Equity Issues in
Collaborative Group Assessment:
Group Composition and Performance

CSE Technical Report 457

Noreen Webb, Kariane Nemer, Alex Chizhik
University of California, Los Angeles

Brenda Sugrue
University of Iowa

June 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 90095-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Center Program, PR/Award Number R305B60002, as administered by the Office of Educational Research Improvement, U. S. Department of Education.
The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U. S. Department of Education.

EQUITY ISSUES IN COLLABORATIVE GROUP ASSESSMENT:

GROUP COMPOSITION AND PERFORMANCE [1,2]

Noreen Webb, Kariane Nemer, Alex Chizhik
University of California, Los Angeles

Brenda Sugrue
University of Iowa

Abstract

Large-scale assessment programs increasingly use group assessment tasks, in which small groups of students collaborate to solve problems or complete projects, often in combination with tasks that students perform individually. Whether some group compositions may be more advantageous than others is an important equity issue. The present study investigated the effects of group ability composition on group processes and outcomes in science performance assessments in which students worked on a series of assessments first individually, then in groups, and finally individually again. Group composition had a major impact on the quality of group discussion in group work and on students' achievement test scores, both during group work and on the subsequent individual test. Groups with above-average students produced more correct answers and generated a greater number of high-quality explanations of how to solve the test problems than did groups without above-average students. The higher level of discussion translated into an advantage on the achievement tests for below-average students working in groups with above-average students compared to below-average students working in groups without above-average students: the former students performed better on the achievement test completed during group work *and* on the subsequent achievement test completed individually than did the latter students. High-ability students performed equally well in heterogeneous groups, homogeneous groups, and when they worked alone.

---

[1] We would like to thank the following teachers for participating in this study: Robert Aochi, Carol Binggeli, Marilyn Mounce, Esther Oey, Lynn Scheffield, and Kathryn Stevens. We also thank Rosa Valdes, Jonathan Troper, and Jonah Schlackman for their help in data collection and coding.
[2] A version of this paper was presented at the annual meeting of the American Educational Research Association in New York in April 1996.

Introduction

Large-scale assessment programs and small-scale in-class assessments increasingly use collaborative small-group work in which students work together to solve problems or complete projects. In most cases, students complete one part of the assessment in a collaborative group and other parts of the assessment individually, before and/or after groupwork (see Connecticut's Common Core of Learning Assessment: Baron, 1994; Connecticut State Board of Education, 1987; Lomask, Baron, Greigh, & Harrison, 1992; California Assessment Program: Awbrey, 1992; Bartlett, 1992; Pandey, 1991; California Leaning Assessment System: Saner, McCaffrey, Stecher, Klein, & Bell, 1994; Oregon State Department of Education: Neuberger, 1993; Shavelson & Baxter, 1992). Emerging state and national standards for assessment also include recommendations for the incorporation of small group work (e.g., Kansas State Board of Education, 1993; Mathematical Sciences Education Board, National Research Council, 1993).

An often-cited reason for using group work in assessment is to link assessment more closely to the widespread use of small-group collaboration in classroom instruction (Linn, 1993; Wise & Behuniak, 1993). Based on two decades of research showing that collaborative learning increases student learning and social-emotional outcomes such as social skills, self-esteem, and attitudes towards others (Bossert, 1988-1989; Slavin, 1990), many school districts and state departments of education mandate the use of collaborative work in the classroom (e.g., California State Dept. of Education, 1985, 1992).

Just as students can learn from each other during small group instructional activities, they can learn from each other during collaborative work on assessments. Students can acquire new skills, ideas, and knowledge by working together to build solutions to problems, resolving disagreements and conflicting perspectives, and giving and receiving help (Webb, 1995; Webb & Farivar, in press; Webb & Palincsar, 1996).

The opportunity to learn from others is an important equity issue, especially when some students have differential access to resources. Neuberger (1993), for example, argues that collaborative work on assessments may help equalize resources among students with different educational backgrounds and make achievement testing more fair.

Among assessment programs that build in opportunities for learning from collaborative groups is the Connecticut Common Core of Learning Alternative Assessment Program (Baron, 1994). Baron (1994) describes an example science assessment from that program. In the first part of the assessment, students provide information about their science knowledge by individually listing the factors that influence yeast's activity in food. In the second part, students work in three-person groups to design, carry out, interpret, and summarize (in writing and orally to the class) an experiment investigating the activity of yeast in food. The third part requires students to independently analyze and critique another group's report. The assessment explicitly recognized the possibility students would learn from each other: the assessment provided students with an opportunity to "deepen their understanding of the concepts and skills being assessed" (Baron, 1994, p. 3).

A few studies have produced evidence of students learning from collaborative work during assessments. In a study of a science assessment with a similar individual–group–individual work design, Saner et al., (1994) found that working in pairs to carry out a scientific investigation influenced students' scores on the subsequent individual part of the assessment, which involved interpreting the results of the pair's investigation. Carry-over effects were strongest for students who scored low on the first individual part of the assessment (short-answer questions about content knowledge).

Learning can occur even without extensive opportunities for collaboration. A ten-minute discussion in the middle of a 90-minute test improved tenth-grade students' language arts performance on the Fall 1992 pilot of the Connecticut Academic Performance Test developed by the Connecticut State Department of Education. Some test forms required students to engage in a 10-minute small group discussion after reading the story and individually answering two questions, but prior to completing the remaining questions. Compared to students without an opportunity to discuss the story, the post-discussion responses of students who engaged in group discussion demonstrated improved understanding of the story (more correct interpretations of plots and characters; Wise & Behuniak, 1993; Fall, Webb & Wise, 1995).

In addition to providing learning opportunities, collaborative assessments can be used to measure group productivity. Due to the importance of teamwork in the workplace, potential employers want to know what students can accomplish in teams (Hackman, 1990). Including collaboration in assessments allows for the gathering of information about group productivity and group effectiveness.

In addition to measuring group productivity, including collaboration on assessments allows for the measurement of a student's ability to work with others. For workforce readiness, in addition to academic skills, students need interpersonal and teamwork skills such as coordination, communication, conflict resolution, decision making, problem solving, and negotiation (O'Neil, Allred, & Baker, 1992; Salas, Dickinson, Converse, & Tannenbaum, 1992). The Secretary's Commission on Achieving Necessary Skills (SCANS) listed the following interpersonal abilities as workforce readiness skills: participating as a member of a team (working cooperatively and productively with others), teaching others new skills, serving clients and customers, exercising leadership, negotiating, and working with diverse groups (O'Neil, Allred, & Baker, 1992; SCANS, 1991). Using self-evaluations or observations, some achievement assessment programs have measured collaborative behavior such as participating in a group, offering useful ideas, involving others, and communicating clearly (Awbrey, 1992; The Connecticut Common Core of Learning Alternative Assessment Program, described in Baron, 1994).

Incorporating collaborative group work into an assessment introduces a complexity not relevant to individual assessments: group composition. Little is known about how group composition influences the performance and processes of collaborative groups working on assessments and the learning of individuals within those groups, and whether some group compositions may be more "fair" than others. Some combinations of students may confer an advantage over others, in terms of group productivity (group performance) or in relation to student learning (individual performance following collaboration).

Previous research in social psychological and organizational settings demonstrates that group productivity (group performance) on intellectual problem-solving tasks often correlates with either the average ability or the highest ability level in the group (e.g., Hastie, 1986; Laughlin & Branch, 1972; Moreland & Levine, 1992). For example, Laughlin and Branch (1972) systematically compared group and individual performance among 1008 college students taking the Terman Concept Mastery Test. Based on their individual scores, students were categorized as high (H), medium (M), or low (L), and were then assigned to one of the following four-person group compositions to take the test again: HHHH, HHHM, HHHL, HHMM, HHML, HHLL, HMMM, HMML, HMLL, HLLL, MMMM, MMML, MMLL, MLLL, LLLL. Laughlin and Branch found that group performance corresponded to both the level of the highest group member and the number of individuals at that ability level. Furthermore, working in a group with persons of higher-ability conferred an advantage over working alone. However, working in a group with persons of lower ability did not.

Most studies of group productivity involving different group compositions use problems that have obvious answers that can easily be explained or demonstrated by a single competent group member (Hastie, 1986; Laughlin & Ellis, 1986). Having a competent person in the group increases the group's chances of successfully solving the problem. Cohen (1994; Cohen & Cohen, 1991) argues, in contrast, that if groups work on a complex task which cannot be completed by an individual or if it has an ill-structured solution instead of a single correct answer, then having a competent person in the group is not sufficient. Because no individual is likely to have all of the necessary expertise to solve a complex problem, groups in which members exhibit a variety of perspectives and areas of competence may perform better than groups with an "expert" (Cohen, 1994; Cohen & Cohen, 1991). Although not all of the tasks in the present study have a single correct answer, the tasks have well-structured solutions and can theoretically be solved by individuals who are knowledgeable about concepts in the domain. Consequently, having an "expert" in the group may be better for group performance than having a range of competencies.

Studies of group composition and learning generally show that when students actively participate in group collaboration, low-ability students learn best in groups with high-ability students, high-ability students perform well in any group composition, and medium-ability students learn most in relatively homogeneous groups (Lou et al., 1996). The advantages of particular group compositions correspond to the nature of students' participation in group collaboration. Low-ability students in heterogeneous groups have an opportunity to receive help from higher-ability students (e.g., Azmitia, 1988; Mugny & Doise, 1978; Webb, 1980). High-ability students typically participate actively and perform well whether they work with other high-ability students or with lower-ability students (e.g., Azmitia, 1988; Hooper & Hannafin, 1988; Hooper, Ward, Hannafin, & Clark, 1989; Skon, Johnson, & Johnson, 1981), although one study found that high-ability students learned more in heterogeneous groups than in homogeneous groups because they assumed the role of teacher and gave more explanations (Webb, 1980). For medium-ability students, some studies have shown that they learn less in heterogeneous groups than in homogeneous groups. Possible causes for this result are that they may be excluded from the teacher-learner relationships that develop between highs and lows, they may not receive the help they need, or they may not be allowed to actively participate (Webb, 1991). Whether the effects of group composition on learning depend on the nature of the task or problem is unknown.

Previous research on the effects of group composition on group productivity and individual learning yields complex results. They do not allow for simple predictions about the role of group composition in collaborative assessments. The present study, in the context of science performance assessment, investigated the impact of group composition on the performance of groups, the quality of group discussion, and the achievement of individual students. This study examined the impact of group composition both on the performance of individual students while working in collaborative groups (related to group productivity) and on the performance of students after they worked in groups (related to learning from group collaboration).

Method

Sample

The study consisted of 662 seventh-grade and eighth-grade students (21 classes) from five schools in Los Angeles County (n = 501 students had complete data on the variables used here). Six teachers taught the 21 classes: two teachers came from one school and the other teachers came from four schools. The selected schools represent a wide range of demographic characteristics. Table 1 gives a summary of the distribution of student characteristics at each school. The schools had very different ethnic compositions, ranging from predominately minority (Black, Hispanic, or both) to predominately white. Schools also differed in distribution of ability (mixture of verbal and nonverbal reasoning, vocabulary, and science knowledge; described in a later section). The predominantly white schools had higher proportions of high-ability students than the other schools.

Table 1

Percent of Students in Each Ethnic and Ability Category by School

School	Ethnic Background				Ability Level			
	White	Black	Hispanic	Asian	Low	Low-Medium	Medium-High	High
1	2	0	96	2	38	28	25	9
2	88	4	3	6	9	9	26	56
3	67	18	9	4	10	23	31	36
4	0	33	67	0	44	35	18	3
5	54	3	31	12	29	30	23	18

Procedures

The assessment was designed to measure students' understanding of voltage, resistance, and current, and the relationships among them in the context of simple electric circuits. At the beginning of the study, prior to instruction in electricity and electric circuits, students were administered three pretests: vocabulary (Vocabulary Test V-1, Buros, 1974), verbal reasoning (the New Jersey Test of Verbal Reasoning Form B, Shipman, 1985), and nonverbal reasoning (Raven's Progressive Matrices). Following administration of these tests, all teachers conducted a three-week unit on electricity and electric circuits in their classrooms. The teachers were given the freedom to teach the topic in their own way, using their own instructional materials. Teachers were told that the assessments to be administered at the end of instruction would target students' knowledge of the concepts of voltage, resistance, and current, as well as the relationships among them, but teachers were not shown copies of the tests.

