APPLIED COGNITIVE PSYCHOLOGY
Appl. Cognit. Psychol. **24**: 1183–1195 (2010)
Published online 15 September 2009 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/acp.1630

The Testing Effect: The Role of Feedback and Collaboration in a Tertiary Classroom Setting

MARIJA VOJDANOSKA,
JACQUELYN CRANNEY* and BEN R. NEWELL

University of New South Wales, Sydney, Australia

SUMMARY

Successful retrieval on a test compared to just re-studying material improves long-term retention—a phenomenon called the 'testing effect'. This study investigated the role of feedback and collaborative testing on the retention of course material in a tertiary educational setting. Tested material was better retained relative to non-tested material (testing effect), and feedback facilitated correction of errors. Group testing produced higher performance on the initial, but not final test performance, compared to individual testing. This set of findings suggests that to encourage long-term retention, educators should utilise individual formative testing with feedback; theoretical implications are also discussed.

The core purpose of education and training is the transfer of knowledge to students, such that this knowledge is retained in the long-term. Research has suggested that testing is one way to increase long-term retention, relative to re-studying material or not being tested (Glover, 1989; Spitzer, 1939). This phenomenon is called the 'testing effect'. One prominent theoretical explanation singles out the key role played by retrieval effort in accounting for the testing effect. Internal memorial processes, like retrieval, are proposed to be the direct effect of the testing procedure on memory, in contrast to indirect effects such as motivation for subsequent study (see Roediger & Karpicke, 2006a, for a review). Most of the research on the testing effect has been conducted in the laboratory. Formative testing with feedback, as well as group activities, is becoming more common in educational settings; however the effectiveness of these strategies, particularly in combination, has received little evaluative attention. The current study examines these issues in a tertiary educational context.

The testing effect

Despite the implications for education and training (Chan, McDermott, & Roediger, 2006; McDaniel & Fisher, 1991; Roediger & Karpicke, 2006a), there are few studies on the testing effect in a tertiary educational context (Bangert-Drowns, Kulik, & Kulik, 1991; Leeming, 2002). In a recent classroom study by McDaniel, Anderson, Derbish, and

*Correspondence to: Jacquelyn Cranney, School of Psychology, Sydney, NSW 2052, Australia.
E-mail: j.cranney@unsw.edu.au

Morrisette's (2007) students volunteered to be part of an experiment that was undertaken within the psychobiology unit in which they were enrolled. An advantage of short-answer (cued-recall) testing over merely re-reading material was reported. In a study by Cranney, Ahn, McKinnon, Morris, and Watts (2009; Experiment 2), first-year psychology students watched a psychobiology video in Phase 1. During Phase 2 participants were either tested on a set of short-answer questions (Test), given statements matched to the set of questions and answers and asked to highlight important points (Re-study), or given no task (Control). Phase 3 occurred a week later and all participants completed a cued recall test containing the questions. Cranney et al. found that being tested in Phase 2 led to better long-term retention of the tested material, relative to the re-study (strict criterion) and control (lenient criterion) groups. One aim of the current study was to replicate the lenient criterion test effect, but with different material—a Power Point presentation of a developmental psychology topic.

Feedback

Feedback has been a focus of intense theoretical and empirical work in the educational domain (e.g. Hattie & Timperley, 2007; Sassenrath & Garverick, 1965; Shute, 2008), with Butler and Winne (1995) proposing that feedback 'is information with which a learner can confirm, add to, overwrite, tune or restructure information in memory' (p. 263). Although Kluger and DeNisi (1996) reported greater effect sizes when feedback provided information on correct rather than incorrect information, others have argued that feedback is highly effective when it allows students to learn the correct answer to any questions they answered incorrectly (McKeachie, 1963). In support of this latter notion, McDaniel et al. (2007) found that for items that were incorrect on the initial (Phase 2) test, feedback led to better subsequent test performance than did no feedback. To explore the potential mechanisms underlying feedback effects, our study explicitly manipulated feedback and analysed the effects of feedback following initially correct *versus* incorrect responses, on final test performance. Overall, we expected feedback to result in better long-term retention.

