1.

Second, it is also possible that the discussions simply are not an important component of interteaching. If this is the case, one would not expect differences in group size to produce differences in course performance. Unfortunately, determining the contribution of group discussions to interteaching is not possible from the present study because it lacked a true "no-discussion" control condition. Future researchers could examine this possibility more directly by exposing two groups to interteaching but eliminating the discussions in one group. Specifically, rather than discussing the prep guides with another student, students in the no-discussion group could use class time to re-study the prep guides they completed before class.

There are other factors that future researchers might wish to examine more closely as well. As noted earlier, the institution at which we collected our data is considered by the Carnegie Foundation to be "more selective" with its admissions criteria (meaning that our students typically have strong academic backgrounds). Our samples also consisted largely of

women, who tend to perform better in college than men (e.g., Mau & Lynn, 2001). Although these variables most likely did not *differentially* affect our results (as shown by our demographic analyses), it is possible that having these types of students in our study may have produced a small ceiling effect, which clouded our ability to identify significant differences (see also Saville & Zinn, 2009). In short, it would be interesting to see if our results might change when the samples studied are more diverse in nature.

Ultimately, if future research determines that the discussions *are* an important component of interteaching, instructors might then wish to consider student preferences when deciding whether to use smaller or larger groups (Wolf, 1978). If students prefer one group size to the other, allowing them to determine how many partners they have may increase their enjoyment of interteaching, which may improve course performance further.

Acknowledgements

The authors would like to thank Tracy Zinn and Kevin Apple for their excellent comments on an earlier version of this paper.

Appendix. Shows a Sample Preparation Guide from the Course.

Preparation Guide #10

Based on: Ch. 7, pp. 267-280

1. What is a schedule of reinforcement? Discuss the difference between continuous and intermittent reinforcement schedules. Identify some behaviors that are reinforced continuously and some that are reinforced intermittently. Which of these is more representative of the types of schedule that operate in our daily lives?

2. What is the relation between the response requirement and the postreinforcement pause in fixed-ratio (FR) schedules? Imagine you are a business owner who is trying to get your employees to be more productive. How might you incorporate a FR schedule to do this? Would you use a small FR schedule or a big FR schedule? Explain your answer.

3. In what way are VR and FR schedules similar? Different? Give some real-life examples of behavior maintained by VR schedules. How might you use these schedules to modify your own behavior?

4. How are the patterns of behavior produced by FI and FR schedules different? What are some behaviors that are maintained by FI schedules?

5. If you owned a casino and wanted visitors to gamble a lot, would you program your slot machines to pay off according to a FI, VI, FR, or VR schedule? Be sure to discuss what pattern of behavior (i.e., pulling the "arm" of the slot machine) each schedule would produce.

6. What are noncontingent schedules, and how do they differ from contingent schedules? How do these schedules possibly account for superstitious behaviors? Also, discuss how noncontingent schedules are likely involved in the development of "learned laziness."

7. In recent years, there has been a push to increase children's self-esteem (called the "self-esteem movement") by making sure that, for example, every child gets a trophy or award or even good grades, regardless of how well they actually perform. The belief is that receiving these "rewards" will make children feel good about themselves, which will then result in improved performance. Unfortunately, studies are showing that the "self-esteem movement" is having negative effects on children's performance. Based on what you know about noncontingent schedules of reinforcement, explain why this is not surprising.

8. If most of our daily behaviors are reinforced under complex schedules (e.g., conjunctive schedules, adjusting schedules, chained schedules), why do you think psychologists have spent so much time studying simple schedules?

References

Boyce, T. E., & Hineline, P. N. (2002). Interteaching: A strategy for enhancing the user-friendliness of behavioral arrangements in the college classroom. *The Behavior Analyst, 25,* 215-226.

Engelmann, S., & Carnine, D. W. (1982). *Theory of Instruction: Principles and Application.* New York: Irvington.

Goto, K., & Schneider, J. (2010). Learning through teaching: Challenges and opportunities in facilitating student learning in food science and nutrition by using the interteaching approach. *Journal of Food Science Education, 9,* 31-35. doi:10.1111/j.1541-4329.2009.00087.x

Guerin, B. (1994). *Analyzing Social Behavior: Behavior Analysis and the Social Sciences.* Reno, NV: Context Press.

Johnson, D. W., & Johnson, R. T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. *Educational Researcher, 38,* 365-379. doi:10.3102/0013189X09339057

Keller, F. S. (1968). Good-bye teacher… *Journal of Applied Behavior Analysis, 1,* 79-89. doi:10.1901/jaba.1968.1-79

Latané, B., Williams K. D., & Harkins S. G. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology, 37,* 822-832. doi:10.1037/0022-3514.37.6.822

Lindsley, O. R. (1964). Direct measurement and prosthesis of retarded behavior. *Journal of Education, 147,* 62-81.

Mau, W-C., & Lynn, R. (2001). Gender differences on the Scholastic Aptitude Test, the American College Test and college grades. *Educational Psychology, 21,* 133-146.

Saville, B. K., Cox, T., O'Brien, S., & Vanderveldt, A. (2011). Interteaching: The impact of lectures on student performance. *Journal of Applied Behavior Analysis, 44,* 937-941. doi:10.1901/jaba.2011.44-937

Saville, B. K., Lambert, T., & Robertson, S. (2011). Interteaching: Bringing behavioral education into the 21st century. *The Psychological Record, 61,* 153-166.

Saville, B. K., & Zinn, T. E. (2009). Interteaching: The effects of quality points on exam scores. *Journal of Applied Behavior Analysis, 42,* 369-374. doi:10.1901/jaba.2009.42-369

Saville, B. K., Zinn, T. E., & Elliott, M. P. (2005). Interteaching versus traditional methods of instruction: A preliminary analysis. *Teaching of Psychology, 32,* 161-163. doi:10.1207/s15328023top3203_6

Saville, B. K., Zinn, T. E., Neef, N. A., Van Norman, R., & Ferreri, S. J. (2006). A comparison of interteaching and lecture in the college classroom. *Journal of Applied Behavior Analysis, 39,* 49-61. doi:10.1901/jaba.2006.42-05

Scoboria, A., & Pascual-Leone, A. (2009). An 'interteaching' informed approach to instructing large undergraduate classes. *Journal of the Scholarship of Teaching and Learning, 9,* 29-37.

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

Williams, K. D., Harkins, S. G., & Latané, B. (1981). Identifiability as a deterrent to social loafing: Two cheering experiments. *Journal of Personality and Social Psychology, 40,* 303-311. doi:10.1037/0022-3514.40.2.303

Wolf, M. M. (1978). Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis, 11,* 203-214. doi:10.1901/jaba.1978.11-203