Group Discussion and Large-Scale Language Arts Assessment: Effects on Students' Comprehension

CSE Technical Report 445

Randy Fall and Noreen Webb
CRESST/University of California, Los Angeles

Naomi Chudowsky
U.S. Department of Education
Office of Educational Research and Improvement

August 1997

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90024-6511
(310) 206-1532

The work reported herein was supported in part by the Academic Senate on Research, Los Angeles Division, University of California; and in part under the Educational Research and Department Center Program cooperative agreement R117G10027 and CFDA Catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

GROUP DISCUSSION AND LARGE-SCALE LANGUAGE ARTS ASSESSMENT: EFFECTS ON STUDENTS' COMPREHENSION[1]

Randy Fall and Noreen Webb
CRESST/University of California, Los Angeles

Naomi Chudowsky
U.S. Department of Education

Abstract

Large-scale assessment programs are beginning to design group assessment tasks in which small groups of students collaborate to solve problems or complete projects. Little is known, however, about the effects of collaboration on students' cognitive processes and performance on such tests. The present study compared student performance on language arts tests in which they either were or were not permitted to discuss the story they were required to read and interpret. The analyses compared the quality of student responses on test forms with and without collaboration, examined qualitative changes in students' responses before and after collaboration, and examined students' reflections about the impact of collaboration on their understanding of the story. The results show that a 10-minute discussion of the story in three-person groups had a substantial impact on student performance. Implications for the design and interpretation of large-scale testing with collaboration are discussed.

---

[1] Randy Fall, Graduate School of Education & Information Studies, UCLA; Noreen M. Webb, Graduate School of Education & Information Studies, UCLA; Naomi Chudowsky, U.S. Department of Education Office of Educational Research and Improvement.

Most of the analyses reported in this paper were conducted while the third author was a Consultant with the Student Assessment Office at the Connecticut State Department of Education. However, all statements in this paper are those of the authors and do not necessarily reflect the views of the Connecticut State Department of Education or of the Center for Research on Evaluation, Standards, and Student Testing.

We would like to thank Cecile Alacayan, Annie Minas, Humberto Iglesias, and Jessie Sziebl for their help in test coding.

A version of this paper was presented at the annual meeting of the American Educational Research Association in San Francisco in April 1995.

Correspondence concerning this article should be addressed to Randy Fall, 160 S. Michigan Ave. #204, Pasadena, CA 91106; email: RFall@ucla.edu.

Introduction

Large-scale assessment programs are increasingly starting to include collaborative small-group work (e.g., Connecticut's Common Core of Learning Assessment: Baron, 1994, Connecticut State Board of Education, 1987, Lomask, Baron, Greig, & Harrison, 1992; California Assessment Program: Awbrey, 1992, Bartlett, 1992, Pandey, 1991; California Learning Assessment System: Saner, McCaffrey, Stecher, Klein, & Bell, 1994; Oregon State Department of Education: Neuberger, 1993; Shavelson & Baxter, 1992). Recommendations to include small-group work on tests have also started appearing in efforts toward developing state and national standards for assessment (e.g., Kansas State Board of Education, 1993; Mathematical Sciences Education Board, National Research Council, 1993). While incorporating group work into achievement tests is becoming more widespread, little systematic research on the effects of collaboration on performance has been carried out. A major unanswered question is whether the opportunity to work with other students during a test improves student performance compared to purely individual work. The purpose of this study, then, is to examine in detail the influence of group collaboration on student performance on a language arts test at the secondary level.

Collaborative work is often incorporated into tests to help link assessment more closely to instruction (Linn, 1993; Wise & Behuniak, 1993). Collaborative small-group work is used in classroom instruction because it can increase student learning, self-esteem, and prosocial attitudes (Bossert, 1988; Slavin, 1990). Students can learn new ideas, skills, and knowledge by solving problems with others, by resolving disagreements due to different points of view, by giving help to other students, and by receiving help (Webb, 1995; Webb & Palincsar, 1996). Peer collaboration is often used in language arts instruction to improve reading comprehension and text recall, including, for example, Book Club, a literature-based reading program with student-led discussions (McMahon, 1992; Raphael et al., 1992), reciprocal teaching that incorporates the reading comprehension strategies of predicting, question generating, summarizing, and clarifying (Brown & Palincsar, 1989; Palincsar, 1986; Palincsar & Brown, 1984), and scripted cooperative work in which students alternatively engage in summarizing and active listening roles (Hythecker, Dansereau, & Rocklin, 1988).

Research on collaborative group work in language arts has shown several benefits of peer-group discussions of works of literature for classroom learning.

Engaging in group discussions helps students gain understanding of the meaning of a story (Noll, 1994; Reid et al., 1994; Leal, 1993; Eeds & Wells, 1989), helps students understand alternate points of view (Brown & Palincsar, 1989), helps students to make connections between a piece of literature and their own personal experience or prior knowledge (Reid et al., 1994; Leal, 1992), improves students' motivation to understand a piece of literature (Almasi, 1994; Noll, 1994), and helps to teach students that social interaction is a normal part of understanding literature (Samway et al., 1991).

Although much research shows that collaborative work in the classroom increases student learning (e.g., Slavin, 1990), the effects of tests with collaboration on student performance have rarely been studied. In particular, little is known about how collaboration during one part of an assessment influences performance on subsequent portions of a test that students work on individually, a popular structure of achievement tests with collaboration (e.g., Baron, 1994; Saner, McCaffrey, Stecher, Klein, & Bell, 1994; Wise & Behuniak, 1993). Saner et al. (1994) found that some students' performance improved after working in pairs on a science assessment developed by the California Learning Assessment System. In another study of science assessment, Webb, Nemer, Chizhik, and Sugrue (1996; see also Webb, in press) found that below-average students benefited more from working in three-person heterogeneous groups than working in homogeneous groups or alone, and that above-average students performed equally well whether they worked in homogeneous groups, heterogeneous groups, or alone. Wise and Behuniak (1993) found that students given an opportunity to collaborate obtained higher overall achievement test scores than students who did not have an opportunity to collaborate.

Previous studies of assessment in groups have compared scores of students working in collaborative groups with those of students working alone, but have not analyzed the changes in students' knowledge and understanding that occur as a result of group collaboration. The present study, therefore, carried out detailed analyses of students' responses from the Wise and Behuniak (1993) study of a large-scale statewide pilot assessment in language arts to determine the impact of collaboration on the nature of students' changes in performance and understanding of a piece of literature. The analyses examined students' understanding of the facts of the story, their interpretations of the events in the story, and their attitudes toward the story.

Method

Sample and Design

Approximately 5,000 10th-grade students from Connecticut public high schools participated in a pilot of a 90-minute language arts test, Response to Literature, designed to measure their ability to interpret a piece of literature, make connections to their own lives, and take a critical stance. Stratified random sampling was used to ensure that the sample represented the statewide population. Students were administered one of nine test forms that varied by story (three) and condition (three). The three conditions for each story were (a) discussion toward the beginning of the test, (b) discussion toward the end, and (c) no discussion. Approximately 500 students took each form. Stratified random sampling was also used to select the students who were administered each form.

On all test forms (three test forms for each story), students read the story for 40 minutes and then answered questions individually. Each form included six questions which were the same for all three forms for a story. Two forms per story gave students an opportunity to discuss the story in three-person groups for 10 minutes; the third form had no group collaboration.

On one form per story, the students read the story individually, answered the first two questions on the test individually and then engaged in a 10-minute discussion in three-person groups. Immediately after the small-group discussion, each student individually answered a question about how the group discussion affected their ideas about the story. Then they answered the remaining four questions on the test individually. This form was called "discussion toward the beginning" in the original report (Wise & Behuniak, 1993). On another form per story, the 10-minute discussion took place after students had answered the first four questions individually (called "discussion toward the end" in the original report). On this form, students read the story individually, answered four questions individually, engaged in the small-group discussion for 10 minutes, answered the question about how the group discussion affected their ideas about the story, and answered the remaining two questions on the test individually. On the third form per story, students read the story individually and then answered the six questions individually. This form was called "no discussion" in the original report.