Collaborative testing

Although there has been some investigation of the role of collaboration during the initial recall stage (Phase 2 in the testing effect paradigm), little research has examined what effect this initial collaboration has on later tests that are undertaken individually. This is important to examine because although students in a classroom setting may collaborate initially, the final test they undertake is often an individual test. Cranney et al. (2009) found that working as a group (*cf*. individual) led to better performance on both initial and final tests, relative to students who performed the initial test individually, and students who read and highlighted material. Collaborative test-taking may give students the opportunity to correct misinformation, or to learn items that they did not initially encode (Blumen & Rajaram, 2008; Sainsbury & Walker, 2008).

Conversely, other research on the role of memory conformity suggests negative effects of collaboration on later recall. For example, Roediger, Meade, and Bergman (2001) argue that people falsely remembered information introduced in group discussion because they misattribute it to the original learning phase. Additionally, Blumen and Rajaram (2008) suggest that the retrieval disruptions due to collaboration that occur during the initial test

DOI: 10.1002/acp

(resulting in 'collaborative inhibition') might impair later memory. Thus, the research on memory conformity and collaborative inhibition suggests that collaboration may result in less accurate memory at a later stage.

In terms of performance on the initial test, although Cranney et al. (2009) reported an advantage of group over individual performance, this does not exclude the possibility of collaborative inhibition occurring. To test for collaborative inhibition, one needs to compare the performance of collaborative groups with nominal groups. In nominal groups, scores are derived by pooling individual correct responses on each question (excluding overlaps). Thus if any one individual in a nominal group gets a question correct, it is counted as correct. These scores are combined to form a total correct and this total is then compared to the single total number of correct responses produced by a collaborative group. Weldon and Bellinger (1997) found that nominal groups remembered more than collaborative groups, and that collaborative groups remembered more than individuals during an initial test. The difference between the nominal and collaborative performances suggests that a process loss occurs when people combine their information collaboratively (as opposed to nominally). This may be due to group interaction disrupting individual retrieval strategies (Basden, Basden, Bryner, & Thomas, 1997) or negative group dynamics such as social loafing (Latane, Williams, & Harkins, 1979). In contrast to the findings of Cranney et al. (2009), then, it is possible that if collaborative inhibition is occurring during Phase 2, as revealed by nominal group analyses, then students who were in the group condition compared to those in the individual condition in Phase 2 will perform worse on a final retention test.

The current study

This study aimed to replicate the basic testing effect in the classroom, and in addition, examine the long-term memory effects of feedback and collaborative testing. We predicted that there would be a main effect for the testing condition (Feedback, No Feedback, Control), with the Feedback condition yielding better memory performance than the No Feedback condition (congruent with laboratory studies such as Pashler, Cepeda, Wixted, & Rohrer, 2005), which in turn would yield better performance than the Control condition (congruent with the general testing effect findings; Roediger & Karpicke, 2006a). In addition, we expected to replicate the McDaniel et al. (2007) finding that feedback leads to greater correction of originally incorrect items than does no feedback.

For the initial test, we expected to replicate Cranney et al.'s (2009) finding that collaborative groups would outperform individual participants. We also expected to find this difference in the final individual test. In contrast, if we find evidence of collaborative inhibition in the initial test (i.e. nominal groups perform better than collaborative groups), this could have a negative impact on the final (individual) test performance for those students that collaborated on the initial test (Blumen & Rajaram, 2008).

METHOD

Participants

One hundred and thirty-three first year psychology students from the University of New South Wales participated in the experiment as part of their course-work during tutorials in

DOI: 10.1002/acp

weeks 4 and 5 of the first semester. Although all enrolled students participated in the experiment, data were used from those in the first six tutorials in the week. This was because it was possible that students in later tutorials would learn about the final test, leading them to specifically study, thus contaminating the results (i.e. the communication problem; Meltzoff, 1998). Data from three participants were removed because they chose not to provide consent to the data being used. Data from a further 24 participants who did not attend a tutorial in week 5, and thus did not complete the experiment, were also removed. Of the final one hundred and six ($n = 106$) students whose data were included, 27 were male, 62 were female and 17 did not disclose their gender. Their ages ranged from 17 to 28 ($M = 19.27$, $SD = 2.19$).