At the end of the instructional unit, students were administered two performance tests measuring their understanding of voltage, resistance, and current, and the relationships among them. They completed both tests individually, without help from other students or from the teacher. The hands-on test required students to assemble batteries, bulbs, wires, and resistors into electric circuits, draw diagrams of their circuits, and then answer questions (multiple choice and written explanations) about them. All circuits that students assembled were videotaped.

The paper-and-pencil written test contained some items analogous to the hands-on test (see Appendix), but required students to draw the circuits instead of assemble them. The other test items did not have direct analogues on the hands-on test; these items asked students to identify circuits with the highest voltage, resistance, or current, or asked students to predict the effect on voltage, resistance, and current of, for example, adding a battery or bulb to an existing circuit.

One month later, with no intervening instruction or review, students were re-administered the same two science tests (hands-on and written). Students completed the hands-on test first and the written test the next day. For the hands-on test, approximately 80% of the students in each class worked on the test in collaborative three-person groups; as a control, the remaining 20% of the students worked individually at separate desks with no interaction with others. All students completed the written test individually.

In the group condition, in which students worked on the hands-on test in collaborative groups, each group was given one set of manipulatives (batteries, bulbs, wires, resistors) but every student had to fill out his or her own test paper. Students were instructed to work together to complete the tasks, to help each other understand how to solve the problems, to ask questions of each other, and to assist group members who seemed confused or appeared to need help. Five groups in each class were videotaped while they completed the test (70% of the groups in this study were videotaped).

Design

The testing design, then, had three phases. In Phase 1, students completed the verbal and nonverbal pretests and the hands-on and written science tests individually. In Phase 2, students completed the hands-on test again either in collaborative three-person groups or individually. In Phase 3, all students individually completed the written test. Table 2 summarizes the design of the study.

Scoring of Science Achievement Tests

Hands-on test. The hands-on test consisted of two tasks. For each task, students were given a bag of materials containing 9-volt and/or 1.5-volt batteries, wires, bulbs, and graphite resistors (Task 1: two 9-volt batteries and two 1.5-volt

Table 2
Design of the Study

Phase 1	Phase 2	Phase 3
INDIVIDUAL	GROUP	INDIVIDUAL
(100% of sample)	(81% of sample)	(100% of sample)
Pretests	Hands-on science test	Written science test
Vocabulary		Analogue to hands-on test items
Verbal reasoning	INDIVIDUAL	Non-analogue items
Nonverbal reasoning	(19% of sample)	
Hands-on science test	Hands-on science test	
Written science test		
Analogue to hands-on test items		
Non-analogue items		

batteries, three bulbs in bulb holders, and seven wires with metal clips on the ends; Task 2: two 9-volt batteries, two bulbs in bulb holders, three graphite resistors, and seven wires with metal clips on the ends). Students were asked to assemble pairs of circuits so that the bulb in one circuit was brighter (or dimmer) than the bulb in the other circuit. After circuit construction, students were asked to draw diagrams of their circuits, and answer three multiple-choice items about which of their two circuits had higher voltage, resistance, and current. Further, the test asked students to write an explanation to justify each multiple-choice answer. In addition, students were asked to explain why one circuit was brighter than the other (see Appendix).

Students assembled a variety of circuits. Because different pairs of circuits gave rise to different correct answers, the multiple choice items and the explanations were scored according to the circuits that students assembled (as shown on the videotapes of their circuits). For example, if a student assembled two circuits which each contained a nine-volt battery and a 1.5-volt battery, both circuits would have equal voltage. If a student assembled one circuit with two 9-volt batteries and the other with two 1.5-volt batteries, the first circuit would have a higher voltage than the second.

Each multiple choice item received a score of incorrect or correct (0,1) depending on the circuit students assembled. Explanations were scored on a 0 to 1 scale according to accuracy and completeness. For example, when asked "Why was voltage in Circuit A higher than in Circuit B?," the following scores were assigned: 1 if a student mentioned the relative number of batteries in the two circuits in addition to the relative power or voltage generated by the batteries, 0.67 if a student mentioned the relative number of batteries in the two circuits but not the strength or voltage of the batteries, 0.33 if a student mentioned batteries but not the relative number or relative strength of them, and 0 if the explanation was irrelevant or if it displayed confusion over cause and effect (for example, "the voltage is higher because it is brighter").

Two coders scored all explanations. The correlation between coders' ratings was very high for all explanation questions, exceeding 0.95.

Written test. The written test had two parts. One part consisted of items exactly analogous to the items on the hands-on test except that students were given pictures of the equipment (batteries, bulbs, resistors) instead of actual manipulatives (see Appendix). Otherwise, on this portion of the test, the same instructions were given. Students were asked to construct two circuits (draw diagrams using the items given in order to make the bulb in one circuit brighter than the bulb in the other) and answer the same multiple choice and explanation questions.

The other part of the test had items that were different from the hands-on analogue items but measured the same concepts. These items all concerned voltage, resistance, current, and Ohm's law, but asked questions in different ways. Some items required students to select which circuit, from a set of pictures, had the highest voltage, resistance, or current. Other items asked students to indicate whether the voltage, resistance, and current in a circuit would increase, decrease, or stay the same if different changes were made to the circuit (i.e., a bulb added or removed). Still other items asked students to write an explanation for why one pictured circuit was brighter than another.

Scores Used in the Analyses

Hands-on test score. The hands-on test score used for analysis was the mean of the three multiple choice items and four explanation items for the two tasks, for a total of 14 items (equally weighted). There was one hands-on score for Phase 1 (internal consistency alpha = .86) and one for Phase 2 (alpha = .83).

Written analogue score. As described above, a portion of the written test was analogous to the hands-on test. The only difference was that students were given pictures of the batteries, bulbs, wires, and resistors instead of the actual equipment. Scores on the analogue items were combined in the same way as on the hands-on test to form a composite score (14 items, equally weighted). There was one written analogue score for Phase 1 (alpha = .85) and one for Phase 3 (alpha= .85).

Written non-analogue score. As previously described, the written test had additional items measuring voltage, resistance, current, and Ohm's law that did not match those on the hands-on test. Twenty-seven of the items were multiple choice, and three required explanations. The written non-analogue score was the mean of the 30 items (equally weighted). There was one written non-analogue score for Phase 1 (alpha = .91) and one for Phase 3 (alpha = .90).

Phase 1 composite "ability" score. A composite score based on all of the tests administered during Phase 1 was computed to serve as a measure of students' competence prior to Phase 2. The composite ability measure was the weighted mean of the following Phase 1 scores: vocabulary, verbal reasoning, nonverbal reasoning, the hands-on test score, the written analogue score, and the written non-analogue score. In the composite, the science scores were weighted more heavily than the verbal and nonverbal scores because, among students who worked individually in all phases of the study, science scores in Phases 2 and 3 were more highly correlated with science scores in Phase 1 (*r*'s ranged from .68 to .69) than with the verbal and nonverbal scores (*r*'s ranged from .40 to .60). The composite measure is called "ability" to distinguish it from the achievement scores in Phase 2 and Phase 3. Although the term "ability" is used here, it should be noted that the composite is a combination of science achievement, verbal and nonverbal reasoning ability, and vocabulary, and so has a heavy achievement component.

Students were classified as low, low-medium, medium-high, or high in ability on the basis of the Phase 1 composite ability score. One quarter of the sample comprised each ability level.

Group compositions

To examine the effect of working with an able student on the performance of less-able students, six students in each class were assigned to two groups in Phase 2 so that one group had one high-scoring student and two low-scoring students and the other group had three low-scoring students. Due to the limited time available for scoring tests between Phases 1 and 2, these group assignments were made on the basis of students' Phase 1 hands-on test scores (multiple choice items only) and verbal and nonverbal pretest scores. The remaining groups were formed randomly. Each classroom had a variety of group ability compositions, including heterogeneous groups (students from different ability levels) and homogeneous groups (students from the same ability level). All groups were formed so that they would be heterogeneous on gender and ethnic background to reflect the class mix of student characteristics.

The 20% of students in each class who worked individually in Phase 2 were purposely selected to represent the range of ability in the class.

Coding of Videotapes

Group discussions. Groups were videotaped as they worked collaboratively on the hands-on test. Group discussions were coded using the same scoring scheme applied to the hands-on test: the accuracy of their multiple choice answers and the quality of their explanations. For each multiple choice item on the test, the group was given a 1 if anyone in the group gave the correct answer, and a 0 otherwise. For each item, the group was given a score for the highest level of explanation articulated based on the contributions of the three students in the group. For example, when looking at an explanation about why one circuit had higher current than the other, if one student explained that Circuit A had higher current because it had more voltage and another student explained that the current was also higher due to lower resistance, the two explanations (one about voltage and one about resistance) were combined to form a single explanation which was then scored. The explanation score, then, reflected the maximum level of discussion in the group, and could have reflected the contributions of one, two, or all three students.

To check the agreement between raters, two coders scored the group discussion variables for 42 groups (two per classroom). The correlation between raters' codes was 1.00 for accuracy of answers and 0.99 for quality of explanations.

Individual behavior of students in groups. Contributions of individual students to group discussion were coded according to the cognitive level of their participation. Adapting Webb, Troper, and Fall's (1995) continuum of activity from most constructive to least constructive, where high-level activity was most predictive of student achievement, we used three levels of participation. High-level participation included making or defending suggestions for how to answer a test item (or adding to or expanding on someone else's suggestion), asking questions about a suggestion, and paraphrasing a suggestion. Medium-level participation included copying someone else's suggestion, repeating it verbatim, or simply agreeing with what was said. Low-level participation included listening or watching without making any substantive contribution or inquiry. The final category was working alone, such as manipulating materials or writing answers without reference to other members of the group. The score given to each student for each level of participation was the number of items (out of eight items that were discussed in the group) that he or she engaged in at that level.

In contrast to the achievement and group discussion variables, which were used in correlational analyses or analyses comparing groups of students (correlations between group discussion and achievement, comparisons between students of different achievement levels—to be reported in the results section), interest in the individual behavior lay in students' absolute level of behavior, not the relative standing of students. Most importantly, interest lay in how much of individual student behavior could be classified as high-level participation, not only whether some students engaged in higher level participation than other students. Consequently, a stricter standard of rater agreement was used for individual behavior than the correlations between raters used for other variables. To check the agreement between raters when coding individual behavior, a generalizability analysis was conducted for each level of participation: two raters coded a random sample of 10 students and coded the number of items (out of eight) that each student engaged in at that level. Each analysis produced an estimated index of dependability, a reliability-like coefficient that showed the consistency of raters in coding students' absolute behavior, not only the relative standing of students (see Brennan, 1992;

Shavelson & Webb, 1991). Consistency among raters was high: the estimated indices of dependability ranged from .87 to .98 across the four levels of participation.

Results

Preliminary Analyses to Determine Group Composition Types

We examined the influence of the lowest and highest and mean ability of the group on the performance of students in the group. Regression analyses were performed using the three variables based on ability (highest, lowest, mean) as predictors of student achievement and group discussion scores. For students in the lowest three quarters, only the highest ability level in the group significantly predicted achievement outcomes. For students in the highest quarter, only the lowest ability level in the group predicted achievement outcomes. Therefore, these are the only group composition variables reported here.

Impact of Group Ability Composition on Science Achievement Test Scores

Tables 3 through 7 give the mean achievement test scores for students of each ability level according to the composition of their Phase 2 group or whether they worked individually in Phase 2. Most of the results reported combine all schools. The results of analyses conducted for each school separately were similar to those of the combined analyses but were not always statistically significant due to low power (small sample sizes in some analyses).

Low-ability students. Table 3 and Figures 1, 2, and 3 present the results for low-ability students (those in the bottom quarter on the composite ability variables) working in different group compositions or individually. The two rows of the table corresponding to the hands-on test give the mean scores for the hands-on test that low-ability students completed individually in Phase 1 before group work and the hands-on test that they completed while working in collaborative groups in Phase 2 (presented graphically in Figure 1). As can be seen in Table 3 and Figure 1, low-ability students who worked with above-average students performed better on the Phase 2 hands-on test than did low-ability students who worked with below-average students.