The method for assigning groups was decided by the individual teacher. Some teachers divided students according to their place on an alphabetical list, others put together students who happened to be sitting near one another.

In the original analyses reported by Wise and Behuniak (1993), a random sample of 300 of the 500 responses for each test form was scored holistically on a 4-point scale. One score was assigned to the whole test based on the following factors: initial understanding, interpretation, critical stance, and connections. No scores were assigned to individual test questions. For the study reported here, detailed analyses of students' responses were performed for two forms for each of two of the three stories. For each story, we contrast the form with discussion toward the beginning (discussion after question 2) and the form with no discussion. The test form with discussion toward the end (discussion after question 5) was not analyzed because the discussion occurred too late in the test to have much impact on student performance: on this test form, students answered most of the questions on the test before the discussion took place. Because the effect of collaboration was minimal for this form, we decided to use the form with no discussion as the contrast with the form with discussion toward the beginning.

The current paper reports on the analyses of two of the three stories. The stories we analyzed are conceptually the most demanding of the stories; informal inspection of students' responses suggested that these stories would the most interesting material for student discussion and would provide more detailed responses to test questions than the other story. These two stories also provide an interesting contrast: one primarily required students to follow a complex plot, the other primarily required students to interpret a character's thoughts and feelings.

The present study analyzed the test papers of 504 students: 251 students from the first story "A Story of an Hour," and 253 students from the second story, "An Ordinary Woman." The mean holistic scores for the original samples and the scored tests from the samples used here are presented in Table 1. The means for the original samples and the samples analyzed in this study are very similar, showing that the samples analyzed here are representative of the original statewide population.

Table 1

Mean Holistic Scores[a] for the Original Statewide Samples and the Samples Analyzed Here

	Original statewide sample		Sample analyzed here	
	M	n	M	n[b]
Story 1: "Story of an Hour"				
Discussion toward the beginning	2.33	300	2.39	127
No discussion	2.08	300	2.00	124
Story 2: "An Ordinary Woman"				
Discussion toward the beginning	2.49	300	2.52	90
No discussion	2.50	300	2.49	105

[a] 4-point scale based on performance on the whole test (see text).

[b] Due to random sampling of tests for holistic scoring, holistic scores were not available for some tests analyzed in this paper. Consequently, sample sizes here are smaller than in subsequent tables.

Descriptions of the Stories

The first story we analyzed was "The Story of an Hour" by Kate Chopin, written and set in the late 1800s. The story begins with Louise being informed that her husband Brently was killed in a train accident. Louise's sister is careful to break the news to her gently, because of concerns about Louise's heart trouble. Louise is initially saddened by the news, and she locks herself in an upstairs room. As she sits in the room, Louise comes to realize that her husband's death has made her free, and she begins to feel joy in her new-found freedom. Meanwhile, her sister waits outside, concerned that Louise's grief will be too much for her to handle. As Louise finally accedes to her sister's entreaties and emerges from her room, the front door begins to open and suddenly Brently arrives, very much alive. Louise is so shocked at the sudden loss of her freedom that she dies. Ironically, the doctors later explain that Louise died of heart disease, "of joy that kills."

The most common misinterpretation of the story for students was to fail to see Louise's joy at her freedom, instead believing that Louise died because she was so sad about her husband's death, or so happy to see him alive again. Other common difficulties were uncertainty about who had died at the end (Louise),

whether Brently was truly dead, and misinterpretations about what was happening to Louise as she sat alone in the room upstairs.

The second story we analyzed was "An Ordinary Woman" by Bette Green. This story consists almost exclusively of one woman's thoughts as she prepares for her day. The sad events of the woman's past are slowly revealed in her thoughts as the story progresses.

The story begins as a teacher, Mandy Brooks, is calling her school to let them know that she may be late for work. She is waiting for a locksmith to come. The woman thinks about her daughter, who she recently discovered to be a junkie. Mrs. Brooks is reminded of her husband's death 22 months earlier. Once the locksmith arrives, the woman stops at her daughter's room, and stares at the burn in the rug and the soot on the furniture that resulted from the fire the previous night. She recalls a fireman's comment that she had been lucky that the fire had not spread to the mattress, and muses over her reputation for being lucky in contrast to the dismal events in her life. She thinks about her daughter Caren, and what she might have done wrong as a mother. She hopes that if Caren returns and finds the lock changed that Caren will know that her mother is not barring Caren, but what Caren has become. When the locksmith has finished changing the lock, he comments that Mrs. Brooks is lucky that she won't be late for school. Mrs. Brooks responds, "People have always said that about me."

Unlike the first story, students only rarely expressed misconceptions about the facts of this story. Instead, students were challenged by the task of interpreting the woman's feelings and the meanings of her actions.

Test Questions

For both stories, the first question on the test was very open, asking students to write their first impressions about the story. The second question was more substantive: for the first story, asking students to describe an event in the story and explain its importance, and for the second story, asking students how Caren feels about her mother. At this point, after the first two questions, students using the form that included discussion were divided into groups to discuss the story. After the discussion, these students answered a question about any effects the group discussion had on their understanding of the story. All of the other questions were identical for the two forms. For the first story, the fourth question on the test (the third question for the no-discussion form) asked whether Louise undergoes

any changes during the story. For the second story, the fourth question asked about problems or conflicts the main character is experiencing. The fifth, sixth, and seventh questions were similar for both stories. The fifth question asked students to choose one of three quotes and explain its meaning in the story. The sixth question asked students to discuss any similarities between the story and their own experiences, or similarities to other books, movies, television shows, etc. The final question asked students to create a definition of "good" literature, and state whether this story meets that definition. Tables 2 and 3 present all of the questions as they were given on the test.

Coding of Student Performance

The original scoring of the test by the Connecticut State Department of Education assigned a single holistic score to each test, focused on the students' ability to interpret, make connections to the short story, and take a critical stance. This paper coded student performance at a more detailed level. For example, in the coding of interpretations, the Connecticut holistic scorers focused on the presence or absence of an interpretation of the story, while we coded every interpretation given by each student and made distinctions among levels of quality of different possible interpretations of the story.

We coded students' responses on the tests on the following five variables: (a) factual knowledge, (b) interpretations, (c) attitude toward the story, (d) evidence of an effect of group discussion, and (e) type of change. For each test, factual knowledge, interpretations, and attitudes were coded twice: once for the questions before discussion (questions 1 and 2) and again for questions after discussion (questions 3-7 or 3-6 on the forms with and without discussion, respectively). For the test forms with discussion we coded a sixth variable: self-reported change in understanding as a result of the group discussion. This information came from the question immediately following the group discussion, asking students specifically if the group discussion had affected their understanding of the story.

Facts. Preliminary analyses of the first story revealed three key facts which indicated levels of understanding of the story. For this story, the facts coded were: (a) Brently is reported dead; (b) Brently isn't really dead; (c) At the end of the story, Louise is dead. For the second story four facts were coded: (a) Caren uses drugs; (b) Mandy's husband (Steve) died; (c) Caren's drug use led to the fire; and (d) Mandy Brooks is changing the locks to keep Caren out.

Table 2

Questions on Story 1: "Story of an Hour"

Item number		
Discussion	No discussion	Question
1	1	What is your first reaction to this story? Write down any thoughts, opinions or questions you may have.
2	2	There are several events in the story. Choose **one** event that stands out for you and explain why you think that event is important.
3	none	How did the group discussion affect your ideas of the story? If your ideas about the story changed, explain how they changed.
4	3	Think about the main character in this story, Mrs. Mallard. Does she change from the beginning of the story to the end? If you think she does, describe the changes and try to explain what they mean. If you don't think she changes, explain what makes you think so.
5	4	Choose **one** of the following quotations from the story to write about. Explain what you think it means about the characters in the story as well as people in general. A. "There was something coming to her and she was waiting for it, fearfully." (p. 5) B. "But she saw beyond that bitter moment a long procession of years to come that would belong to her absolutely." (p. 6) C. "When the doctors came they said she had died of heart disease—of joy that kills." (p. 6) Circle the letter of the quotation you are writing about: **A B C**
6	5	What does "The Story of an Hour" remind you of? other books or stories you have read? movies? television programs? other people? Discuss why this story reminds you of these other things. If the story doesn't remind you of other things in your life, explain why it doesn't.
7	6	Should this story be considered "good" literature? Briefly make up your own definition of what makes a piece of literature "good," and then explain how this story does or does not fit your definition.