Design

This experiment employed a 2 (Collaboration: Group *vs*. Individual) × 3 (Testing Condition: Feedback *vs*. No Feedback *vs*. Control) between by within subjects design. Within the collaboration factor, participants either collaborated (Group) or not (Individual) on the initial test, and the Testing factor was a combination of two variables: Whether the final test questions were tested in the initial test, and whether participants received feedback on tested material.

The primary dependent variable was long-term retention for the material (as measured by correct responses on the final test). An additional dependent variable was performance (correct responses) on the initial test.

Materials

A PowerPoint presentation, approximately 10 minutes in duration, and including short videos, was constructed to contain at least 24 items of information relevant to adult development. Each item was expressed in cued recall question form as a statement with a key word missing, (e.g. *Erikson described _____stages of psychosocial development, from birth to late adulthood*). Note that this form of questioning, rather than a mulitple choice question format, was chosen because of the memory error issues associated with the latter in formative assessment (Roediger & Marsh, 2005). These items were randomly allocated to one of three item sets: A ($n = 8$), B ($n = 8$) or C ($n = 8$), which were pilot-tested to check whether the item sets were of equal difficulty. The per cent accuracies achieved on the items assigned to set A ($M = 50.00$, $SD = 25.62$), set B ($M = 47.00$, $SD = 19.44$) and set C ($M = 49.00$, $SD = 25.89$) were not significantly different, $F\ (2,\ 20) = 0.23$, $p > 0.05$, $p\eta^2 = 0.02$, indicating that the sets were of similar difficulty. However, we note that our pre-testing sample was small ($n = 11$). Each of these sets corresponded with a retrieval condition, and these were counterbalanced across tutorials such that there were three orders: ABC, BCA and CAB. In each of these set orders, the first letter denotes the Feedback items, the second letter denotes the No Feedback items and the third letter denotes the Control items.

Two tests were designed: An initial test and a final test. The initial test contained 16 test items, half of which received feedback and half which did not. The final test contained all 24 questions, including the items from the initial test as well as an additional 8 Control items (not previously tested). Each of the tests was presented on the back of a piece of paper, which had other questions on the front (see Procedure). The tests were in cued recall format and responses were written.

DOI: 10.1002/acp

Procedure

Phase 1
The Group condition participants were divided into groups of 4–6, and told that there would be a group activity later. All participants then viewed the PowerPoint presentation. They were instructed not to take notes during the presentation as it would be available to them on the WebCT course site later on.

Phase 2 (Immediately after Phase 1)
Students in the Group condition joined their groups of 4–6 and collaboratively answered the 16 quiz items. In these groups, there was much animated, but low-volume discussion during the quiz (so as not to 'give away' answers to surrounding groups—intergroup competition was encouraged). Students in the Individual condition answered the 16 quiz items individually and without discussion. They were given 8 minutes to complete the quiz, after which time the responses were collected. This period of time appeared sufficient for participants to recall what they knew, as in the last couple of minutes of this retrieval period, most students were no longer responding. Feedback was then given in a PowerPoint presentation for the Feedback items only (half of the tested items). Feedback involved the visual presentation of the question on a Power Point slide and the tutor reading this out aloud, and waiting for an oral answer response from the students (approximately 15 seconds). The answer was then visually presented and read out loud (approximately 5 seconds). The answer was greeted by expressions of glee from some students, and moans from others. The next question was then presented, and so on. Only two or three students overtly queried why only some items were given feedback, and tutors told them that time had run out, and that all answers would be posted on the course website prior to the mid-session examination. This feedback process took approximately 4 minutes. All students were told that the test was part of a class experiment on which their assignments would be based, and that there would be more discussion of it next week. They were asked not to talk to other students in the course about the experiment because it could have a confounding effect on the results. They were not informed that they would be tested on the presentation material in the following week. This was to ensure that there was a minimal effect of extra study hours.

Phase 3
In the same tutorials exactly 1 week after Phase 1 and 2, students were given an unexpected final test. Before sighting the questions, students were asked to rate how well they thought they would do on the test and to indicate how many hours of study on the topic they had undertaken during the previous week. They then completed the final test, which contained both previously tested and untested items (total = 24 items). They were given 12 minutes to complete the final test. At the end of the test students were given information about the experiment. They were then given the option of allowing their data to be included in the study by writing a short statement of consent and signing it. Papers were then collected, and feedback given. Students were told that they should not tell their classmates about the test and that the PowerPoint material would be released at the end of the week, in time to study for the mid-session exam.