Table 3

Achievement Test Scores for Low-Ability Students Who Worked in Different Group Compositions During Phase 2

				Highest Ability Level in the Group											
	Worked Individually			Low			Low-Medium			Medium-High			High		
Achievement Test Scores[a]	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*
Phase 1 ability composite	.21	.05	22	.23	.04	21	.22	.08	27	.23	.05	28	.22	.05	15
Hands-on															
Phase 1 (individual)	.14	.08	21	.15	.09	17	.16	.11	26	.14	.11	25	.12	.10	15
Phase 2 (group)	.30	.19	21	.30	.15	17	.36	.17	26	.43	.21	25	.64	.17	15
Written analogue[b]															
Phase 1 (individual)	.17	.13	21	.17	.12	17	.15	.11	24	.15	.11	25	.15	.08	14
Phase 3 (individual)	.25	.18	21	.26	.14	17	.37	.21	24	.36	.17	25	.47	.21	14
Written non-analogue															
Phase 1 (individual)	.27	.17	21	.34	.11	17	.27	.15	24	.30	.14	25	.25	.18	14
Phase 3 (individual)	.37	.16	21	.36	.11	17	.38	.15	24	.38	.17	25	.48	.17	14

[a]Proportion Correct

[b]Analogue to hands-on test; see text

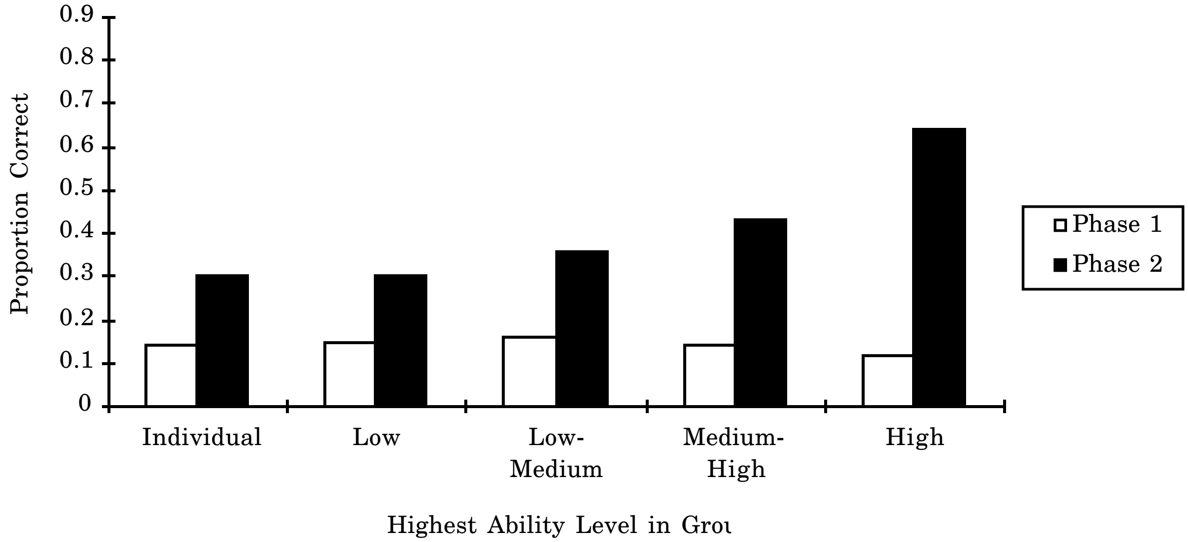


Figure 1. Mean hands-on scores in Phases 1 and 2: Low-ability students in different group compositions.

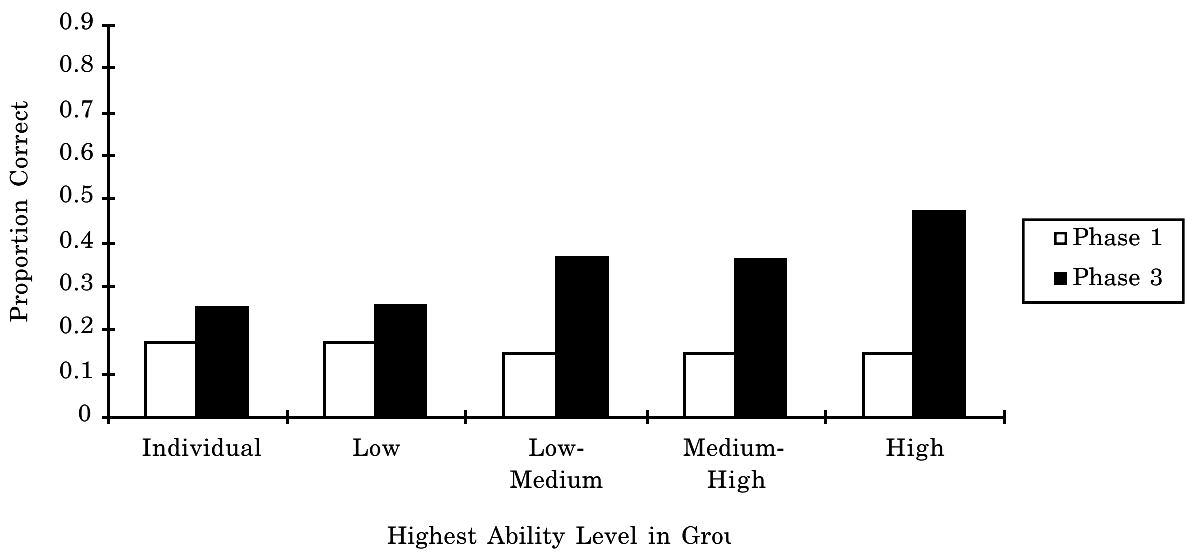


Figure 2: Mean written analogue scores in Phases 1 and 3: Low-ability students in different group compositions.

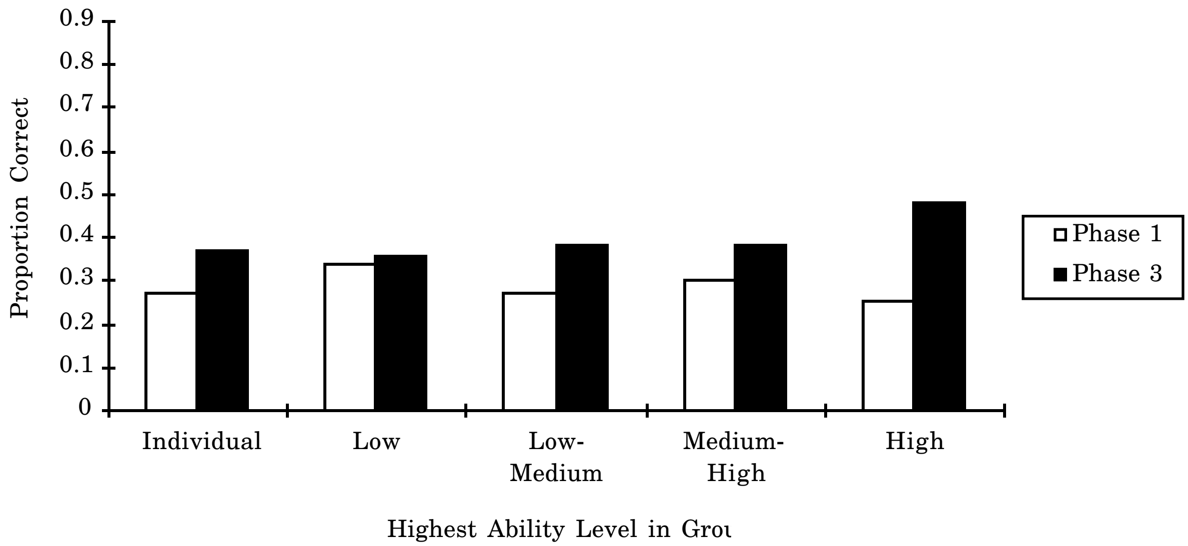


Figure 3: Mean written non-analogue scores in Phases 1 and 3: Low-ability students in different group compositions.

To test the difference among group compositions, an analysis of covariance was performed with the ability composite and the corresponding Phase 1 hands-on test score as covariates. The difference between means was statistically significant ($F(3,77) = 12.14$, $p < .0001$). Moreover, Table 3 and Figure 1 show that low-ability students who worked with below-average students did no better than students who worked individually.

The experience of working with above-average students carried over to the written test that was completed individually the next day. The middle two rows of Table 3 (see also Figure 2) show the improvement in the written analogue scores of low-ability students who worked with high-ability students ($F(3,74) = 4.09$, $p = .01$). The bottom two rows of Table 3 (see also Figure 3) show the improvement on the non-analogue written test items of low-ability students who worked with high-ability students ($F(3,74) = 3.06$, $p = .03$). Working with high-ability students brought about greater gains on the written analogue and on the written non-analogue items than working with students of less than high ability or working individually.

The improvement in test performance from Phase 1 to Phases 2 and 3 shown by students who worked individually is evidence of a practice effect.

Table 4

Achievement Test Scores for Low-Medium-Ability Students Who Worked in Different Group Compositions During Phase 2

				Highest Ability Level in the Group								
	Worked Individually			Low-Medium			Medium-High			High		
Achievement Test Scores[a]	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*
Phase 1 ability composite	0.35	0.03	21	0.34	0.03	36	0.37	0.03	23	0.36	0.03	24
Hands-on												
Phase 1 (individual)	0.27	0.13	21	0.29	0.13	32	0.32	0.12	22	0.30	0.16	24
Phase 2 (group)	0.40	0.13	21	0.40	0.16	32	0.55	0.22	22	0.66	0.18	24
Written analogue[b]												
Phase 1 (individual)	0.29	0.13	17	0.29	0.12	31	0.31	0.11	20	0.34	0.14	19
Phase 3 (individual)	0.43	0.20	17	0.37	0.15	31	0.57	0.19	20	0.63	0.17	19
Written non-analogue												
Phase 1 (individual)	0.38	0.16	17	0.36	0.18	31	0.50	0.10	20	0.40	0.20	19
Phase 3 (individual)	0.50	0.17	17	0.44	0.17	31	0.52	0.19	20	0.61	0.17	19

[a]Proportion Correct

[b]Analogue to hands-on test; see text

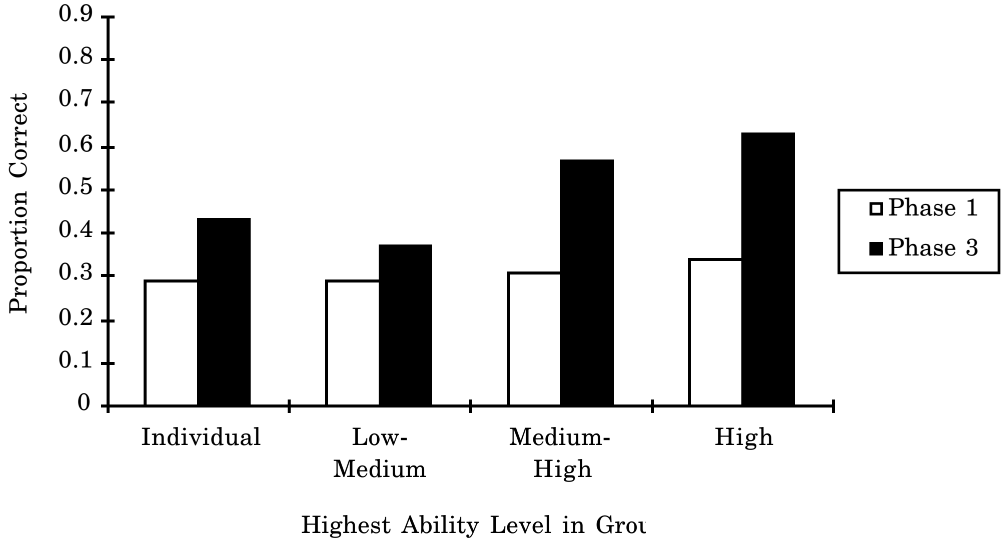


Figure 4. Mean hands-on scores in Phases 1 and 2: Low-medium students in different group compositions.