Table 3

Questions on Story 2: "An Ordinary Woman"

Item number		
Discussion	No discussion	Question
1	1	What is your first reaction to this story? Write down any thoughts, opinions or questions you may have.
2	2	Caren, Mrs. Brooks' daughter, never actually appears in the story but the author provides clues as to what she may be like. How do you think Caren feels about her mother? What makes you think so?
3	none	How did the group discussion affect your ideas about the story? If your ideas about the story changed, explain how they changed.
4	3	Describe the problems or conflicts the main character in the story is experiencing.
5	4	Choose **one** of the following quotations form the story to write about. Explain what you think it means about the characters in the story as well as people in general. A. "Probably very early on, I should have warned you that your mother was a very ordinary woman with not a single extraordinary power to her name." (p. 8) B. "Understand, at least, that I am not barring you, but only what you have become." (p. 8) C. "'You know, you're a lucky lady, Mrs. Brooks,' he says, dropping a single brass key into my hand . . . . People have always said that about me." (p. 9) Circle the letter of the quotation you are writing about: **A B C**
6	5	What does the author say about human nature? Think about **your** world—the people you know and the experiences you've had. In what ways does this short story relate to your world and experiences? Explain why. If the story doesn't relate to your world, explain why it doesn't.
7	6	Should this story be considered "good" literature? Briefly make up your own definition of what makes a piece of literature "good," and then explain how this story does or does not fit your definition.

Students usually stated the facts of the story as a part of their answers, e.g., for Story 1, "I was surprised by Louise's reactions to Brently being reported dead." Oblique references to facts were only coded when the fact was inescapable from the statements. For example, "Brently was reported dead, but he comes in the door at the end" was coded as an indication of knowledge of both 1 and 2. The number of facts that each student included was coded separately for "before discussion" (called Part 1 of the test) and "after discussion" (called Part 2 of the test).

Interpretations. In our coding, interpretations referred to students' understanding of the feelings and motives of the characters or the meaning of the story. Of the 14 interpretations coded for the first story, 6 were correct. Four of the six were central to understanding the story and two were correct but not essential elements of the story. For the second story, 15 interpretations were coded. Thirteen of the interpretations were correct, and 11 of these were central to the story.

The central interpretations were given twice as much weight (+2) in the analyses as the correct but not essential interpretations (+1). All incorrect interpretations were equally important and were given equal weights in the analyses (-1). For example, one central interpretation in Story 1 was, "Louise is initially very upset about her husband's death, but quickly becomes happy about it, knowing she will only have to live for herself in the future." A correct, but not central interpretation was, "Louise is sad about her husband's death but realizes she will have to go on/sees a positive side." Erroneous interpretations included, "Louise dies because she is so sad about her husband's death," and "What is happening upstairs is Louise dying." Complete lists of coded interpretations are presented in Tables 4 and 5.

For each student, all of the interpretations were coded for each question. If a student gave the same interpretation in more than one question before discussion, it was only counted once. Any interpretation repeated in several questions after discussion was also only counted once. The number of different interpretations coded for any individual student ranged from a maximum of six (three before discussion and three after), to a minimum of two (one before discussion and one after).

Table 4

Coding of Interpretations in Story 1: “Story of an Hour”

Correct and central to understanding the story (coded +2)
1. Louise is initially very upset about her husband’s death, but quickly becomes happy about it, knowing that she will only have to live for herself in the future.
2. (Upstairs in her room) The something coming at Louise was the feeling of freedom.
3. Having experienced (1) (5) or (6), Louise dies of shock at the loss of her freedom.
4. Story reflects the time is was written.
Correct but not central interpretations (coded +1)
5. Louise is happy about her husband’s death.
6. Louise is sad about her husband’s death but realizes she will have to go on / sees a positive side.
Incorrect interpretations (coded -1)
7. (Upstairs in her room) The something coming at Louise was death.
8. (Upstairs in her room) The something coming at Louise was God.
9. (Upstairs in her room) Louise goes up to heaven /dreams she goes to heaven.
10. What’s happening upstairs was Louise dying/ accepting her death.
11. Louise dies because she is so sad/upset about her husband’s death / happy he came back.
12. Louise is sad about her husband’s death.
13. Don’t know what was happening upstairs.
14. “Joy that kills” means strength and faith in herself.

Table 5

Coding of Interpretations in Story 2: "Ordinary Woman"

Correct and central to understanding the story (coded +2)
1. Mandy Brooks never really dealt with Steve's death / Mandy Brooks is grieving about Steve's death.
2. Mandy Brooks hides her feelings.
3. Steve's death, or Mandy Brooks' lack of attention led to Caren's drug use.
4. Caren feels neglected, disappointed, unloved by Mandy Brooks.
5. Mandy Brooks isn't sure what she did wrong with Caren.
6. Mandy Brooks loves, cares about Caren.
7. Caren loves, cares about Mandy Brooks.
8. Mandy Brooks is alone, feels lonely.
9. Mandy Brooks blames herself for all the problems.
10. Caren doesn't love / doesn't care much about / resents / hates Mandy Brooks.
11. Mandy Brooks feels unlucky.
Correct but not central interpretations (coded +1)
12. Mandy Brooks tries to avoid dealing with things / refused to deal with Caren's drug problem.
13. Mandy Brooks shouldn't lock Caren out / should put Caren in rehab.
Incorrect interpretations (coded -1)
14. Caren feels suffocated / Mandy Brooks is overprotective toward Caren.
15. Other incorrect

Attitudes. Each student's attitude toward the story (if any) was also coded, once for answers before discussion and again for the answers after discussion. Students' attitudes were coded as 1 (positive), for statements such as "I liked this story," 0 (no attitude given), or -1 (negative), for statements like, "This story sucked." No students expressed both positive and negative attitudes toward the story in the questions before discussion or in the questions after discussion. The coding in this category was clearly affected by the questions on the test. Only the final question in the test asked for the student's appraisal of the story, "Should this story be considered good literature?," so while only a few students expressed an opinion about the story before discussion (19% of students for Story 1, 24% for Story 2), most students expressed an opinion after discussion (84% of students for Story 1, 81% for Story 2).

Self-reported changes in understanding. We also coded whether students made any statements about changes in their understanding of the story. This coding primarily pertained to the forms of the test with discussion, since the question following discussion asked specifically about any changes in understanding (see Tables 2 and 3). Both general and specific responses were coded as self-reports that the discussion did have an effect, ranging from "Now I get it!" to "I didn't know Brently Mallard wasn'ts dead."[2] The category of self-reported changes was coded simply yes or no for the presence of self-reports of an effect of group discussion.

Evidence of change in facts and interpretations. As a higher level of evidence of an effect of group discussion on students' understanding, we coded evidence of change in facts and interpretations from the questions before discussion to the questions after discussion. Evidence of change in facts and interpretations was defined more narrowly and more stringently than the self-report categories, and required two conditions. Evidence of change in facts and interpretations was coded only when (a) there was an improvement in facts or interpretations from the first part of the test to the second part of the test, and (b) that fact or interpretation had been addressed by the student in questions both before and after discussion. This category, then, does not count as change those situations in which students simply chose to address different issues before and

[2] All quotes retain students' grammar, spelling, and punctuation.

after discussion, and thus presented different facts or interpretations before and after.