Data reduction and nominal groups

For the Individual and Group conditions, the total number of items correct (out of 8) for each of the Feedback, No Feedback and Control sets was determined and converted to a

DOI: 10.1002/acp

percentage. As well as comparing the memory performances of collaborative groups and individuals, we also generated nominal group scores from the individual scores. This was accomplished through random assignment of individuals to nominal groups, with the condition that the mean number of individuals in nominal groups ($M = 4.45$, $SD = 0.93$) was matched with the mean number in collaborative groups ($M = 4.07$, $SD = 1.21$), $t(23) = -0.87$, $p > 0.05$, $d = 0.35$. Moreover, this matching occurred for each question set order grouping.

RESULTS

Retention analyses

Initial test

We analysed retention of the 16 items using a 2 (collaboration) × 3 (set order: ABC, BCA, CAB) univariate ANOVA. There was a significant difference between Groups ($M = 67.41$, $SD = 13.91$) and Individuals ($M = 39.54$, $SD = 15.01$) such that, on average, Individuals performed worse on the test than Groups, $F\ (1,\ 57) = 45.23$, $p < 0.05$, $p\eta^2 = 0.44$, thus supporting our hypothesis (see also Figure 1). In additional analyses, nominal groups ($M = 77.84$, $SD = 11.31$) performed better on the initial test than collaborative groups ($M = 67.41$, $SD = 13.91$); however this trend was not significant, $F(1,\ 19) = 4.21$, $p = 0.054$, $p\eta^2 = 0.18$. There were no significant set order main or interaction effects for either of these sets of analyses.

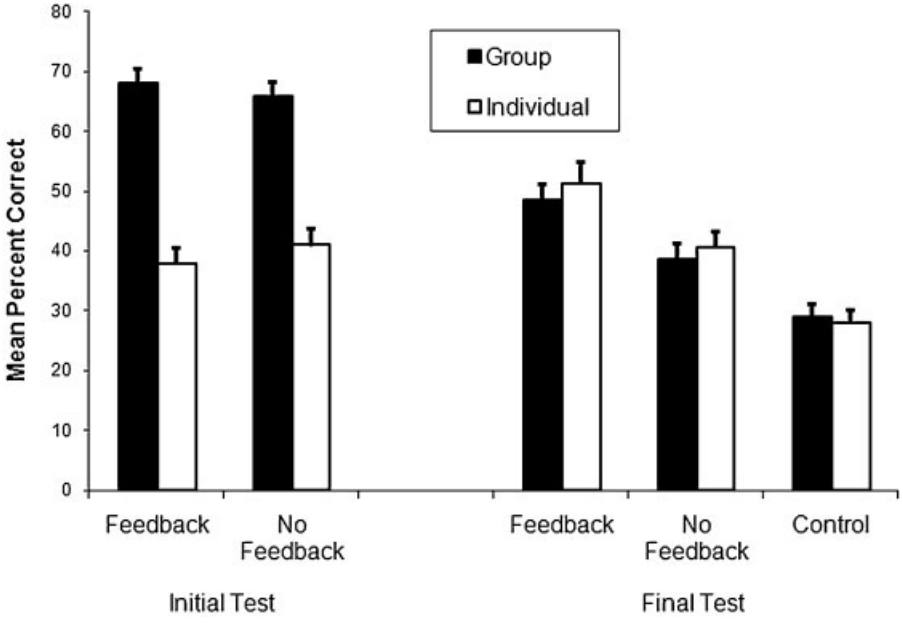

Figure 1. Mean per cent correct as a function of collaboration, testing condition and initial/final test conditions (error bars are standard error of the mean)