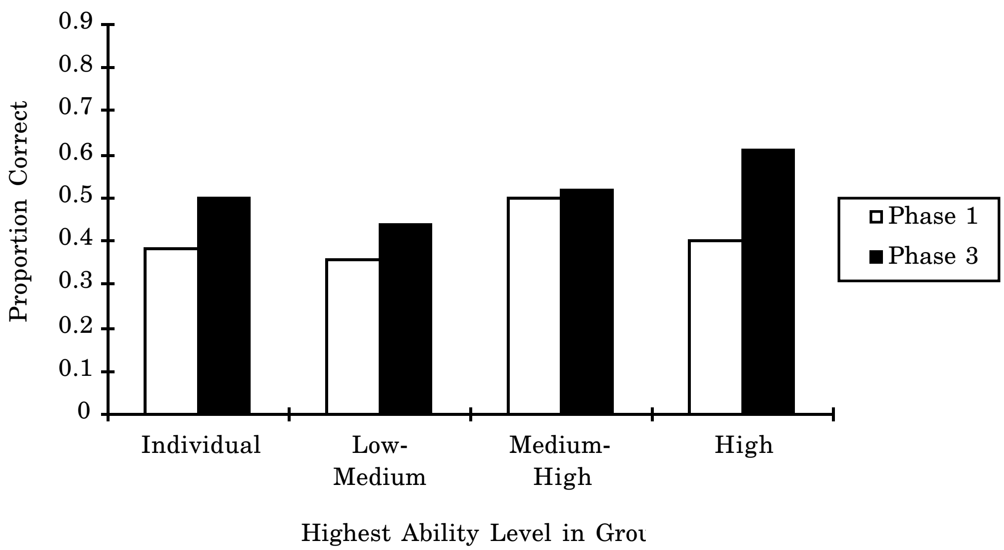


Figure 5: Mean written analogue scores in Phases 1 and 3: Low-medium students in different group compositions.

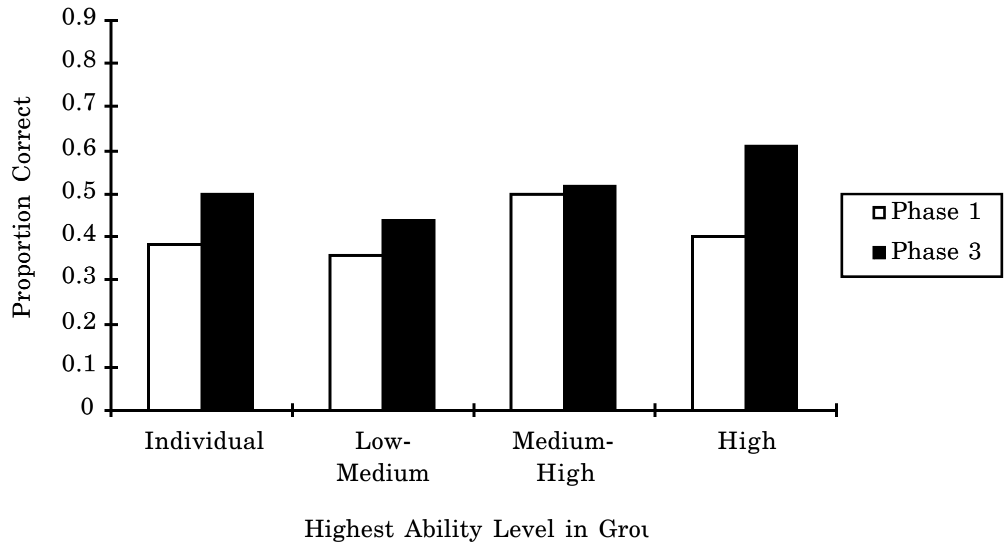


Figure 6. Mean written analogue scores in Phases 1 and 3: Low-medium students in different group compositions.

Nearly all students showed improvement in their scores the second time they took the test (Phases 2 and 3) regardless of whether they worked individually or in groups during Phase 2. However, for all group compositions except homogeneous low-ability, working in groups brought about greater gains than working individually.

Low–medium-ability students. Table 4 and Figures 4, 5, and 6 present the results for the low–medium-ability students (those in the second quarter on the composite ability variable). Working with above-average students (both medium–high-ability and high-ability) benefited low–medium-ability students' performance on the group hands-on test ($F(2,73) = 11.48$, $p < .0001$), and carried over to their scores on the individually performed written analogue ($F(2,65) = 13.82$, $p < .0001$) and non-analogue items ($F(2,65) = 5.43$, $p = .007$). Working with below-average students provided no benefit over working alone for low-medium-ability students.

Medium–high-ability students. Table 5 and Figures 7, 8, and 9 present the results for medium–high-ability students. Working with high-ability students was beneficial for group hands-on test performance ($F(1,85) = 8.82$, $p = .004$) and

Table 5

Achievement Test Scores for Medium-High-Ability Students Who Worked in Different Group Compositions During Phase 2

				Highest Ability Level in the Group					
	Worked Individually			Medium-High			High		
Achievement Test Scores[a]	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*
Phase 1 ability composite	.49	.05	24	.49	.05	48	.50	.04	49
Hands-on									
Phase 1 (individual)	.50	.16	23	.48	.17	46	.49	.17	43
Phase 2 (group)	.60	.15	23	.58	.20	46	.70	.11	43
Written analogue[b]									
Phase 1 (individual)	.47	.13	20	.46	.19	36	.40	.16	46
Phase 3 (individual)	.56	.24	20	.62	.17	36	.67	.17	46
Written non-analogue									
Phase 1 (individual)	.40	.22	20	.53	.15	36	.54	.17	46
Phase 3 (individual)	.62	.18	20	.64	.18	36	.69	.15	46

[a]Proportion Correct
[b]Analogue to hands-on test; see text

Figure 7. Mean hands-on scores in Phases 1 and 2:

Medium-high students in different group compositions.

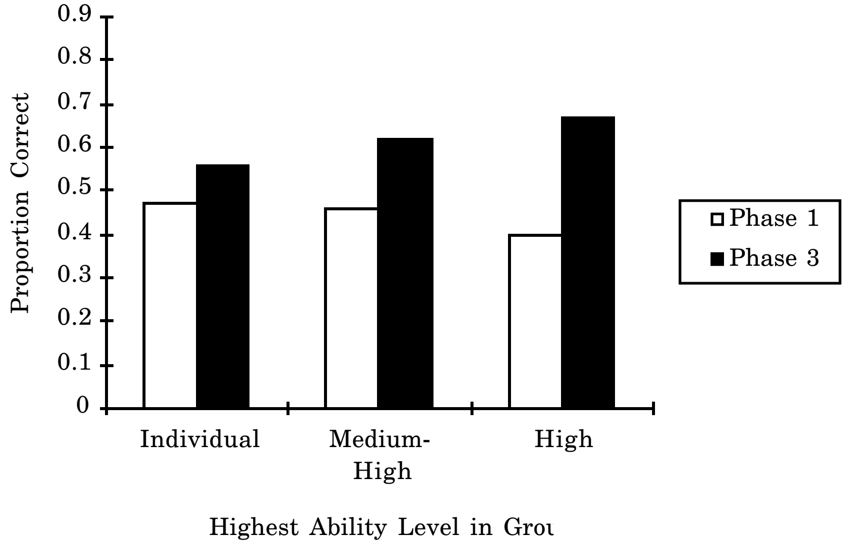


Figure 8. Mean written analogue scores in Phases 1 and 3:
Medium-high students in different group compositions.

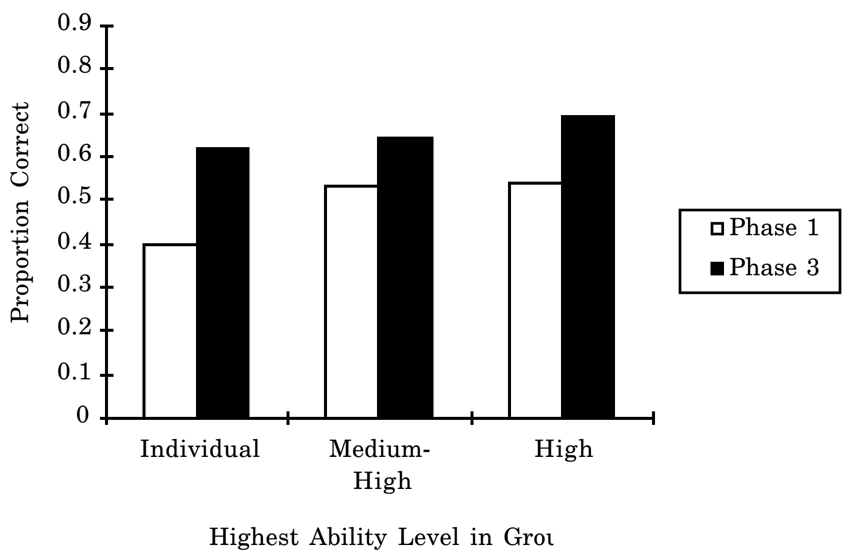


Figure 9. Mean written non-analogue scores in Phases 1 and 3:
Medium-high students in different group compositions.

Table 6

Achievement Test Scores for High-Ability Students Who Worked in Different Group Compositions During Phase 2 (Excluding School 2)

	Worked Individually			Lowest Ability Level in the Group: Low			Low-Medium			Medium-High			High		
Achievement Test Scores[a]	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*
Phase 1 ability composite	.68	.09	14	.68	.09	8	.65	.04	15	.70	.09	26	.68	.08	9
Hands-on															
Phase 1 (individual)	.69	.15	14	.75	.11	8	.66	.16	15	.71	.15	26	.65	.18	9
Phase 2 (group)	.73	.14	14	.72	.15	8	.68	.17	15	.71	.11	26	.80	.10	9
Written analogue[b]															
Phase 1 (individual)	.65	.24	11	.65	.18	8	.61	.20	16	.65	.21	24	.68	.16	9
Phase 3 (individual)	.76	.12	11	.71	.17	8	.75	.11	16	.72	.13	24	.79	.12	9
Written non-analogue															
Phase 1 (individual)	.69	.17	11	.68	.14	8	.66	.21	16	.73	.19	24	.75	.13	9
Phase 3 (individual)	.81	.23	11	.72	.13	8	.78	.09	16	.75	.15	24	.79	.12	9

[a]Proportion Correct

[b]Analogue to hands-on test; see text

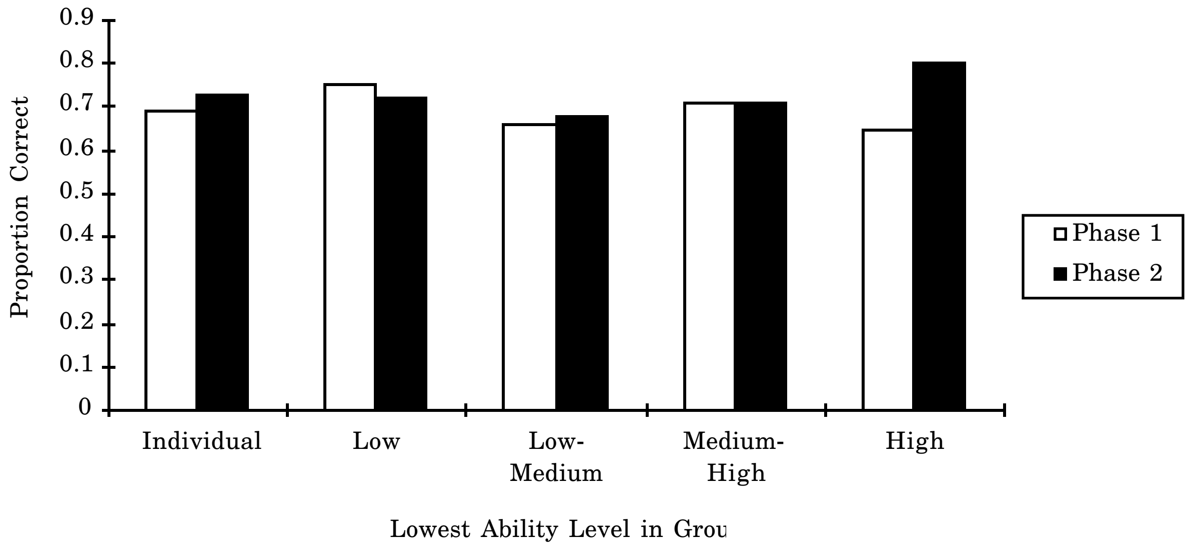


Figure 10. Mean hands-on scores in Phases 1 and 2: High-ability students in different group compositions (School 2 excluded).