One example of evidence of change in facts and interpretations in the first story comes from a student who was initially unsure about what had happened, "Did Mrs. Mallard really die? Why? Why was Mr. Mallard's name on that list if he didn't even get hurt?" After discussion, the student acknowledged that the group discussion did affect his understanding of the story: "The discussion cleared my thoughts about the story. My group didn't think the same way about the story that I did. We all discussed it thoroughly now the story is more clear." The improvement is evident in later answers, as he now understands that Louise died (although he still misinterpreted Mrs. Mallard's emotions), "When her husband came through the door she was so happy that she had a heart attack."

Interrater agreement. Four raters coded a random sample of 35 tests from Story 2. For each variable coded on a quantitative or ordered scale (number of facts, number of correct and incorrect interpretations, attitude toward the story), a generalizability analysis was conducted to assess the degree of agreement among raters' coding. Each analysis produced an estimated index of dependability, a reliability-like coefficient that showed the consistency of raters in coding students' absolute performance. This is a stricter indicator of rater agreement than is a correlation among raters which indicates only the relative standing of students (see Brennan, 1992; Shavelson & Webb, 1991). Consistency among raters was substantial: the estimated indices of dependability ranged from .73 to .95 across the student performance variables. Because the variable "evidence of change in facts and interpretations" was categorical (whether a student showed an improvement from the first part of the test to the second part in knowledge of facts or interpretations or both), rater agreement was assessed using the average percent of exact agreement among all pairs of raters. Average exact rater agreement was 82%.

Analyses Comparing Test Forms With and Without Discussion

The analyses compared student performance on the test forms with and without discussion. As described earlier, the first two and last four questions on the test were identical on the two test forms for each story, but the form with discussion had an additional question concerning the effects of group discussion on students' ideas about the story (question 3: see Tables 2 and 3). The question

arises about whether students' responses to the additional question should be included in the analyses of student performance. It could be argued that including the additional question may create a difference in results between the two test forms. For example, if the additional question elicited interpretations of the story that would not have emerged in responses to the other questions on the test, students' interpretation scores on the form with discussion could be higher than interpretation scores on the form without discussion (if the interpretations elicited by question 3 were correct) or they could be lower (if the interpretations elicited by question 3 were incorrect). On the other hand, excluding the question could also create differences between the two test forms. For example, once students discussed particular interpretations in question 3 (correct or incorrect), they may have decided not to repeat them in their responses to subsequent questions. Excluding the additional question could result in the omission of important information about students' understanding (or misunderstanding) of the story.

Given the complexity of the issue, all analyses were carried out twice: with and without question 3. In all of the comparisons across the two forms of the test, the results of the two analyses were identical or nearly identical. Thus, to preserve all of the data available, the results are presented with the additional question included.

Results

The original analyses of the entire statewide sample showed that, for the first story, overall performance on the test (using a 4-point holistic scale) was higher when students discussed the story than when students did not; for the second story, holistic scores were nearly the same on test forms with and without discussion (Table 1; see also Wise & Behuniak, 1993). The detailed analyses presented here corroborate the overall finding for the first story and provide important insights into the effects of collaboration on students' understanding of both stories.

Facts and Interpretations

Table 6 presents the means and standard deviations for number of correct facts, number of correct interpretations, number of incorrect interpretations, and the weighted sum of all of a student's interpretations. For Story 1, as can be seen in Table 6, students who had an opportunity to discuss the story improved in

Table 6

Test Performance on Test Forms With and Without Discussion

	Form with discussion[a]				Form without discussion[b]			
	Part 1[c]		Part 2[d]		Part 1[c]		Part 2[e]	
Performance variable	M	SD	M	SD	M	SD	M	SD
Story 1: "Story of an Hour"								
Number of correct facts	2.20	.92	2.51	.76	2.00	1.05	1.89	.96
Number of correct interpretations	.65	.83	1.00	.82	.52	.80	.59	.75
Number of incorrect interpretations	.40	.58	.35	.55	.40	.60	.46	.63
Average level of interpretations[f]	.35	1.16	.85	1.22	.18	1.04	.34	1.21
Story 2: "An Ordinary Woman"								
Number of correct facts	1.93	.87	2.24	.85	1.45	.80	1.95	.90
Number of correct interpretations	1.97	1.20	1.98	1.24	1.60	.96	1.29	1.02
Number of incorrect interpretations	.18	.43	.07	.26	.10	.30	.06	.24
Average level of interpretations[f]	.84	.58	1.11	.59	.94	.62	.93	.73

[a] 10-minute discussion took place after students responded to the first two items on the test. $\underline{n} = 127$ (Story 1), $\underline{n} = 123$ (Story 2).

[b] $\underline{n} = 124$ (Story 1), $\underline{n} = 130$ (Story 2).

[c] First two items, see text.

[d] Last five items, see text.

[e] Last four items, see text.

[f] Weighted average of interpretations: each interpretation was scored as 2 = correct and central to understanding the story; 1 = correct but not central to understanding the story; 0 = irrelevant interpretation or no interpretation; -1 = incorrect interpretation.

presentation of facts and interpretations from the first part of the test (Part 1: first two items) to the second part of the test (Part 2: last five items) more than did students who had no opportunity to discuss the story. First, students who discussed the story stated more correct facts on the items after discussion than on the items before discussion, whereas the reverse was true for students who did not discuss the story. Mean scores on the two forms (with and without discussion) were compared statistically using analysis of covariance with scores on the second part of the test as the outcome measure and performance on the first part of the test as the covariate. The difference between forms was statistically significant ($F(1,248) = 30.09$, $p<.0001$). Second, students who discussed the story stated more correct interpretations on the items after discussion than on the items before discussion; students who did not discuss the story showed little change from the first part to the second part of the test ($F(1,248) = 19.50$, $p<.0001$). Third, students who discussed the story decreased the number of incorrect interpretations from the first part to the second part of the test, whereas the reverse was true for students who had no opportunity to discuss the story, but the difference between forms was not statistically significant ($F(1,248) = 2.60$, $p=.11$). Finally, the average level of interpretations of students in the discussion condition improved more than that of students in the no-discussion condition ($F(1.248) = 10.48$, $p=.001$).

For Story 2, as can be seen in Table 6, students on both forms showed comparable increases in the number of correct facts from the first part to the second part of the test ($F(1, 250) = 1.90$, $p=.17$). The two forms showed significant differences for interpretations, however. Students who discussed the story gave a similar number of correct interpretations on the first and second parts of the test, while students who did not discuss the story gave fewer correct interpretations on the second part than on the first part ($F(1, 250) = 16.98$, $p<.0001$). Furthermore, students who discussed the story showed an increase in the level of interpretations from the first part of the test to the second, while students who did not discuss the story showed no increase from the first part of the test to the second ($F(1,250) = 5.38$, $p=.02$). The difference between forms in the number of incorrect interpretations was not statistically significant ($F(1, 250) = 0.06$, $p=.80$).

In summary, for Story 1, students who engaged in group discussion showed an increase in understanding from the first part of the test to the second part of the test whereas students who did not discuss the story showed a similar level of

understanding on both parts of the test. For Story 2, students who discussed the story showed an increase in understanding whereas students who did not discuss the story showed the same level of understanding or even a decrease in understanding from the first part to the second part of the test.

Analyses of specific kinds of improvement from the first part of the test (before discussion) to the second part of the test (after discussion) help to illustrate the impact of discussion on students' understanding. In Story 1, for example, on the first part of both test forms, a similar proportion of students showed partial but incomplete understanding of the facts of the story: 58 students who were administered the form with discussion (46% of the sample for that form) and 56 students who were administered the form without discussion (45% of the sample for that form) reported one or two facts of the story, but not all three facts of the story. Among the students who had an opportunity to discuss the story, more than half (32 of 58) reported all three facts of the story on the second part of the test. Only a few students who did not have an opportunity to discuss the story showed such an improvement (8 of 56).

A second example for Story 1 is the improvement in overall quality of interpretations from the first part to the second part of the test. On the first part of the test, 23 students who received the form with discussion (18% of the sample for that form) and 18 students who received the form without discussion (15% of the sample for that form) gave interpretations that scored a total of three points or higher, the equivalent of two correct interpretations, one central to understanding the story. On the second part of the test, the number of students scoring three points or higher increased by 61% (23 to 37) among students in the discussion condition, but decreased by 11% (18 to 16) in the no-discussion condition.