DOI: 10.1002/acp

Final test

A 2 (collaboration) × 3 (set order) × (3) (testing condition) repeated measures ANOVA of final test performance confirmed that there were no significant main or interaction effects for collaboration (i.e. no difference between Groups and Individuals). As suggested in Figure 1, there was a significant effect for testing condition, $F(2, 200) = 46.39$, $p < 0.05$, $p\eta^2 = 0.32$, with participants showing higher performance for Feedback items than for No Feedback and Control items (Feedback *vs*. No Feedback, $F(1, 100) = 23.87$, $p < 0.05$, $p\eta^2 = 0.19$; Feedback *vs*. Control, $F(1, 100) = 82.26$, $p < 0.05$, $p\eta^2 = 0.45$). Additionally, performance was better for No Feedback items than for Control items, $F(1, 100) = 25.99$, $p < 0.05$, $p\eta^2 = 0.21$, which constitutes a replication of the basic testing effect. It should be noted that although there was a significant main effect for set order, $F(2, 100) = 6.78$, $p < 0.05$, $p\eta^2 = 0.12$, further analyses revealed that this was due to Set A being easier than the other two sets, despite the pilot work attempt to equalise difficulty across sets. Importantly, the main effect for testing was present in separate analyses of performance on each of the sets (not reported here).

Feedback analyses

Although the final test results suggest there was no effect of collaboration, a visual comparison of the right and left panels of Figure 1 suggests that collaboration and feedback operate differently at the initial and final tests. Figure 1 suggests that for the Group condition, there is a decrease in performance on the final test (relative to the initial test), although less so for feedback items. For the Individual Condition, there appears to be an increase in performance on the final test, but only for the feedback items. A complete exploration of this pattern would necessitate a 2 (Collaboration) × 2 (Feedback) × 2 (Initial, Final Test) analysis, with expectation of a three-way interaction. However, this analysis, which would have allowed us to explore possible processes occurring during collaborative test-taking, could not be undertaken; including the Initial Test group data violates assumptions of independence of data points, because each 'group' student would need to be assigned their group score. Thus, future experimentation with different designs may allow more detailed examination of the processes occurring during collaborative test taking and later retrieval.

We were, however, able to examine the effect of feedback on the initial test on performance of *individuals* in the final test. Analyses using a (2) (Initial test: correct *vs*. incorrect) × (2) (Feedback: Feedback *vs*. No Feedback) repeated measures ANOVA, examined how individuals' response outcome on the initial test influenced the production of correct responses on the final test, as a function of feedback (see Figure 2). There was a main effect of initial test response, $F\ (1, 46) = 309.22$, $p < 0.05$, $p\eta^2 = 0.87$, such that initially correct responses were more likely to be correct on the final test than initially incorrect responses. There was also a significant interaction such that receiving feedback had no effect on the proportion of initially correct items that remained correct, compared to not receiving feedback; however, feedback led to better final test performance of initially incorrect items, than did no feedback, $F(1, 46) = 6.23$, $p < 0.05$, $p\eta^2 = 0.12$.

Study time

In order to determine whether study time between the initial and final tests influenced performance (a potential indirect effect of the testing procedure; Roediger & Karpicke,

DOI: 10.1002/acp

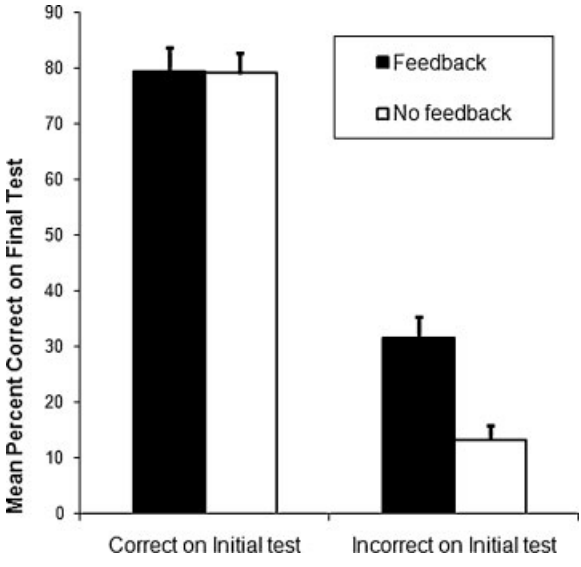

Figure 2. Mean proportion correct in the final test as a function of initial test response (correct–incorrect) and feedback, for individual participants only (error bars are standard error of the mean)

2006a), we examined whether there was a difference in the reported number of hours studied by the students who did the initial test in groups and those who did so as individuals. An independent samples *t*-test comparing Groups ($M = 0.81$ hours, $SD = 1.10$) to Individuals ($M = 0.52$, $SD = 0.69$) showed no significant difference between the two, $t(104) = 1.56$, $p > 0.05$, $d = 0.31$.