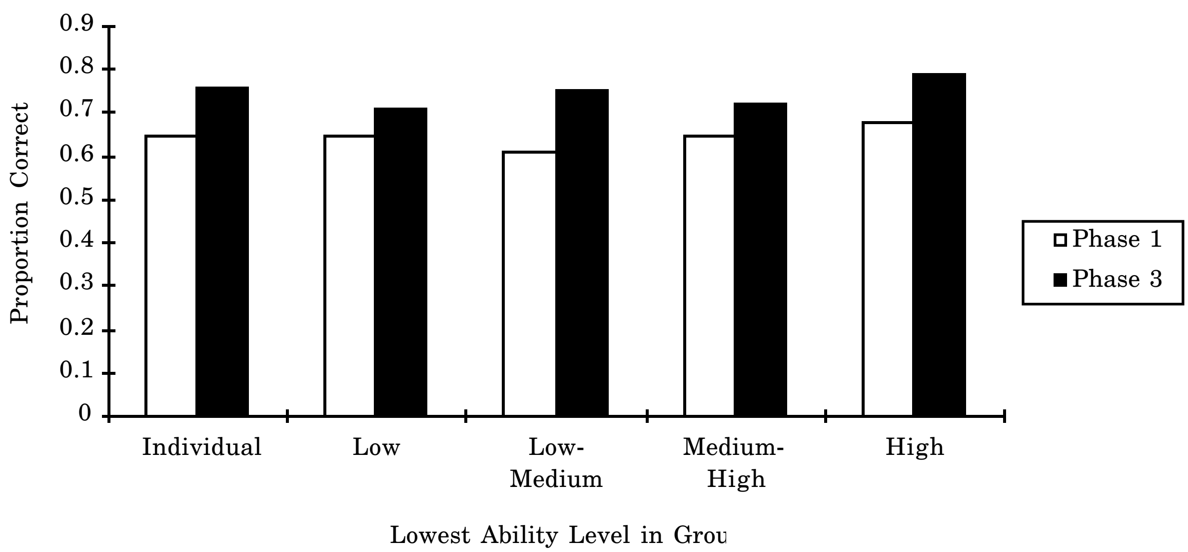


Figure 11. Mean written analogue scores in Phases 1 and 3: High-ability students in different group compositions (School 2 excluded).

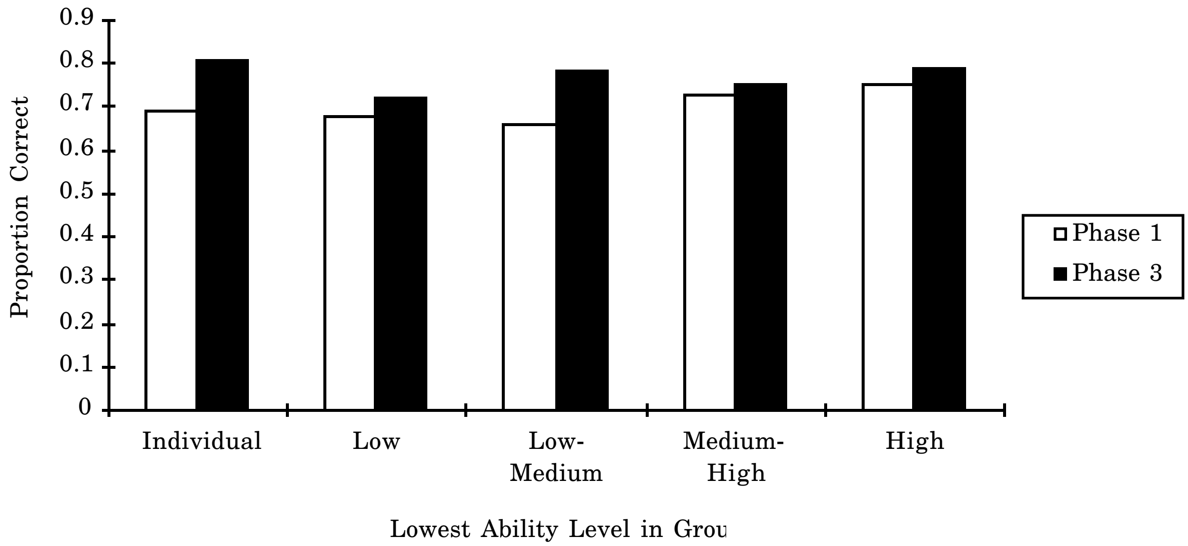


Figure 12. Mean written non-analogue scores in Phases 1 and 3: High-ability students in different group compositions (School 2 excluded).

on the individual written analogue portion ($F(1,78) = 3.54$, $p = .02$) but did not carry over to the non-analogue portion of the test ($F(1,78) = 1.87$, $p = .18$).

High-ability students. Because the majority of homogeneous–high groups (70%) came from the highest-achieving school, the results for high-ability students from this school (in which homogeneous–high groups were concentrated) are presented separately from those of the other schools. Table 6 and Figures 10, 11, and 12 present the results for all high-ability students except those from the highest-achieving school. As this table shows, none of the differences in group compositions were statistically significant (group hands-on test: $F(3,52) = 1.53$, $p = .22$; individual written analogue: $F(3,51) = 1.84$, $p = .15$; individual written non-analogue: $F(3,51) = 0.93$, $p = .43$). Although high-ability students in homogeneous–high groups showed a tendency toward higher mean performance than high-ability students in other group compositions on the group hands-on test, this result was not statistically significant due to the small number of homogeneous–high groups.

Table 7 and Figures 13, 14, and 15 present the results for high-ability students at the highest-achieving school. A significant difference among group compositions appeared for the group hands-on test ($F(3,44) = 9.39$, $p < .0001$) and for the individual written analogue items ($F(3,42) = 7.29$, $p < .0001$) but not for the written non-analogue items ($F(3,42) = 1.02$, $p = .39$). The significant difference among group compositions on the hands-on and written test are due to the low scores of high-ability students who worked in groups where the lowest student had medium–high ability. Post hoc comparisons of group compositions on the hands-on test and the written analogue scores showed that high-ability students who worked in groups in which the lowest student had medium–high ability performed significantly lower than high-ability students who worked in other group compositions ($t = 3.43$, $p = .001$ and $t = 3.56$, $p = .001$, respectively). A later section of this paper will explore possible explanations for this result. Furthermore, differences between working in groups and working individually were not statistically significant for any test for high-ability students (hands-on test: $F(1,53) = 0.61$, $p = .44$; written analogue items: $F(1.51) = 1.65$, $p = .21$; written non-analogue items: $F(1,51) = 0.57$, $p = .46$). Overall, then, the results for high-ability students in Tables 4 and 5 suggest that neither working in groups, nor specifically working in groups with lower-ability students, disadvantages high-ability students.

In summary, these results show that for below-average students (low-ability and low–medium-ability), working in a group with above-average students (medium–high-ability or high-ability) conferred a great advantage over working in a below-average group for both group performance and, more importantly, for performance on the individual test. For medium–high ability students, working in a group with high-ability students conferred an advantage for performance on the group test and written analogue items but not for performance on the non-analogue items. For high-ability students, while there were significant differences between some group compositions, working in a group with lower-ability students did not pose a disadvantage compared to working in a group with higher-ability students or compared to working alone.

Quality of Group Discussion in Different Group Compositions

One possible explanation for the increased scores of below-average students (low-ability and low–medium-ability students) who worked in groups with above-average students, compared to those who worked with below-average students, is that they were exposed to higher quality group discussion. Furthermore, an argument that might be posed against heterogeneous grouping is that high-ability students will be exposed to lower quality work in

Table 7

Achievement Test Scores for High-Ability Students Who Worked in Different Group Compositions During Phase 2 (School 2 Only)

	Worked Individually			Lowest Ability Level in the Group											
				Low			Low-Medium			Medium-High			High		
Achievement Test Scores[a]	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*
Phase 1 ability composite	.72	.07	7	.74	.09	5	.74	.09	6	.70	.06	19	.75	.06	21
Hands-on															
Phase 1 (individual)	.75	.22	7	.81	.10	5	.67	.18	6	.66	.10	18	.73	.12	21
Phase 2 (group)	.83	.08	7	.76	.08	5	.81	.14	6	.71	.11	18	.89	.07	21
Written analogue[b]															
Phase 1 (individual)	.72	.21	7	.62	.28	5	.78	.14	6	.72	.12	18	.73	.17	19
Phase 3 (individual)	.86	.05	7	.83	.13	5	.82	.11	6	.72	.10	18	.89	.07	19
Written non-analogue															
Phase 1 (individual)	.66	.25	7	.80	.09	5	.81	.12	6	.71	.11	18	.79	.10	19
Phase 3 (individual)	.79	.15	7	.83	.06	5	.86	.13	6	.80	.10	18	.88	.07	19

[a]Proportion Correct

[b]Analogue to hands-on test; see text

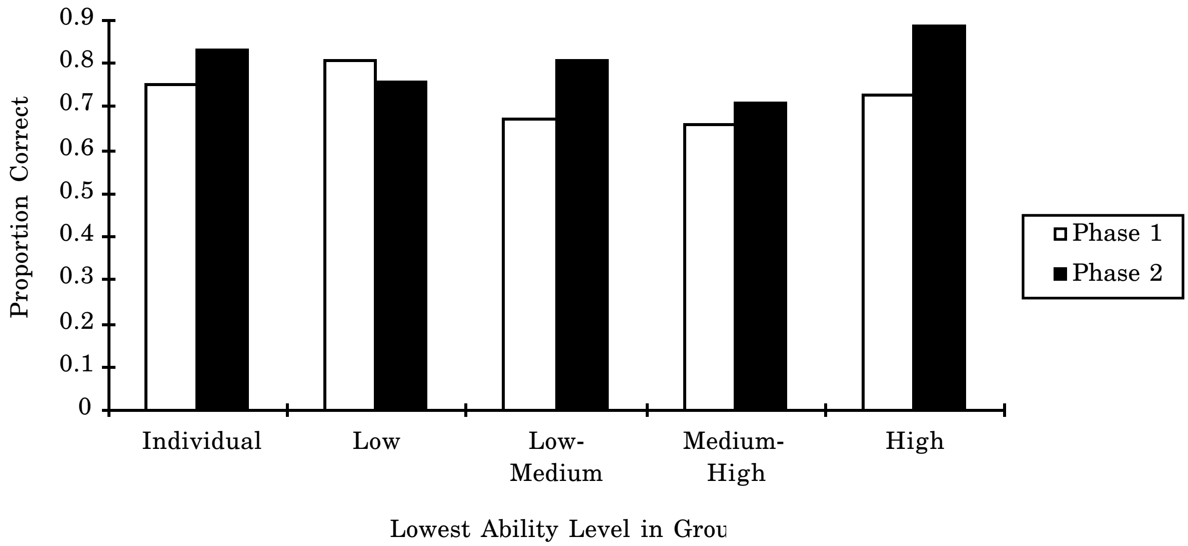


Figure 13. Mean hands-on scores in Phases 1 and 2: High-ability students in different group compositions (School 2 only).

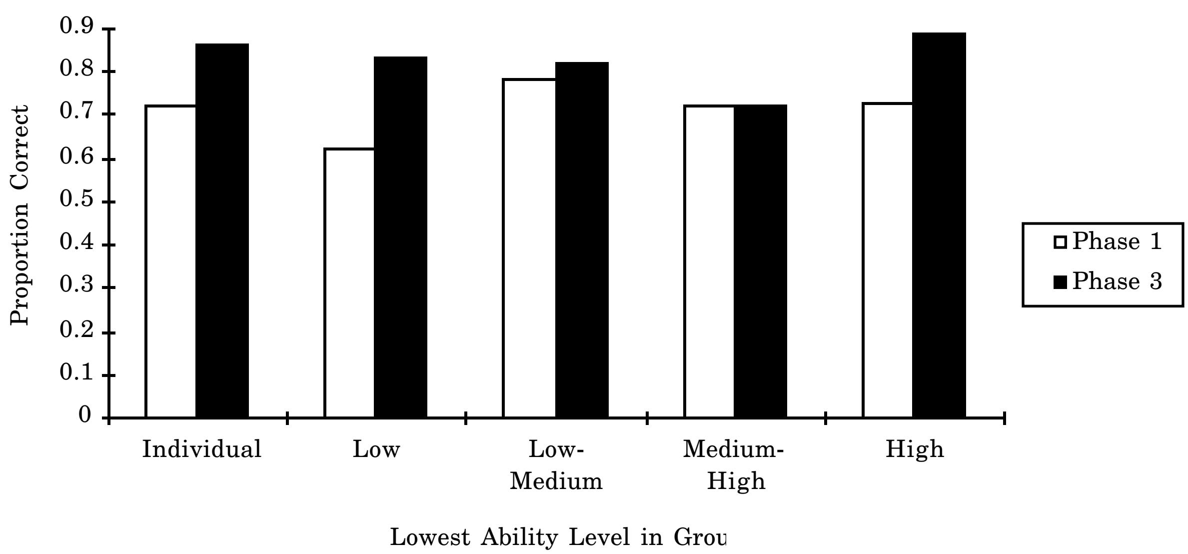


Figure 14. Mean written analogue scores in Phases 1 and 3: High-ability students in different group compositions (School 2 only).