A more specific example is the appearance of a particularly important insight into the story: that Louise dies from the shock at the loss of her freedom. For example,

> I know strongly think that Mrs. Mallard died because she was not longer her own person.

Among students who had no opportunity to discuss the story, about the same number mentioned this insight on the first and second parts of the test (11 before, 10 after). Among students who had an opportunity to discuss the story, in

contrast, more students mentioned this insight after discussion (26) than before (17).

A final example from Story 1 of the effects of group discussion is its role in helping students to eliminate misconceptions and generate correct interpretations. As described above, the most common misconception among students in interpreting Story 1 was the belief that Louise was primarily sad about her husband's death, or that she died of joy when she saw him alive. Similar numbers of students held this misconception prior to discussion, 36 in the no-discussion condition, and 35 in the discussion condition. Of these students, 75% (27) of those in the no-discussion condition continued to hold this misconception, or generated another misconception after discussion, compared to only 43% (15) of those in the discussion condition. Also, of the students who held this initial misconception, 54% (19) of those in the discussion condition gave a correct interpretation after discussion, compared to only 19% (7) in the no-discussion condition.

In Story 2, there was no single common misconception, nor was there any single insight that reflected a deeper understanding of the story. Instead, the effects of discussion are seen in students' efforts to understand the characters' motives and feelings and to interpret the meaning of the story.

As described earlier with respect to Table 6, for the no-discussion condition for Story 2, students who did not discuss the story showed less evidence of understanding of the story (fewer correct interpretations) in the second part of the test than on the first part of the test. This finding does not mean that students understood the story less in the second part of the test; instead, it seems to reflect the fact that students often chose not to repeat information they had presented in prior answers. This reticence to repeat ideas was evident for students who discussed the story as well, but the students who discussed the story often added new interpretations of the story or revised their understandings, leading them to show higher levels of understanding in the second part of the test. The discussion often raised new issues to be discussed, even if students believed their understanding of the story had not changed. For example: "My ideas didn't really change. One of the others brought up that the daughter felt resentment towards her Mom because Caren was competing with her Mom's high school image. That may have brought about the drug use. Besides that my thoughts and theirs were basically the same."

Attitudes Toward the Story

Table 7 shows the frequencies of positive and negative attitudes for both forms of the test for both stories. For both stories, the majority of students expressed no opinion about the story in the first part of the test and expressed some opinion (either positive or negative) on the second part of the test. For Story 1, while some students shifted to a positive attitude on the second part of the test,

Table 7

Frequencies of Positive and Negative Attitudes Toward the Story on Test Forms With and Without Discussion

	Form with discussion[a]		Form without discussion[b]					
	Part 1[c]		Part 2[d]		Part 1[c]		Part 2[e]	
Attitude toward story	n	%	n	%	n	%	n	%
Story 1: "Story of an Hour"								
Positive	14	11	44	35	8	6	27	22
None	96	76	19	15	107	86	22	18
Negative	17	13	64	50	9	7	75	60
Story 2: "An Ordinary Woman"								
Positive	16	13	44	36	26	20	87	67
None	100	81	65	53	96	74	23	18
Negative	7	6	14	11	8	6	20	15

[a] 10-minute discussion took place after students responded to the first two items on the test; $\underline{n} = 127$ (Story 1), $\underline{n} = 123$ (Story 2).

[b] $\underline{n} = 124$ (Story 1), $\underline{n} = 130$ (Story 2).

[c] First two items, see text.

[d] Last five items, see text.

[e] Last four items, see text.

the greatest shift was to a negative attitude toward the story. The results were similar on both test forms: chi-square analyses of attitudes revealed no significant differences in students' attitudes toward the story between the discussion and no-discussion forms on either the first part of the test ($c^2(2)= 4.65$, $p=.10$) or on the second part of the test ($c^2(2)= 5.15$, $p=.08$). While students' self-reports (in question 3) make it seem likely that the discussions had some effects on some students' attitudes toward the story, the fact that the questions before discussion did not lead most students to express an opinion leaves us without much evidence of changes that may have occurred.

Students reading Story 2, in contrast, showed a shift toward more positive opinions on the second part of the test than on the first. This shift was especially marked on the test form with no discussion. The majority of students taking the no-discussion test form expressed positive opinions toward the story on the second part of the test. For Story 2, students' attitudes in the discussion and no-discussion forms were not significantly different on the first part of the test ($c^2(2)= 2.34$, $p=.31$) but were significantly different on the second part of the test ($c^2(2)= 35.05$, $p<.0001$). Here it is possible that students who liked the story initially may have been talked out of it by other members of their groups. The lack of clear evidence of students' attitudes toward the story prior to discussion makes further analysis difficult.

Evidence of Change in Facts and Interpretations

Evidence of change in facts and interpretations was our most strictly defined variable measuring the effects of group discussion. Evidence of change in facts and interpretations was coded only when a student's answers demonstrated a change in facts or interpretations before and after discussion, and that change was regarding an issue that the student addressed both before and after discussion. For Story 1, in the no-discussion condition, only two students (2 out of 124, or 1.6% of the sample for that condition) showed evidence of change in facts and interpretations. That is, two students gave evidence of a real change in understanding of issues in the story that they addressed in both part 1 (questions 1-2) and part 2 (questions 3-6). In contrast, nearly half of the students in the discussion condition (55 out of 127 students, or 43%) provided this level of evidence of change in facts and interpretations. The difference between proportions was statistically significant ($z=7.88$, $p<.0001$). Similarly, for Story 2, 17 students out

of 130 (13%) in the no-discussion condition showed this evidence of change, whereas nearly half of the students (55 out of 123 students, or 45%) showed evidence of change in facts and interpretations. The difference between proportions was statistically significant ($z=5.57$, $p<.0001$).

Table 8 gives the breakdown of types of change shown by students who had an opportunity to discuss the story. Because so few of the students who did not have an opportunity to discuss the story showed any change, they are not included in the table or discussed here. The table shows that three categories (1, 2, and 3) account for most of the types of change among students who were coded as showing evidence of change in facts and interpretations of change. Most students either came to understand basic facts about the story, changed from having no interpretation of the story to having an interpretation, or gained an improved understanding of characters' motives or emotions. These three categories account for 85% of the cases with evidence of change in facts and interpretations in Story 1 and 95% in Story 2.

Effects of Group Discussion on Understanding: Self-Reports

This variable was coded only for the discussion condition because this form of the test included a question following the group discussion that asked about any effects that the group discussion may have had on the student's understanding. About half of the students reading Story 1 (62 out of 127) and slightly less than half of the students reading Story 2 (51 out of 123) reported that the group discussion had affected their understanding of the story.

Some students simply stated that the group discussion had caused them to change their ideas about the story. For example, from Story 1:

> Some people in the group had different views of the story. Some liked it, some didn't. My ideas about the story changed. One group member explained the story, and now I understand it more. It's the kind of story that you'd have to read over a few times.

And from Story 2:

> We had a great group discussion! We were very deep and looked for the "hidden meanings" in the story. We each spoke - a lot and it was useful to get everyone's input. It also received some tension from taking a test. As a group, we had the same ideas, just from a slightly different point of view.

Table 8

Categories of Change in Students' Understanding from Part 1 to Part 2[a] of the Test (Forms With Discussion Only)

Type of change	Number of students	
	Story 1: "Story of an Hour" (n = 127)	Story 2: "An Ordinary Woman" (n = 123)
Students showing one kind of change		
1 Comes to understand basic facts	12	17
2 Changes a prior misunderstanding	1	0
3 Gives no interpretation in Part 1, gives an interpretation in Part 2	6	4
4 Greater understanding of motives or feelings	11	25
5 From a greater misunderstanding to a lesser misunderstanding	1	0
6 Adds a misunderstanding	1	0
7 Sharpens a vague interpretation	2	1
Students showing two types of change		
1 and 3 (Facts and interpretations)	8	1
1 and 4 (Facts and motives)	5	4
1 and 5 (Facts and lesser misunderstandings)	1	0
2 and 4 (Prior misunderstanding and motives)	2	0
3 and 4 (Interpretations and motives)	5	1
4 and 7 (Motives and sharper interpretations)	0	2
Total number of students showing change	55	55

[a] Part 1 = first two items on test; Part 2 = remaining items on test.