DISCUSSION

This study aimed to expand research on the testing effect in a classroom setting, particularly by examining the effects of feedback and collaborative testing during an initial test on the final individual test performance. As hypothesised, feedback led to better long-term retention than no feedback, which in turn yielded better long-term retention than no initial testing. We expected that collaboration would lead to better performance on both the initial and final tests, replicating Cranney et al. (2009). However, although better performance was found on the initial test, there was no difference on the final test.

The testing effect

Our study demonstrates the testing effect in a classroom setting, with better memory for items that were tested (but did not receive feedback), compared to untested items, thus meeting the 'lenient criterion' for the testing effect. The 'strict criterion' for the testing effect requires students in the control condition to read the material rather than being tested during the second phase; such a condition was not employed in this study (*cf.* Cranney et al., 2009). This finding replicates previous experimental work in both tertiary educational environments (McDaniel et al., 2007) and in the laboratory (e.g. Karpicke &

DOI: 10.1002/acp

Roediger, 2007; Roediger & Karpicke, 2006b). These findings reflect a direct, positive effect of repeated testing on memory, potentially mediated by repeated retrieval. Specifically, repeated effortful retrievals involve deeper processing and elaboration, which in turn increases the strength of the memory trace and the number of retrieval routes (Carpenter & DeLosh, 2006; Roediger & Karpicke, 2006a). In addition, transfer-appropriate processing (TAP) may play a role (Morris, Bransford, & Franks, 1977; McDaniel & Masson, 1985; Roediger & Karpicke, 2006a; Wheeler & Roediger, 1992). Practicing the skills needed for retrieving information in the initial test, enhances those skills when they are required in the final test. However, a *strict* TAP account would predict an advantage when the initial and final test situations overlapped (i.e. when a participant was tested individually at both times.) There is no indication in the results of such a benefit—that is, no advantage in the final (individual) test for those tested individually in Phase 2 (see Figure 1). The absence of such an effect suggests that elaborative encoding and its effect on strengthening memory traces and increasing retrieval routes might have had the greater influence in our experiment.[1]

Feedback

Previous laboratory research has reported that feedback after testing leads to better long-term retention than does no feedback (Butler, Karpicke, & Roediger, 2007; Butler & Roediger, 2008; Kang, McDermott, & Roediger, 2007; Pashler et al., 2005; *cf*. Butler & Roediger, 2007). We extended this laboratory research by finding similar results in a classroom setting using a cued recall test. The finding that feedback significantly improved individual performance on the final test for initially incorrect items supports the notion that feedback provides an opportunity for correction of these errors (Butler et al., 2007; McDaniel et al., 2007). This indicates that the positive effect of feedback may be due more to the indirect mechanism of error correction than to direct effects of the testing procedure.

For initially correct items, feedback did not lead to improved individual performance on the final test compared to the initial test. Although it is not possible for items that are correct on the initial test to become 'more correct' during the final test, it is possible for such items to become incorrect on the final test, as a result of forgetting or lack of encoding into long-term memory (see Figure 2, where the mean for initially correct items is no longer 100%). Thus, it is possible that percentage correct for initially correct items could drop on the final test, and that this drop could be mitigated by the feedback procedure (see Butler & Roediger, 2007; McDaniel et al., 2007, Table 4). Butler, Karpicke, and Roediger (2008) have suggested a role for confidence as a mediator in such effects—specifically, that participants' highly confident correct responses are impervious to feedback. The current study did not measure confidence in the initial test; future classroom studies could examine the proposition that low confidence corrrect responses could benefit from feedback (see Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Kulhavy & Stock, 1989).

[1]One might argue that a TAP *disadvantage* was shown whereby those tested in groups initially fare worse in the final (individual) test than those tested as individuals. Though a comparison of the left and right sides of Figure 1 suggests this pattern, as we note in the Results section, such a comparison should be cautioned against because of the problem of assigning the same group score to every individual for the initial test. The simplest conclusion one can draw is that collaboration (individual *vs*. group) in the initial test had no effect on performance in the final test.