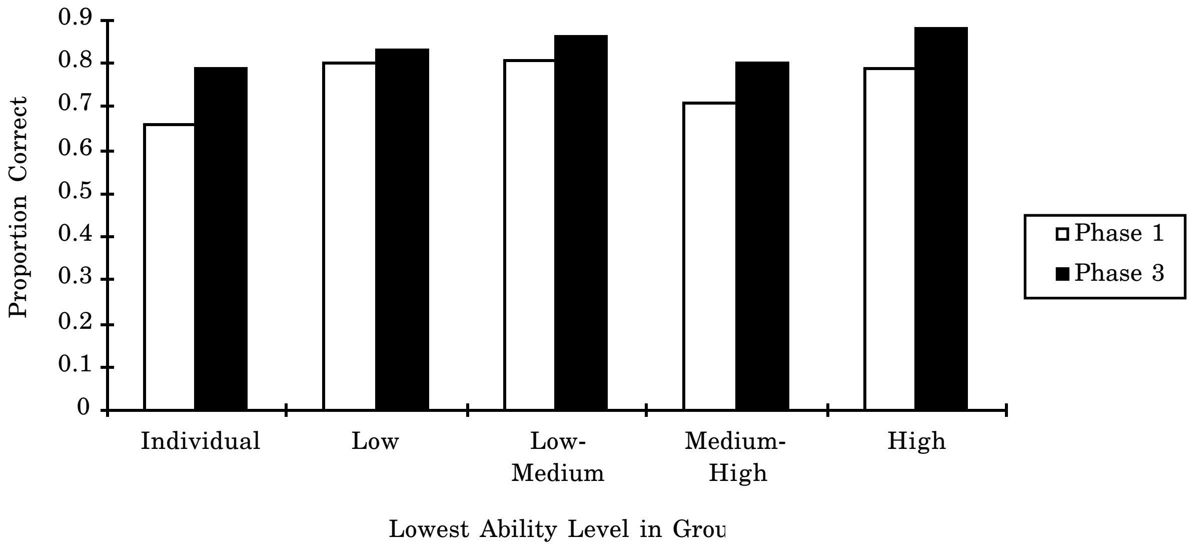


Figure 15. Mean written non-analogue scores in Phases 1 and 3: High-ability students in different group compositions (School 2 only).

heterogeneous groups. To test these possibilities, the level of answer accuracy and the quality of explanations generated were examined for each group composition.

To examine the relationship between group composition and the quality of group discussion, we calculated correlations between the two group discussion variables (accuracy of answers and quality of explanations) and the maximum ability level and minimum ability level in the group (the highest and lowest score of a group member on the ability composite). The maximum and minimum scores were more precise than the four–category ability level variables used in the previous section. Because there was one score per group on both accuracy of answers and quality of explanations, the group was the unit of analysis.

Table 8 gives the correlations between maximum and minimum ability in the group and the accuracy of answers produced in the group and the quality of explanations generated in the group. For groups with low-ability students, groups with low–medium-ability students, and groups with medium–high-ability students, the maximum ability in the group was significantly related to

Table 8

Correlations[a] Between Group Ability Level and Quality of Group Discussion

	Accuracy of Answers Produced by Group[b]		Quality of Explanations Generated in Group[c]	
	Minimum Ability[d]	Maximum Ability[e]	Minimum Ability	Maximum Ability
Low-ability (n=51)[f]	-.10	.37**	-.01	.43**
Low-medium-ability (n=65)	.16	.49***	.20	.50***
Medium-high ability (n=64)	.15	.42**	.03	.39**
High-ability (n=45)	-.06	.03	-.08	.10

[a]Unit of analysis is the group.

[b]Proportion of answers correct.

[c]Completeness of explanations: see text for details.

[d]Lowest score of a group member on the Phase 1 ability composite.

[e]Highest score of a group member on the Phase 1 ability composite.

[f]Number of groups containing this ability level.

* $p < .05$ ** $p < .01$ *** $p < .001$

both indicators of the quality of group discussion. These results show that students in the lowest three quarters were exposed to more high–quality group discussion in groups with relatively able students than in groups with less able students.

In groups with high-ability students, neither the maximum ability level in the group nor the minimum ability level in the group were related to the quality of group discussion. These results show that high-ability students were exposed to the same quality of group discussion whether or not their groups contained students of lower ability.

Relationship Between Quality of Group Discussion and Student Achievement

Test Scores

The results in the previous two sections showed that below-average students in the lowest three quarters performed better on the achievement tests

Table 9

Zero-Order and Partial[a] Correlations Between Quality of Group Discussion and Achievement Test Scores

	Hands-On (Phase 2)		Written Analogue (Phase 3)		Written Non-Analogue (Phase 3)	
Quality of Group Discussion	*r*	Partial *r*	*r*	Partial *r*	*r*	Partial *r*
Low-Ability Students (n = 74,66)[b]						
Accuracy of answers produced by group[c]	.49***	.54***	.33**	.34**	.24*	.24
Quality of explanations generated in group[d]	.55***	.57***	.47***	.49***	.42***	.42***
Low-Medium-Ability Students (n = 62,56)						
Accuracy of answers produced by group	.61***	.57***	.38**	.36**	.32**	.23
Quality of explanations generated in group	.70***	.67***	.51***	.49***	.37**	.30*
Medium-High-Ability Students (n = 65,57)						
Accuracy of answers produced by group	.36**	.29*	.23	.13	.11	.03
Quality of explanations generated in group	.45***	.40**	.29*	.21	.14	.07
High-Ability Students (n = 70,66)						
Accuracy of answers produced by group	.07	.04	.16	.10	.17	.16
Quality of explanations generated in group	.20	.21	.30*	.28*	.23	.17

[a]Controlling for Phase 1 ability composite and Phase 1 science achievement score.

[b]First n: Hands-on test. Second n: Written test.

[c]Proportion of answers correct.

[d]Completeness of explanations: see text for details.

* $p < .05$ ** $p < .01$ *** $p < .001$

when they worked in groups with higher-ability, rather than lower-ability, students, and that they were exposed to a higher quality of group discussion in the higher-ability group compositions. This section examines whether students' achievement test scores were significantly related to the quality of group discussion.

Table 9 presents the zero-order and partial correlations between the group discussion variables (accuracy of group answers and quality of explanations) and students' scores on the achievement tests (score on the group hands-on test, Phase 3 individual written analogue score, Phase 3 individual written non-analogue score). The partial correlations control for ability to help disentangle the effects of group discussion from the effects of their previous ability on students' subsequent achievement.

As can be seen in Table 9, the patterns of correlations between quality of group discussion and student achievement differ across ability levels. For low-ability and low–medium-ability students, the accuracy of answers and the quality of explanations generated in the group usually were significant predictors of their achievement test scores both during and after group work. For medium-high students, the group discussion variables were significant predictors of scores on the group hands-on test but not scores on the individual written test. For high-ability students, the group discussion variables did not significantly predict scores on any of the achievement tests. (Because the results for high-ability students were similar for the highest–achieving school and the other schools, Table 8 presents the results for all high-ability students combined.)

The results in Table 9, then, show that the quality of group discussion was a significant predictor of achievement for below-average ability students but was usually not a significant predictor of achievement for above-average ability students. To provide additional information on the relative contributions of students' ability scores and the quality of group discussion on their achievement scores, regression analyses were performed predicting achievement scores from two ability scores (the Phase 1 ability composite and the score on the specific Phase 1 individual test that corresponded to the test used as the outcome in the regression equation) and the two quality of group discussion scores (accuracy of answers and quality of explanations generated in the group). Table 10 presents multiple regression coefficients when all predictors were entered into the equation along with the proportion of variance in achievement test scores (r-squared) accounted for by students' ability and quality of group discussion. For low-ability and low–medium-ability students, the quality of group discussion usually contributed more to their achievement test scores than did their own ability scores. Typically the strongest predictor of their achievement test scores was usually the quality of explanations generated in the group. For medium–high-ability and high-ability students, in contrast, students' ability scores usually contributed more to their achievement test scores than did the quality of group discussion. Below-average students, who had not mastered much of the material, had more to gain from the group discussion than above-average students, who already had some competence in the subject matter.

Students' Participation in Group Discussion

Below-average students. The results in the previous sections suggest that below-average students benefited from the higher quality of group discussion in groups with above-average students. Not only did they show higher scores on the group test, but they also seemed to learn from the group discussion and apply what they learned to the next day's individual test. This section examines the behavior of students in the group to shed more light on how this may have occurred.

Table 11 lists the types of behavior that below-average students exhibited in groups with above-average students. This analysis was based on 28 groups: the group from each classroom that was purposely formed with one high-scoring student and two low-scoring students plus 7 additional groups selected at random that had low-scoring and high-scoring students. Because the achievement scores of the students included in Table 11 are very similar to those of the entire sample in Tables 1 and 2, the students included in this analysis are representative of the whole sample. Table 11 shows that on most items, below-average students participated in group discussion at a high level: making suggestions or defending or elaborating suggestions that other students had made, asking questions of other students, and paraphrasing other students' suggestions. Out of eight items that were discussed in the group, below-average students engaged in high-level discussion on an average of 4.86 (or 61%) items. The predominance of high-level participation appeared for both low-ability students (56%) and low-medium-ability students (66%).

Table 10

Regression Coefficients Predicting Achievement Test Scores from Student Ability and Quality of Group Discussion

	Hands-On (Phase 2)			Written Analogue (Phase 3)			Written Non-Analogue (Phase 3)		
Predictor	*b*	*B*	R^2 Chg.	*b*	*B*	R^2 Chg.	*b*	*B*	R^2 Chg.
Low-Ability Students									
Phase 1 scores			.01			.05			.10*
Ability composite	-.15	-.04		1.10*	.28*		.46	.15	
Science test	.47*	.23*		-.20	-.11		.21	.20	
Group discussion scores			.35***			.23***			.18**
Answers[a]	.16	.26		-.04	-.07		-.09	-.19	
Explanations[b]	.37*	.39*		.50**	.54**		.42**	.55**	
Low–Medium-Ability Students									
Phase 1 scores			.11			.02			.36***
Ability composite	-.04	-.01		.16	.02		.15	.02	
Science test	-.17	-.09		-.02	-.01		.62***	.54**	
Group discussion scores			.40***			.24***			.06
Answers	.07	.11		-.06	-.10		-.01	-.02	
Explanations	.51***	.59***		.49**	.58**		.21	.27	

Table 10 (continued)

Regression Coefficients Predicting Achievement Test Scores from Student Ability and Quality of Group Discussion

	Hands-On (Phase 1)			Written Analogue (Phase 3)			Written Non-Analogue (Phase 3)		
Predictor	*b*	*B*	R^2 Chg.	*b*	*B*	R^2 Chg.	*b*	*B*	R^2 Chg.
Medium–High-Ability Students									
Phase 1 scores			.10*			.16**			.27***
Ability composite	.79	.18		1.51*	.37*		1.21*	.29*	
Science test	-.09	-.08		-.29	-.26		.35**	.34**	
Group discussion scores			.14**			.04			.00
Answers	.02	.03		-.01	.03		-.02	-.04	
Explanations	.34*	.38*		.18	.15		-.08	-.09	
High-Ability Students									
Phase 1 scores			.14**			.26***			.19**
Ability composite	.72***	.41***		.90***	.56***		.19	.14	
Science test	-.14	-.15		-.11	-.18		.20*	.33*	
Group discussion scores			.05			.06			.03
Answers	-.07	-.12		-.02	-.10		.03	.07	
Explanations	.17	.26		.13	.30*		.06	.11	

Note. *b* and *B* are unstandardized and standardized multiple regression coefficients, respectively.

[a]Proportion of answers correct.

[b]Completeness of explanations: see text for details.