Many other students not only reported a change, but gave some insight into the change that had occurred.

> It made me think about the story in a whole new way. I now think that Louise wasn't going to die, but she now realized that she was free from her husband. That feeling that she had wasn't necessarily death, but was the feeling out self-independence. She really didn't love her husband, and now she was free. At the end of the story she died because the fact that he was supposably dead, and now he's alive killed her.

From Story 2:

> Before discussion I saw Caren as an unfeeling person, stone cold in her solitary lifestyle. However now I was reminded on the mug Mrs. Brooks was given. At one time it seemed Caren was a loving daughter that wanted to give her Mom a chance at being part of her life.

Relationship Between Self-Reports and Actual Change

The final analysis examined the correspondence between students' self-reports of change and evidence of change in their understanding of facts and interpretations as coded from their responses to test questions. Among students who showed no evidence of change in facts and interpretations, there was a strong relationship between their actual performance on the test and their self reports of change. For both Story 1 and Story 2, 92% of the students who showed no evidence of change in facts and interpretations reported that discussion had no effect on their understanding. There was less agreement between actual performance and self reports among students who did show changes in facts and interpretations from the first part to the second part of the test. For Story 1, of the 55 students whose answers revealed evidence of change in facts and interpretations, only 73% of them agreed that discussion influenced their understanding of the story. A substantial proportion (27%) claimed that group discussion did not have an effect on their understanding. The results were almost identical for Story 2. Of the 64 students who showed evidence of change, only 72% agreed that discussion influenced their understanding of the story; 28% did not.

This discrepancy between self-reports and evidence of change in facts and interpretations is found even among students who seem to have benefited the most: those who learned basic facts about the story, or who were able to make an

interpretation after discussion even though they had no interpretation of the story prior to discussion. For example, one student who read Story 1 phrased all of her answers as questions before the discussion.

> The event I picked was (1) Why was she saying 'Free free'? I think this is important because this was mostly my question and it got confusing everytime it would come up. Did she know that she was going to die or was she happy that her husband died (When he really wasn't? I want to know was she unhappy in her marriage?

After discussion, she wrote in response to the question about the effect of the group discussion:

> It didn't effect my ideas at all about the story. We all had the same ideas Realy.

But her answers to other questions after discussion reflect new understanding. For example:

> Well, I think she was happy that he died, probably because he abuse her . . . or maybe even he didn't let her do anything.

Another student initially thought that Brently was truly dead, and believed that Louise had died because she had been so distraught over Brently's death:

> Sure, some wives might be disstraught but, none of them are going to die because their husbands died in a freak accident.

After discussion, this student stated clearly that the group discussion had no effect:

> The group discussion affected my ideas about the story in no way whatsoever.

But the student's answers following the group discussion show substantial improvement in his understanding of the story. For example:

> I can't think of anything where the husband dies, the wife mourns but is then happy about his death because she didn't love him that much were she dies because she thinks he's dead but he is really alive.

Discussion of Results

In this study, a 10-minute discussion of a story during a 90-minute language arts test had a significant impact on students' understanding. The discussion

helped students understand basic facts of the story, helped them understand the characters' feelings and motives, and helped them understand the meaning of the story. Across individual students and between the two stories, though, there was substantial variation in the effects of group discussion. This section discusses how and why some students benefited from discussion while others did not, first for Story 1, then for Story 2.

Effects of Discussion in Story 1

For Story 1, the most important reason for the variation in the effect of the discussion is that students came into the group discussion with different levels of understanding of the story. For some students, the story was initially very confusing, and the group discussion was responsible for any understanding of the story that they were able to achieve. At the other extreme, some students had no difficulty understanding the story and found that the group discussion had little effect on their understanding.

Students who understood the story. For students who understood the story well prior to discussion, the discussion tended to be most useful in helping them form their interpretation of the story, understand characters' motives, and broaden their understanding of the meaning of the story. For example, some students who understood the facts of the story misinterpreted the feelings of the characters. One student wrote before discussion, "She must've really loved her husband to feel that much emotion." This misunderstanding was corrected after discussion, "At the beginning, she was still under her husband's beliefs and authority. But at the end, she totally realized that she could be her own person and she was now in complete control of her decisions, not her husband. She could now run her own life, it belonged to her."

Other changes among those who understood the story were more subtle, in details or inferences about the meaning of the story. For example, in several cases students adopted another group member's idea that the story reflected common marital practices at the time the story was written. In a few other groups the discussion brought forward issues of women's status in society. Others developed clearer ideas about the emotions or motives of the characters, or changes in the characters over the course of the story.

For those who understood the story prior to discussion and did not benefit from discussion, the most common reason seems to have been their confidence in

their understanding of the story. There is evidence in some cases that the group members learned early on in their discussion that all members of the group understood the story similarly, and thus believed that they had little left to discuss. For example, "The group discussion didn't affect my ideas of the story at all. Most of the people in my discussion group had all of the same ideas about the story." The fact that the test itself did not state a clear purpose for the discussion may have contributed to some groups' tendency to limit their discussion to basic facts about the story.

Other groups who found themselves fundamentally in agreement pursued interesting discussions, but did not ultimately find that these discussions had much impact on their understanding of the story, as the following two examples show:

> My partners had an idea more along the lines that she had to cope with death but at the end the roles were reversed and the husband had to cope with death. I don't believe in this idea as much as my own. The discussion wasn't very helpful.

> It's nice to hear other ideas and views on the story to get a broader idea of what it's actually saying. Many had common ideas so discussion went well. However, my own view remained the same and were not really affected by my groups discussion.

Others who understood the story found that other group members did not have as complete an understanding of the story as themselves, and used the discussion period to assist others, usually without clear changes in their own understanding. For example, one student wrote, "My ideas about the story didn't change. It made me realize I understand the story more then others."

Students with a partial understanding of the story. The majority of students seemed to come to the group discussion with a partial understanding of the story. They often understood one or two of the essential facts of the story, but not enough to grasp the full meaning of the story. The following student, for example, understood most of the story, but did not understand the death at the end that provided the ironic twist. Before discussion this student wrote, "How does that fit the story - Who died of heart disease? Louise? Richards wife? This story left me wondering about what happened." After discussion, the student understood the facts of the story and wrote, "But in the end of the story her husband comes home and she probaly died because she couldn't believe that her husband wasn't really dead, and she was probaly so happy then it had some effect on her."

In a number of cases, students' partial understanding of the story led them to misinterpret the story, and the group discussion often helped them to correct these misunderstandings. For example, a number of students who initially believed that Louise died because she was so happy to see her husband alive were able to grasp Louise's feelings much more clearly after discussion. For many students, the group discussion seems to have helped them fill in gaps in their understanding of the story.

Some students with a partial understanding of the story did not benefit from group discussion. First, some groups simply chose not to discuss the story. A few students remarked about it, for example, "We didn't talk about the story so it didn't help at all." Informal observations of the administration of the pilot test showed that there were substantial differences in students' engagement in the group discussion between groups, across classrooms, and across schools. Many of the groups we observed spent a majority of their discussion time on topics other than the story at hand. Without videotape or audiotape records, it is impossible to know what proportion of students spent what proportion of time discussing the story, but it seems evident that some students might have benefited more from the discussion if their groups had been more conscientious.

A second reason why some students who did not fully understand the story did not benefit from the group discussion may have been that the other group members had similar difficulties with the story, or made similar misinterpretations. In these cases, the group collaboration may have served to convince students that their misunderstanding was correct.