DOI: 10.1002/acp

Collaborative testing

In contrast to Cranney et al.'s (2009) findings, collaborative testing did not lead to an advantage in the final test, despite an advantage in the initial test. This finding could be explained in terms of the indirect effects of collaborative inhibition during the initial test, as suggested by the trend towards superior nominal group performance compared to collaborative group performance (Weldon & Bellinger, 1997). This may constitute one reason for the lack of replication of Cranney et al. (2009); specifically, there may have been a higher level of collaborative inhibition in the current study. Unfortunately a nominal group analysis was not undertaken in that previous study, so we cannot test that hypothesis. If one assumes there was a difference in the amount of collaborative inhibition, this may have been due to differences in group dynamics, in particular, group cohesion. In the Cranney et al. (2009) study, the groups had previously been formed and had already undertaken some tasks together, while in the current study, the groups were formed for the first time just prior to the initial test. Thus, students may have been more hesitant in offering answers, or may have been more likely to engage in social loafing (Karau & Hart, 1998). Future studies could attempt to record group interaction, and manipulate the way in which students interact in the group in order to reduce collaborative inhibition. For example, in one group condition participants could take turns recalling what they remember, without being able to access what the others have previously recalled, whereas in the other group condition there could be no control over recall procedure (see Wright & Klumpp, 2004, regarding the nature of collaborative inhibition).

Recent research by Blumen and Rajaram (2008) compared collaborative and individual retrieval over three recall phases. Interestingly, although they did not find a benefit of initial test collaboration on the second recall stage (equivalent with our final test), they did find that collaboration at either the second test, or both first and second tests, led to better recall at a third individual test than did two prior individual tests. Hence, other laboratory research suggests initial collaboration can have a positive effect in some situations. Future research could attempt to extend the findings of Blumen and Rajaram (2008) to a classroom setting.

A further consideration is the level of initial test performance, which was much higher in Cranney et al. (2009), and the maintenance of the difference between Group and Individual conditions from Initial to Final Test (A. Butler, personal communication, 12 February 2009). That is, the greater initial difference was maintained in the Cranney et al. (2009) study but not in the current study. Future exploration of this notion requires a direct manipulation (e.g. task difficulty) that would produce significant differential performance in the initial test.

Implications for education and training

Test-taking is common in educational institutions and in organisational training and development. However in educational settings, testing is often used as a means of summative assessment, rather than as a tool for learning (formative assessment), and so corrective feedback is not always given to students. Our results clearly suggest that not only is testing an important way of encouraging long-term retention, but feedback after testing is important because it allows learners to correct their memory errors. This means that in order to encourage accurate long-term retention, educators should include testing with feedback as an essential part of their teaching strategy. Teamwork is often used in

DOI: 10.1002/acp

commercial organisations and increasingly in educational institutions, and our results on collaborative testing have implications for learning in a group context. Specifically, although groups perform better than individuals at initial recall, this does not necessarily translate to later recall. This may be because of collaborative inhibition, or because members of a group inaccurately judge their learning based on the initial group performance, leading them to think they know more than they do, and potentially put less effort into remembering.[2] Clearly, further research on collaborative testing is required to determine what factors lead to different outcomes for long-term retention.

CONCLUSIONS

The present study found evidence of the testing effect in a tertiary classroom setting, replicating studies in this area (e.g. Cranney et al., 2009; McDaniel et al., 2007). Feedback was important for long-term retention because it allowed students to correct misinformation or to acquire the correct information. Although collaboration led to better performance on the initial test, this was not the case for the final test, with groups scoring similarly to individuals. These findings are important as they support the potential for laboratory findings to be generalised to the reality of the classroom setting (Jaffe, 2008; Worrell et al., 2009) and point to directions for further research on the testing effect, feedback and collaboration.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support of the School of Psychology, the input of the tutors and students in the first year psychology course at the University of New South Wales, and the reviewers' constructive comments.

REFERENCES

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research*, *85*, 89–99.

Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*, 213–238.

Basden, B. H., Basden, D. R., Bryner, S., & Thomas, R. L. III, (1997). A comparison of group and individual remembering: Does collaboration disrupt retrieval strategies? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 1176–1191.

Blumen, H. M., & Rajaram, S. (2008). Influence of re-exposure and retrieval disruption during group collaboration on later individual recall. *Memory*, *16*, 231–244.