* $p < .05$ ** $p < .01$ *** $p < .001$

Table 11

Mean Behavior of Below-Average Students in Groups with Above-Average Students

	All Students (n = 44)		Low Students (n = 23)		Low-Medium Students (n = 21)	
Level of Participation in Group Discussion	Number of Items[a]	Percent of Items	Number of Items	Percent of Items	Number of Items	Percent of Items
High[b]	4.86	61	4.48	56	5.29	66
Makes or defends suggestion	3.43	43	2.91	36	4.000	50
Asks question	2.07	26	2.17	27	1.95	24
Paraphrases other student's suggestion	0.59	7	0.43	5	0.76	10
Medium	1.02	13	1.3	16	0.71	9
Repeats verbatim or copies another student's answer	0.59	7	0.74	9	0.43	5
Agrees with another student's suggestion	0.43	5	0.56	7	0.29	4
Low: listens or watches	0.68	9	0.52	7	0.86	11
Works alone	1.02	13	1.09	14	0.95	12
Unknown or no group discussion[c]	0.42	5	0.61	8	0.19	3

Note. Percents do not always sum to 100 due to rounding.

[a]Groups worked on 8 items.

[b]Some students engaged in more than one category of high-level participation on each item.

[c]Student either copied another student's work or worked alone, or the group did not complete the task.

The example below demonstrates how participating at a high level seems to help a below-average student learn from group discussion. In this excerpt, the group was working on the question "Which circuit has higher resistance? Why?"; the correct answer was Circuit A because it contained graphite resistors (the two circuits were the same in all other respects). The low-ability student challenged and questioned the high-ability student, as well as paraphrasing his suggested answer:

High student: "A [has the highest resistance]."

Low student: "A doesn't have higher resistance."

High student: "Graphites stop current."

Low student: "Because what? The graphites slow down the current?"

High student: "Graphites resist current."

Low student: "No, this [Circuit A] has higher resistance because of all the graphite. The graphite rods slow down the current."

High student: "It resists the current. The graphite resists it."

On the individual test prior to group work, the low-ability student erroneously linked resistance to voltage and did not recognize the relevance of the graphite. On the Phase 3 individual test, the low-ability student gave a correct response describing how the graphite increased the circuit's resistance, showing that he had learned from group work.

On only a minority of items did below-average students participate at less than a high level. Students repeated other students' answers verbatim, copied answers onto their papers, or agreed with other students' suggestions without elaboration on 13% of the items. Students watched or listened to group discussion without making any contribution on 9% of the items. Students worked alone, without seeming to pay attention to group discussion, on 13% of the items.

The results of this section show that below-average students in heterogeneous groups generally exhibited behavior that is conducive to learning: making or defending suggestions, asking questions, and paraphrasing other students' suggestions. They were less likely to exhibit non-active behavior such as repeating or copying other students' answers, or watching and listening without contributing to the group discussion.

High-ability students. The results in previous sections showed no systematic effect of group composition on the achievement of high-ability students, but did reveal the unexpected finding that, at the school with an overall highly achieving student population, high-ability students working in groups in which the lowest student was a medium-high student performed worse than high-ability students working in other group compositions. This section examines the behavior of high-ability students in different group compositions at that school in order to explain this puzzling result.

Table 12 gives the mean behavior of high-ability students at the high-achieving school for different group compositions. The sample sizes for the group compositions in Table 12 are slightly lower than those in Table 7 (which reports achievement test results) because some students were not videotaped (out of approximately eight groups in each classroom, five were videotaped). Because the achievement scores of the students included in Table 12 are very similar to those of the entire sample in Table 7, the students who were videotaped are representative of the whole sample.

In contrast to the achievement results, Table 12 shows that high-ability students working in groups in which the lowest student had medium-high ability did not show a pattern of participation that was different from high-ability students working in other group compositions. In fact, high-ability students working in groups in which the lowest student had medium-high ability showed similar patterns of participation to high-ability students working in homogeneous-high groups (contrasts between these two group compositions on all participation variables were not statistically significant).

The significant differences in Table 12 concerned other group compositions. Specifically, high-ability students who worked in groups with below-average students (low or low-medium) exhibited more high-level participation than did high-ability students who worked in groups with only above-average students (medium-high or high; $F(3,30) = 4.44$, $p = .01$). This result confirms those of a previous study which found that high-ability students in heterogeneous groups assumed a teaching and explaining role and participated more actively than did high-ability students in homogeneous groups (Webb, 1980). That earlier study

Table 12

Mean Behavior of High-Ability Students in Groups (School 2 Only)

Level of Participation in Group Discussion	Lowest Ability Level in the Group: Low ($n = 5$[a])		Low-Medium ($n = 2$[a])		Medium-High ($n = 7$)		High ($n = 19$)	
	Number of Items[b]	Percent of Items	Number of Items	Percent of Items	Number of Items	Percent of Items	Number of Items	Percent of Items
High[c]	6.60	83	5.00	63	3.11	39	4.61	58
Makes or defends suggestion	6.40	80	5.00	63	2.78	35	3.89	49
Asks question	1.80	23	.50	6	.89	11	.94	12
Paraphrases other student's suggestion	.20	3	2.00	25	.11	1	.11	1
Medium	.20	3	2.50	31	.67	8	.56	7
Repeats verbatim or copies another student's answer	.00	0	.50	6	11.00	1	.28	4
Agrees with another student's suggestion	.20	3	2.00	25	.56	7	.28	4
Low: listens or watches	.40	5	.50	6	1.67	21	1.61	20
Works alone	.80	10	.00	0	2.56	32	1.11	14

Note. Percents do not always sum to 100 due to rounding.

[a]This number represents all students in this group composition that were videotaped.

[b]Groups worked on 8 items.

[c]Some students engaged in more than one category of high-level participation on each item.

found that achievement results mirrored students' behavior: high-ability students working in heterogeneous groups learned more than those high-ability students working in homogeneous groups. Second, Table 12 shows that high-ability students who worked in groups with low-medium students exhibited more medium-level participation than did the other group compositions ($F(3,30) = 4.52$, $p = .007$). Finally, high-ability students who worked in groups in which the lowest student was above-average (medium–high-ability or high-ability) worked alone more often than high-ability students who worked in groups in which the lowest student was below average (low-ability or medium–low-ability; $F(1,30) = 3.51$, $p = .03$).

In the current study, the lower rate of participation corresponded to lower achievement for high-ability students who worked with medium-high ability students, but not for high-ability students who worked only with others of high ability. Why the relationship between participation and achievement held for high-ability students who worked with medium–high-ability, but not high-ability, students is unclear. It could be a chance result, given the small number of students analyzed. If this result is replicated in future studies, students should be interviewed to provide insights into their behavior and achievement.

Achievement and Behavior of "Middle" Ability Students

The results described above suggest that, for students at most ability levels (especially below-average), working in heterogeneous groups proves more beneficial than working in homogeneous groups due to the greater likelihood of having an able student in the group. While this conclusion is consistent with previous findings concerning low-ability students, it contradicts some previous findings concerning medium-ability students. As noted in the introduction to this paper, previous research has shown that medium-ability students may participate less and learn less in heterogeneous groups with a wide range of ability than in homogeneous groups (groups with all medium-ability students) or in heterogeneous groups with a narrow ability range (medium-ability and low-ability students in a group; or medium-ability and high-ability students in a group). Previous studies have shown that medium-ability students may be left out of the teacher-learner relationships that develop in heterogeneous groups, and that they tend to participate more actively in groups that are more homogeneous (where there are no clear "highs" who may become "teachers" and no clear "lows" who may become "learners").

To investigate whether the results in the present study contradict those in previous studies, this section makes similar comparisons among group compositions. Specifically, this section compares the achievement and behavior of students in the middle two quarters who worked in (1) homogeneous groups (all students were at the same ability level), (2) groups in which they were the lowest students (one or both of the other students in the group were at a higher ability level), (3) groups in which they were the middle students (the other two students in the group were at a lower and higher ability level, respectively), and (4) groups in which they were the highest students (one or both of the other students in the group were at a lower ability level). To be consistent with previous research, students working in groups in which they were the "middle" student would have had to participate less and learn less than students working in the other group compositions.

Table 13 presents the mean achievement test scores for low-medium students and medium-high students according to their position in the group. In contrast to past research, students who were in the "middle" of heterogeneous groups did not perform worse than students who were the highest or lowest in their groups or who worked in homogeneous groups. Rather, the results in Table 13 are consistent with the finding from Tables 2 and 3 that performance was related to the group's highest ability level, regardless of the student's position in the group. Low-medium-ability students did better in groups where they were the lowest student or the middle student (groups with at least one above-average student) than in groups where they were the highest student or in homogeneous groups (groups without any above-average students). The differences in achievement of medium-ability students across the four positions in the group were statistically significant for all tests (hands-on test: $F(3,72) = 10.25$, $p < .0001$; written analogue: $F(3,64) = 13.65$, $p < .0001$; written non-analogue: $F(3,64) = 2.85$, $p = .04$). For medium-high-ability students, position in the group had a significant effect only on the group hands-on test: they did less well in groups in which they were the most able, rather than the middle or lowest, member ($F(3,83) = 3.53$, $p = .02$). On the other achievement tests, position in the group was not significantly related to achievement test scores. These results show that being the "middle" student in the group was not a disadvantage.

Table 13

Mean Achievement Test Scores for Low-Medium and Medium-High Students by Relative Position in the Group

			Position in the Group in Heterogeneous Groups					
	Homogeneous Groups		Lowest		Middle		Highest	
Achievement Test Scores[a]	*M*	*n*	*M*	*n*	*M*	*n*	*M*	*n*
Low–Medium-Ability								
Phase 1 ability composite	0.34	6	0.36	31	0.37	15	0.34	26
Phase 2 hands-on	0.34	6	0.66	31	0.49	15	0.42	26
Phase 3 written analogue	0.37	5	0.66	26	0.48	13	0.37	26
Phase 3 written non-analogue	0.33	5	0.60	26	0.50	13	0.46	26
Medium–High-Ability								
Phase 1 ability composite	0.49	3	0.51	35	0.49	8	0.49	43
Phase 2 hands-on (Phase 2)	0.62	3	0.68	35	0.75	8	0.58	43
Phase 3 written analogue	0.58	2	0.66	37	0.72	9	0.62	34
Phase 3 written non-analogue	0.70	2	0.69	37	0.69	9	0.63	34

[a]Proportion Correct

As noted above, one reason why medium-ability students in previous studies have shown lower performance in the "middle" of heterogeneous groups than in other group compositions is that they participated less actively. To compare the behavior of "middle" students in the current study with "middle" students in previous studies, Table 14 presents the behavior of a random sample of low–medium-ability students in different group compositions. As can be seen in Table 14, the profiles of middle student behavior were comparable to students with other positions in their groups (differences across positions in the group were not statistically significant on any participation variable: F statistics ranged from $F(3,29) = 0.23$, $p = .87$ to $F(3,29) = 1.95$, $p = .14$). Most relevant for learning, the frequency of high-level participation was similar across the group positions of students. Students who were in the "middle" of their groups, then, showed just as much high-level participation in group work as did students who were "highest" or "lowest" in heterogeneous groups or who worked in homogeneous groups.

In contrast to previous research, then, students who were in the "middle" of heterogeneous groups did not participate less than students of comparable ability who worked in homogeneous groups or who were the highest or lowest in heterogeneous groups. There are at least two reasons for this result. First, a teacher-learner relationship between high and low students that excludes middle students may take more time to develop than the 50 minutes provided for group work in the present study. Second, the fact that groups were working on an assessment rather than problems in a normal classroom setting may have motivated students in the middle of their groups to be more active than they would be otherwise.