Students who did not understand the story. From a research perspective, those who did not understand the story initially are the best test of the usefulness of group discussion, because with them the presence or absence of change is usually very clear. In general, we have found that the less students understood before discussion, the stronger was the evidence that the discussion had an effect.

Those who did not understand the story independently had the most to gain from the group discussion. As might be expected, students who initially did not understand the story gained in understanding of the basics of the story, primarily facts and simple interpretations. While a handful of students changed from a complete or nearly complete misunderstanding of the story to a complete

understanding, most showed some incremental benefit, understanding somewhat more that they understood originally, but perhaps not as much as those who began with an accurate understanding of the story. For a student who was initially completely baffled by the story, a partial understanding was a very meaningful improvement. One student, for example, wrote before discussion, ". . . What was the point of the whole story was beyond me. To many words I didn't understand. The whole story was confusing." But after discussion, this student understood some facts of the story, and was able to put together an interpretation to fit the facts he understood, "I thing she did change because after her husband died all she wanted to do was die."

For those students who initially did not understand the story but did not benefit from the group discussion, one reason may have been that they were so discouraged by the difficulty of the story that they chose not to participate in the discussion. For some students, the story's archaic language, the plot twists, and the vague descriptions made the story so opaque that they simply gave up. It is also possible that some who did not understand the story had further difficulties understanding their peers' discussion of the story. For example, one student wrote about the group discussion, "It confused me more. They [ideas abut the story] didn't change at all cause I'm still confused but now Im confused more."

Effects of Discussion in Story 2

In contrast to Story 1, students reading Story 2 did not usually have much difficulty understanding the basic facts and plot of the story. Thus the discussions did not usually function to help students dispel misconceptions or learn new facts, and distinctions among students who initially understood more or less about the story are less useful. Instead, most students who benefited from discussion seem to have made an incremental improvement in their understanding, a new perspective, a new insight, or a new appreciation for the characters' feelings.

For Story 2, students' engagement with the story seems to have been a primary factor in the effectiveness of the group work. Unlike Story 1, Story 2 often struck an emotional chord among students, being described as "realistic" or "like my life," or similar to someone the student knows. Several students described in their answers incidents in their own lives similar to events in the story.

For students who were engaged by the story, the most common evidence of changes due to discussion were in students' understandings of the characters' motives and feelings, for example:

> Before discussion I saw Caren as an unfeeling person, stone cold in her solitary lifestyle. However now I was reminded on the mug Mrs. Brooks was given. At one time it seemed Caren was a loving daughter that wanted to give her Mom a chance at being part of her life.

The effects for other students were more subtle, in which they did not change their interpretation of the story, but gained a new perspective or a new insight into the story:

> The group discussion showed me a few more incidents in the story that I had not noticed the significance of. I think our group worked effectively because we all got a chance to share our personal ideas about the story. Because of the discussion, I think the theme intended for this short story was largely luck and the relevance of luck.

> My ideas didn't really change. One of the others brought up that the daughter felt resentment towards her Mom because Caren was competing with her Mom's high school image. That may have brought about the drug use. Besides that my thoughts and theirs were basically the same.

Other students seemed to find the discussion useful even though it did not lead them to substantial changes in their thinking about the story:

> We had a great group discussion! We were very deep and looked for the "hidden meanings" in the story. We each spoke - a lot and it was useful to get everyone's input. It also relieved some tension from taking a test. As a group, we had the same ideas, just from a slightly different point of view.

As with Story 1, some students did not find Story 2 to be particularly engaging or relevant. In these groups, discussion appears to have been much less helpful:

> Our group discussion wasn't very involved. We didn't discuss the story very intensely. both my partners said they didn't like the story and weren't interested by it. I would have been willing to discuss more, but I don't think they really wanted to. My ideas didn't change about the story.

While students benefited from discussion for both Story 1 and Story 2, the nature of the stories led to different kinds of learning from each story. Story 1's plot twists and sometimes archaic language led some students to misunderstand the story, and some to understand only parts of the story. For these students the group discussion provided an opportunity to correct misconceptions and fill in gaps in their understanding. Students who read Story 2 were faced with much less difficulty in understanding the events of the story. Their challenge was to understand the motives and feelings of the characters. As a result, the discussion did not serve as a way to correct misconceptions, but as a way to refine and compare understandings and interpretations, and share insights.

Relationship Between Actual and Self-Reported Changes in Understanding

As reported above, a substantial minority of students (more than one-fourth), whose answers demonstrated an improvement in understanding after discussion, claimed that the group discussion did not affect their understanding of the story. Without individual interviews or videotaped records, it is impossible to know with certainty why these students responded in this way. Nevertheless, students' answers give some clues about possible causes for the discrepancy between self-reports and evidence of change in their test responses. First, the fact that some students denied any effect of group discussion and then immediately went on to describe changes in their understanding that resulted from discussion implies that some students may have interpreted the question, "How did the group discussion affect your ideas of the story?," to refer to global or drastic changes, more than the understanding of one new fact, for example. Second, changes that result from group discussion may occur gradually, becoming incorporated slowly into a student's interpretation of the story. Students may not be aware of such gradual changes, and may believe that their understanding has not really changed much at all. Whatever the cause, this discrepancy between self-reported change and evidence of change is an interesting avenue for future research.

Conclusions and Implications

The results of this study show that even a small amount of collaboration on an assessment can have significant influences on students' understanding of the material and their performance on the test. Despite the fact that the group

discussion was relatively brief, group members were often unfamiliar with one another, and the purpose of the group discussion was ambiguous, a substantial portion of the student responses presented clear evidence of improvement after discussion. These findings have several implications for test design, administration, and use of test scores.

First, scores from tests with even a small amount of collaboration should not be interpreted as measuring unassisted student competence. In this study, many students clearly learned from the collaboration and performed better than they would have in the absence of collaboration. That these scores do not measure unassisted student competence does not necessarily make them invalid indicators of achievement, however. Because assessments with collaboration may closely reflect classroom practice, students' scores may be valid and representative indicators of their usual classroom performance. For students who work in groups on a regular basis, and who value collaboration in interpreting stories or writing, it may be most appropriate to test them in a collaborative context.

Second, if a small amount of relatively unstructured collaboration such as that used in the present study can have a measurable impact on student performance, then it is possible that greater amounts of collaboration can have an even larger impact on performance. This raises questions about the duration of group collaboration, and the timing and nature of the collaboration that should be used in assessments. Some possible next steps in research are to identify the nature, timing, and duration of group collaboration that tend to occur during instruction; to develop assessments that mirror naturally-occurring collaboration in the classroom; to investigate the effects of collaboration of this type on student performance; and to compare the effects of different amounts and kinds of collaboration on student performance.

Third, the differences in the effects of group collaboration across the two stories highlight the issue of the function of group collaboration in assessment. Different stories require different interpretive skills, and afford different opportunities to benefit from collaboration. Stories that emphasize characters' perceptions and feelings allow for students to share personal reactions and feelings, leading students to gain a greater understanding of the diversity of perspectives and interpretations of the story. Stories that require skills in decoding actions and statements afford opportunities for discussions about the

facts of a story, leading students toward correcting misconceptions and filling in gaps in understanding.

Fourth, issues arise about the fairness of different group compositions. Small groups will differ on the mix of student characteristics (ability, demographic characteristics, personality, motivation to do the task, experience collaborating with others, relative academic or peer-group status) and on the processes that emerge during collaborative group work. The composition of the group and the group processes that emerge during group work will have effects on how much students learn and on how they perform (for reviews of research, see Webb, 1995; Webb & Palincsar, 1996). In the context of the present study, for example, it would be to a student's advantage to work in a group with students who have good reading comprehension skills (who understand the story), who have good communication skills (who can explain and interpret the story in clear ways), and who are motivated to work with and help others (who are willing to discuss the story and explain their interpretations). It would be to a student's disadvantage to work in a group in which other students have poor reading comprehension skills, cannot communicate well with others, dominate the group or suppress students' participation, lack the motivation to work on the task, or prefer to engage in off-task talk or activities or disrupt group work in other ways.