Butler, A. C., Karpicke, J. D., & Roediger, H. L. III, (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, *13*, 273–281.

[2]Prior to the final test, all participants were asked to estimate how well they would perform in the test. Although there were no significant differences between the Group and Individual conditions in these 'judgments of learning' (Nelson & Dunlosky, 1991), the correlation with the total percentage correct was significant for Individual participants ($r = .33$) but not for Group participants ($r = .07$). It is possible that when students made a judgment of learning based on previous collaborative activities, they displayed overconfidence regarding their current memory for the material.

DOI: 10.1002/acp

Butler, A. C., Karpicke, J. D., & Roediger, H. L. III, (2008). Correcting a metacognitive error: Feedback increases retention of low confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 918–928.

Butler, A. C., & Roediger, H. L. III, (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*, 514–527.

Butler, A. C., & Roediger, H. L. III, (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*, 604–616.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Psychology*, *65*, 245–281.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, *34*, 268–276.

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. III, (2006). Retrieval induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, *135*, 553–571.

Cranney, J., Ahn, M., McKinnon, R., Morris, S., & Watts, K. (2009). The testing effect, collaborative learning, and retrieval-induced facilitation in a classroom setting. *European Journal of Cognitive Psychology*, *21*, 919–940.

Glover, J. A. (1989). The ''testing'' phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*, 392–399.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*, 81–112.

Jaffe, E. (2008). Will that be on the test? *Observer*, *21*, 18–21.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. III, (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*, 528–558.

Karau, S. J., & Hart, J. W. (1998). Group cohesiveness and social loafing: Effects of a social interaction manipulation on individual motivation within groups. *Group Dynamics: Theory, Research, and Practice*, *13*, 185–191.

Karpicke, J. D., & Roediger, H. L. III, (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*, 151–162.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: An historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*, 254–284.

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review*, *1*, 279–308.

Latane, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology*, *37*, 822–832.

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology*, *29*, 210–212.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*, 494–513.

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology*, *16*, 192–201.

McDaniel, M. A., & Masson, M. E. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 371–385.

McKeachie, W. J. (1963). Research on teaching at the college and university level. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 1118–1172). Chicago: Rand McNally.

Meltzoff, J. (1998). *Critical thinking about research: Psychology and related fields*. Washington, DC: American Psychological Association.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning & Verbal Behaviour*, *16*, 519–533.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The ''delayed-JOL effect''. *Psychological Science*, *2*, 267–270.

Pashler, H., Cepeda, N. J., Wixted, J., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 3–8.

DOI: 10.1002/acp

Roediger, H. L., III, & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.
Roediger, H. L., III, & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255.
Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *31*, 1155–1159.
Roediger, H. L., III, Meade, M. L., & Bergman, E. T. (2001). Social contagion of memory. *Psychonomic Bulletin & Review*, *8*, 365–371.
Sainsbury, E. J., & Walker, R. A. (2008). Assessment as a vehicle for learning: extending collaboration into testing. *Assessment & Evaluation in Higher Education*, *33*, 103–117.
Sassenrath, J. M., & Garverick, C. M. (1965). Effects of differential feedback from examinations on retention and transfer. *Journal of Educational Psychology*, *56*, 259–263.
Shute, V. (2008). Focus on formative feedback. *Review of Educational Research*, *78*, 153–189.
Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*, 641–656.
Weldon, M. S., & Bellinger, K. D. (1997). Collective memory: Collaborative and individual processes in remembering. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 1160–1175.
Wheeler, M. A., & Roediger, H. L. (1992). Disparate effects of repeated testing: reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science*, *3*, 240–245.
Worrell, F. C., Casad, B. J., Daniel, D. B., McDaniel, M., Messer, W. S., Miller, H. L., Jr., et al. (2009). Promising principles for translating psychological science into teaching and learning. In D. F. Halpern (Ed.), *Undergraduate education in psychology: A blueprint for the future of the discipline* (pp. 129–144). Washington, DC: American Psychological Association.
Wright, D. B., & Klumpp, A. (2004). Collaborative inhibition is due to the product, not the process, of recalling in groups. *Psychonomic Bulletin and Review*, *11*, 1080–1083.

DOI: 10.1002/acp