Conclusions

In this study, group ability composition had a major impact on all performance and process variables. For all students except those in the highest quarter, working in groups with high-ability students produced higher scores during group problem solving and sometimes higher scores on the individual test administered the next day. The effect of group composition was especially strong for the students in the lowest two quarters (low-ability and low–medium-ability). For these students, working in groups with at least one above-average

Table 14

Mean Behavior of Low-Medium-Ability Students by Relative Position in the Group

			Position in the Group in Heterogeneous Groups[b]					
	Homogeneous Groups (n = 5)		Lowest (n = 11)		Middle (n = 10)		Highest (n = 7)	
Level of Participation in Group Discussion	Number of Items[a]	Percent of Items	Number of Items	Percent of Items	Number of Items	Percent of Items	Number of Items	Percent of Items
High[b]	3.80	48	5.55	69	5.00	63	4.86	61
Makes or defends suggestion	3.00	38	4.27	53	3.70	46	4.29	54
Asks question	2.20	28	2.18	27	1.70	21	1.86	23
Paraphrases other student's suggestion	.60	8	.90	11	.60	8	.14	2
Medium	.80	10	.91	11	.50	6	.71	9
Repeats verbatim or copies another student's answer	.60	8	.55	7	.30	4	.14	2
Agrees with another student's suggestion	.20	3	.36	5	.20	3	.57	8
Low: listens or watches	.80	10	1.45	18	.20	3	.71	10
Works alone	1.80	23	.09	1	1.90	25	1.95	24

Note:. Percents do not always sum to 100 due to rounding.

[a]Groups worked on a maximum of 8 items; some groups did not finish all 8 items.

[b]Some students engaged in more than one category of high-level participation on each item.

student was a huge advantage over working in groups with only below-average students. Working with above-average students helped below-average students produce greater performance on the group test *and* higher scores on the subsequent individual test. Analyses of group discussions and individual students' contributions revealed, first, that groups with above-average students gave more correct answers and more high-quality explanations, and, second, that below-average students learned from the high-quality discussions through high-level participation in group discussion (e.g., making and defending suggestions, asking questions, and paraphrasing other students' suggestions). Below-average students did not merely copy the work of more-able group members.

Importantly, performance of high-ability students was not related to the composition of the group. Working with less-able students did not pose a disadvantage for high-ability students.

These findings raise serious questions about the fairness of using different group compositions on assessments with collaborative group work. The implication for practice is that, to give all students the same advantage, all students should be in a group with a high-achieving student. Classroom-to-classroom and school-to-school differences in distributions of achievement suggest, however, that manipulating group composition to produce equally fair groups across classrooms and schools may be impossible. In the present study, some schools had relatively few high-achieving students (see School 4 in Table 1), limiting the opportunities available for low-achieving students to work with a high-achiever.

Rather than manipulating group composition, which may be an intractable problem, it may be more productive to prepare students for group assessments by giving them practice and training in the processes that lead to high performance. These processes include effective communication skills, high-level elaboration and discussion of ideas, questioning of each others' ideas, and promoting active participation by all group members (Webb & Palincsar, 1996). Giving students skills to work effectively in groups can help them make the best use of the available resources.

While in the present study group composition had a strong effect on performance, an important question for further research is whether the effects of group composition depend on the nature of the assessment tasks. Although the tasks used here (assembling electric circuits with certain properties and answering questions about the circuits' relative voltage, resistance, and current) were difficult and complex for students in this sample, the tasks are theoretically unidimensional (as shown by the high internal consistency reliability coefficients) and unquestionably have correct answers. A competent and knowledgeable student could individually complete all items successfully. As described in the introduction, the tasks used here have well-structured solutions and answers that can be clearly explained and demonstrated. The importance of having a competent group member in this study is consistent with the previous work of social psychologists using well-structured tasks (e.g., Hastie, 1986; Laughlin & Branch, 1972; Laughlin & Ellis, 1986; Moreland & Levine, 1992). On tasks with ill-structured solutions and no clearly correct answer, Cohen (1994; Cohen & Cohen, 1991) argues that heterogeneous groups with multiple perspectives and areas of competence will be more important than having a single competent student in the group. Which group compositions are most advantageous for group performance and individual learning on ill-structured tasks and whether such group compositions can be formed equitably across classrooms and schools are questions still to be studied.

Also to be resolved are inconsistencies between some results of this study and those of previous research. In some previous studies, medium-ability students working in heterogeneous groups learned less than comparable students working in homogeneous groups due to being excluded from the teacher-learner relationships that developed between high-ability and low-ability students (see Webb, 1991). In the present study, in contrast, students in the "middle" of heterogeneous groups did not participate less or learn less than comparable students working in other group compositions or those who worked individually. As noted earlier, reasons for the discrepancy may be that the short duration of group work in the present study did not allow time for teacher-learner relationships to develop and that working on an assessment may serve as a motivation for all students to participate in group work. Whether being in the middle of heterogeneous groups on assessments of longer duration or in groups with a longer history of working together have a negative effect remain to be explored.

In conclusion, this study showed that collaborative group work on assessments poses a significant advantage for below-average students if the group has a student who is knowledgeable and competent in the subject matter. Through increased access to intellectual resources—correct answers and high-quality explanations—low-ability students working with high-ability peers produced higher-quality solutions to problems on the group test and on the subsequent individual test than students working with less competent students or those working alone. While group work can be an advantage for many students, the composition of the group introduces a possible source of inequity that is not present in individual testing situations. Students of comparable ability may perform very differently depending on the composition of the group. One possible recommendation is to make sure that every group has an "expert." But not all classrooms and schools have enough experts to spread around. How to resolve this dilemma is unclear. At the very least, however, group composition must be taken into account when interpreting and comparing scores of different students, classrooms, or schools.

References

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.

Azmitia, M. (1988). Peer interaction and problem solving: When are two heads better than one? *Child Development, 59,* 87-96.

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, LA.

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

Bossert, S. T. (1988-1989). 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.

Buros, O. K. (1974). *Tests in Print II.* New Jersey: Gryphon Press.

California State Department of Education (1985). *Mathematics Framework for California Public Schools, Kindergarten Through Grade Twelve*. Sacramento, CA: California Department of Education.

California State Department of Education (1992). *Mathematics Framework for California Public Schools, Kindergarten Through Grade Twelve*. Sacramento, CA: California Department of Education.

Cohen, B. P., & Cohen, E. G. (1991). From groupwork among children to R&D teams: Interdependence, interaction, and productivity. *Advances in Group Processes, 8,* 205-225.

Cohen, E. G. (1994) Restructuring the classroom: Conditions for productive small groups. *Review of Educational Research, 64,* 1-36.

Connecticut State Board of Education (1987). *Common Core of Learning*. Hartford, CT: Connecticut State Board of Education.

Connecticut State Board of Education (1992). *Connecticut Academic Performance Test*. Hartford, CT: Connecticut State Board of Education.

Fall, J. R., Webb, N. M., & Wise, N. (1995, April). *Group discussion and large-scale language arts assessment: Effects on students' comprehension.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Hackman, J. R. (1990). *Groups that work (and those that don't): Creating conditions for effective teamwork*. San Francisco: Jossey-Bass.

Hastie, R. (1986). Experimental evidence on group accuracy. B. Grofman, & G. Guillermo (Eds.), *Information pooling and group decision making*. Greenwich, CT: JAI Press.

Hooper, S., & Hannafin, M. J. (1988). Cooperative CBI: The effects of heterogeneous versus homogeneous grouping on the learning of progressively complex concepts. *Journal of Educational Computing Research, 4,* 413-424.

Hooper, S., Ward, T. J., Hannafin, M. J., & Clark, H. T. (1989). The effects of aptitude composition on achievement during small group learning. *Journal of Computer-Based Instruction, 16,* 102-109.

Kansas State Board of Education (1993). *Kansas Curricular Standards for Science.* Topeka, KS: Kansas State Board of Education.

Laughlin, P. R., & Branch, L. G. (1972). Individual versus tetradic performance on a complementary task as a function of initial ability level. *Organizational Behavior and Human Performance, 8,* 201-216.

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology, 22,* 177-189.

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.

Lou, Y., Abrami, P. C., Spence, J. C., Poulsen, C., Chambers, B., & d'Apollonia, S. (1996). Within-class ability grouping: A meta-analysis. *Review of Educational Research, 66*, 423-458.

Mathematical Sciences Education Board, National Research Council. (1993). *Measuring Up: Prototypes for Mathematics Assessment*. Washington, DC: National Academy Press.

Moreland, R. L., & Levine, J. M. (1992). The composition of small groups. *Advances in Group Processes, 9,* 237-280.

Mugny, G., & Doise, W. (1978). Socio-cognitive conflict and structure of individual and collective performances. *European Journal of Social Psychology, 8,* 181-192.

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.

O'Neil, H. F., Allred, K., & Baker, E. L. (1992). *Measurement of workforce readiness: Review of theoretical frameworks* (CSE Technical Report 343). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

Salas, E., Dickinson, T. L., Converse, S. A., & Tannenbaum, S. I. (1992). Toward an understanding of team performance and training. In R. W. Swezey & E. Salas (Eds.), *Teams: Their Training and Performance* (pp. 132-160). Norwood, NJ: Ablex.

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.

SCANS (1991, June). *What work requires of schools. A SCANS report for America 2000.* Washington, DC: U.S. Department of Labor, The Secretary's Commission on Achieving Necessary Skills (SCANS).

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.

Shipman, V. C. (1985). *The New Jersey Test of Reasoning Skills*. Totowa, NJ: Totowa Board of Education.

Skon, L., Johnson, D. W., & Johnson, R. T. (1981). Cooperative peer interaction versus individual competition and individualistic efforts: Effects on the acquisition of cognitive reasoning strategies. *Journal of Educational Psychology, 73,* 83-92.

Slavin, R. E. (1990). *Cooperative Learning: Theory, Research, and Practice*. Englewood Cliffs, NJ: Prentice-Hall.

Webb, N. M. (1980). A process–outcome analysis of learning in group and individual settings. *Educational Psychologist, 15,* 69-83.

Webb, N. M. (1991). Task-related verbal interaction and mathematics learning in small groups. *Journal for Research in Mathematics Education, 22,* 366-389.

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

Webb, N. M., & Farivar, S. (in press). Developing productive group interaction in middle school mathematics. In A. M. O'Donnell & A. King (Eds.), *Cognitive Perspectives on Peer Learning.*

Webb, N. M., & Palincsar, A. S. (1996). Group processes in the classroom. In D. Berliner & R. Calfee (Eds.), *Handbook of Educational Psychology* (pp. 841-873). New York: Macmillan.

Webb, N. M., Troper, J. D. , & Fall, R. (1995). Constructive activity and learning in collaborative small groups. *Journal of Educational Psychology, 87,* 406-423.

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

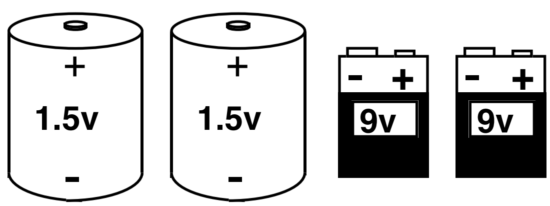
APPENDIX

Example Test Questions

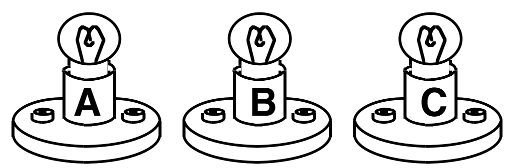
1. (a) Use the items drawn below (batteries and bulbs) to draw **two circuits** in the boxes labeled Circuit A and Circuit B. **Follow these rules:**

- **Bulb A should be in Circuit A. Bulb B should be in Circuit B.**
- **Bulb A should be brighter than Bulb B**.
- **There should be one 9-volt battery in each circuit.**
- **You must draw the wires needed to connect up the items in each circuit.**
- **Use all of the items but do not use any item more than once.** For example, if you put Bulb C in Circuit A, you cannot also put it in Circuit B.

Batteries



Bulbs



Draw the circuits in these boxes:

Circuit A (brighter)	Circuit B (dimmer)

1. (b) Why will Bulb A in Circuit A be brighter than Bulb B in Circuit B? (Try to use scientific terms in your answer.)

1. (c) Which of the two circuits you drew has the **highest voltage**?

Circle one: **CIRCUIT A** **CIRCUIT B** **BOTH CIRCUITS HAVE THE SAME VOLTAGE**

Why? (Try to use scientific terms in your answer.)

____

____

____

1. (d) Which of the two circuits you drew has the **highest resistance**?

Circle one: **CIRCUIT A** **CIRCUIT B** **BOTH CIRCUITS HAVE THE SAME RESISTANCE**

Why? (Try to use scientific terms in your answer.)

____

____

____

1. (e) Which of the two circuits you drew has the **highest current**?

Circle one: **CIRCUIT A** **CIRCUIT B** **BOTH CIRCUITS HAVE THE SAME CURRENT**

Why? (Try to use scientific terms in your answer.)

____

____

____