When testing is done in the classroom and the teacher has control over group formation, the teacher can assign students to groups to try to equalize the mix of ability and other student characteristics across groups, or to form groups that are likely to work productively. When tests are not administered in intact classrooms or the test administrator is not familiar with the students, forming groups with comparable mixes of student backgrounds and abilities or forming groups that are as productive as possible would be very difficult, if not impossible.

One way to lessen the variation in group processes that may arise as a result of different group compositions is to prepare students and teachers for collaborative group work. Students can practice working on tasks collaboratively; they can receive training in effective communications skills; they can be encouraged to actively participate in group work and to encourage the contributions of others; and they can be taught how to help others, seek help, and engage in high-level discussion of ideas, all of which have been shown in previous research in classroom instructional settings to promote group processes that are beneficial for learning (see Webb, 1995). Our informal observations of some small

groups in the present study confirm the importance of previous experience working in collaborative groups. Students who had more previous experience working in groups were more at ease in the testing situation, were able to begin the discussion more quickly, and spent less time negotiating about the purpose of the discussion or which issues to discuss than students who had less previous experience with collaborative group work. Giving students and teachers instruction and training in how to work in groups productively may help all groups function in the most effective ways possible. Although training in productive group work would not eliminate the inequities caused by some groups having particularly skilled members, it may help reduce inequities caused by some groups functioning more effectively than others.

Within the context of the test administration itself, the test administrator can foster beneficial group processes. Our informal observations of small-group work in this study suggested that students were more engaged and group discussion was more fruitful when teachers were actively circulating among groups and encouraging students to share their ideas than when they simply arranged students into groups and told them to follow the instructions in the test booklet. Future research should systematically investigate how different ways of preparing students and teachers for collaborative assessments influence student performance on such assessments.

Finally, important challenges for future research are to document the processes that occur when students collaborate on assessment tasks, to investigate the impact of group processes on test performance, and to investigate the variables that might predict whether beneficial or detrimental group processes will occur during group assessment. The present study did not systematically observe group processes nor did it collect information about variables that might predict group processes, such as the composition of the small groups with respect to ability and previous experience with small-group collaboration. So it is impossible in the present study to discern which group processes or group compositions produced the largest changes in student performance before and after group discussion.

Documenting group processes may also help assuage equity concerns, especially when scores are reported at the individual student level but also when scores are reported at aggregate levels such as school or district. Performance scores could be interpreted in light of information about the group processes. For

example, students may not be judged so harshly for low performance scores when they are members of poorly functioning groups.

The results of this study clearly indicate that collaboration does have an effect on students' performance on assessments. An assessment that includes group discussion cannot be understood in the same way as a traditional assessment, as a measure of how students can perform without assistance from others. At the same time, an assessment that includes collaboration can answer questions about students' social interactions and classroom functioning that cannot be answered by traditional assessments. The challenge to test developers, administrators, and users is to articulate clear objectives that will lay a foundation for the form and purpose of group work in assessments.

References

Almasi, J. F. (1994). The nature of fourth graders' sociocognitive conflicts in peer-led and teacher-led discussions of literature. *Reading Research Quarterly, 29,* 304-306.

Awbrey, M. (1992, September). *History-social science group assessment in California (high school level).* Paper presented at the National Center for Research on Evaluation, Standards, and Student Testing's Conference on "What Works in Performance Assessment," UCLA, Los Angeles, CA.

Baron, J. B. (1994, April). *Using multi-dimensionality to capture versimilitude: Criterion-references performance-based assessments and the ooze factor.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Bartlett, L. D. (1992). Students successfully grapple with lessons of history in innovative group performance tasks. *Social Education, 56,* 101-102.

Bossert, S. T. (1988). Cooperative activities in the classroom. *Review of Research in Education, 15,* 225-252.

Brennan, R. L. (1992). *Elements of generalizability theory* (rev. ed.). Iowa City: American College Testing.

Brown, A. L., & Palincsar, A. S. (1989). Guided, cooperative learning and individual knowledge acquisition. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 393-451). Hillsdale, NJ: Erlbaum.

Connecticut State Board of Education. (1987). *Common core of learning.* Hartford, CT: Connecticut State Board of Education.

Eeds, M., & Wells, D. (1989). Grand conversations: An exploration of meaning construction in literature study groups. *Research in the Teaching of English, 23,* 4-29.

Kansas State Board of Education. (1993). *Kansas curricular standards for science.* Topeka, KS: Kansas State Board of Education.

Leal, D. J. (1993). The power of literary peer-group discussions: How children collaboratively negotiate meaning. *The Reading Teacher, 47,* 114-120.

Leal, D. J. (1992). The nature of talk about three types of text during peer group discussions. *Journal of Reading Behavior, 24,* 313-338.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15,* 1-16.

Lomask, M., Baron, J., Greigh, J., & Harrison, C. (1992, March). *ConnMap: Connecticut's use of concept mapping to assess the structure of students' knowledge of science.* A symposium presented at the annual meeting of the National Association of Research in Science Teaching, Cambridge, MA.

Mathematical Sciences Education Board, National Research Council. (1993). *Measuring up: Prototypes for mathematics assessment.* Washington, DC: National Academy Press.

McMahon, S. (1992). Book club: A case study of a group of fifth graders as they participate in a literature-based reading program. *Reading Research Quarterly, 27*(4), 292-294.

Neuberger, W. (1993, September). *Making group assessments fair measures of students' abilities*. Paper presented at the National Center for Research on Evaluation, Standards, and Student Testing's Conference on "Assessment Questions: Equity Answers," UCLA, Los Angeles, CA.

Noll, E. (1994). Social issues and literature circles with adolescents. *Journal of Reading, 38,* 88-93.

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1,* 117-175.

Palincsar, A. S. (1986). The role of dialogue in providing scaffolded instruction. *Educational Psychologist, 21,* 73-98.

Palincsar, A. S., Brown, A. L., & Martin, S. M. (1987). Peer interaction in reading comprehension instruction. *Educational Psychologist, 22,* 231-253.

Pandey, T. (1991). *A sampler of mathematics assessment.* Sacramento, CA: California Department of Education.

Raphael, T. E., McMahon, S. I., Goatley, V., Bentley, J., Boyd, F. B., Pardo, L. S., & Woodman, D. A. (1992). Literature and discussion in the reading program. *Language Arts, 69,* 54-61.

Reid, L., Cintorino, M. A., Crews, W. M., & Sullivan, A. M. (1994). Making small groups work. *English Journal, 83*(3), 59-63.

Samway, K. D., Whang, G., Cade, C., Gamil, M., Lubandina, M. A., & Phommachanh, K. (1991). Reading the skeleton, the heart, and the brain of a book: Students' perspectives on study circles. *The Reading Teacher, 45,* 196-205.

Saner, H., McCaffrey, D., Stecher, B., Klein, S., & Bell, R. (1994). *The effects of working in pairs in science performance assessments*. Santa Monica, CA: The Rand Corporation. Manuscript submitted for publication.

Shavelson, R. J., & Baxter, G. P. (1992). What we've learned about assessing hands-on science. *Educational Leadership, 49,* 20-25.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage Publications.

Slavin, R. E. (1990). *Cooperative learning: Theory, research, and practice.* Englewood Cliffs, NJ: Prentice-Hall.

Webb, N. M., & Palincsar A. S. (1996). Group processes in the classroom. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology*. New York: Macmillan.

Webb, N. M. (1995). Group collaboration in assessment: Multiple objectives, processes, and outcomes. *Educational Evaluation and Policy Analysis, 17,* 239-261.

Webb, N. M. (in press). Assessing students in small collaborative groups. *Theory into Practice.*

Webb, N. M., Nemer, K., Chizhik, A., & Sugrue, B. (1996, April). *Equity issues in collaborative group assessment: Group composition and performance.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Wise, N., & Behuniak, P. (1993, April). *Collaboration in student assessment.* Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.