The effects of teacher knowledge of rubrics on achievement of high school students were studied. Pairs of teachers were selected from system nominations based on course (algebra, biology, English, and government) and ability level of students taught. One member of each pair was randomly assigned to receive rubric training and the other did not. After a period of instruction, all teachers administered subject matter tests that contained both selected response and constructed-response (CR) items. Results favored the achievement of students whose teachers had received rubric training in 3 of the 8 (2 item types in each of 4 content areas) comparisons; the other 5 comparisons were not significant. The significant results were in algebra for both item types and in biology for CR items.

Careful attention to specification of achievement targets is thought to have utility for both teachers and students (Stiggins, 1997). Students, Stiggins and Conklin (1992) suggested, use the assessment patterns of their teachers to predict what they
will be asked to do in the future. Teacher assessments accordingly are thought to act as operational definitions of achievement for students. If assessments are the vehicle by which students understand achievement goals, then the more clearly they are stated in behavioral terms, the more accurately they will be understood by each student and the greater the consistency of those understandings among groups of students (Wiggins, 1998). These concepts suggest that assessments may be utilized as a mechanism of instruction. In their synthesis of research into formative assessment interventions, Black and Wiliam (1998) concluded that effective assessments are indivisible from instruction. They also may influence student motivation (Brookhart, 1997).

Similarly, teachers who understand achievement goals in assessment (i.e., behavioral) terms are thought to be better equipped than those who do not to design effective instructional experiences (Airasian, 1994). They are also believed to be better able to utilize assessments as part of instruction by explaining learning goals to students through descriptions of the activities they will be expected to engage in, improving accuracy of student understandings of achievement goals and criteria (Shepard et al., 1996). Indeed, Khattri, Reeve, Kane, and Adamson (1996) found evidence that teachers use descriptions of levels of achievement with their students to build proficiency.

Rubrics are vehicles that can provide a useful mechanism to translate achievements into assessment terms. As used here, a rubric is a rating scale that consists of ordered categories, together with descriptions of criteria that may include exemplars, which are used to sort student-produced responses into levels of achievement. Mehrens, Popham, and Ryan (1998) argued that teachers who are clear about the criteria that will be used to identify high achievement levels are better able to design instruction to help students attain those levels. They also recommended that the criteria, expressed as rubrics, be communicated to students. Agreeing with Arter (1993), they suggested that the rubrics should be applicable to a broad range of tasks (generic) as opposed to descriptions of achievement levels that provide evidence for a single task (item specific). They then can be shared, studied, and used in providing feedback to help students understand how to conceptualize high levels of achievement.

The purpose of this research was to study the hypothesis that enhanced teacher knowledge of generic assessment rubrics results in improved student achievement. The hypothesis is based on the speculation that teachers who have better understandings about what is expected will make effective use of that information. In so doing, teachers may design more effective instructional experiences that capitalize on criteria expressed in the rubrics and may describe to students the relevant characteristics of demonstrations of achievement. They may more effectively monitor student progress and provide formative feedback. Students, in turn, may be better able to participate in ongoing evaluation of their own work (self-monitoring of achievement).
This research was one of a series of studies intended to be of use to the Maryland State Department of Education (MSDE) in designing mandatory, end-of-course high school assessments for the state. It had been determined that certain core learning goals (CLGs) would be assessed, and three item formats would be used: selected response (SR), brief constructed response (BCR; 5 to 10 min), and extended constructed response (ECR; 15 to 30 min). Initial assessments were being developed for courses in biology, English, government, and mathematics.

The end-of-course assessments were anticipated to replace an existing series of four assessments that were required for high school graduation as part of the Maryland Functional Testing Program (MSDE, 1980). Two of the tests, in reading and mathematics, tested content and skills that are clearly below the level of those demanded by the CLGs in algebra and English, and most students in the state had passed those tests by the time they entered the ninth grade. The other two existing tests, one in writing and one in citizenship, have more important implications about this study.

The Maryland Writing Test (MSDE, 1990b) consists of writing samples to two prompts that are varied across test administrations. They are scored using generic rubrics that are shared with English teachers throughout the state. There were, therefore, common understandings among teachers of English prior to this study about the nature and use of generic rubrics to a greater extent than there were in any of the other academic fields.

The Maryland Test of Citizenship Skills (MTCS; MSDE, 1990a) was an SR test generally associated with the ninth-grade citizenship course taken at the time in all state public school systems. It consisted primarily of fact-based questions about definitions and principles. The CLGs in government were intended to introduce more reasoning skills into the curriculum, but students in the citizenship courses at the time of this study were nevertheless being prepared in their citizenship courses to take the MTCS.

As part of the development process, generic rubrics were under development to score constructed-response (CR) items in each of four content areas to be tested with the new end-of-course high school assessments. MSDE and local experts decided to develop generic rubrics, in part, to facilitate their use in instruction by teachers and students. Several teacher committees had used the rubrics with student-produced responses to many items in attempts to select exemplars of the rubrics’ score values, to learn from those attempts, and to revise the rubrics. The intent of MSDE, in this study, was to gain experience in using all three item formats in each of the four content areas. In addition, the instructional value of generic scoring rubrics was to be studied through a workshop taken by one half of the participating teachers. These same rubrics would be used for scoring all CRs in the study. It was known that the rubrics would be further modified for eventual use.
with the high school assessment program. There was considerable motivation on
the part of teachers and their supervisors throughout the state to learn as much as
possible about these aspects of the assessments along with the nature of the CLGs.

METHOD

Participants

Several groups of professionals were involved at various phases in this study. Many
participated in developing materials; they are described in the “Materials” section.
Others were teachers who received the experimental manipulations; they are de-
dscribed here.

All school districts in the state were invited to nominate pairs of teachers in
each of the four content areas to participate. The pairs of teachers were to have stu-
dents with the same ability level, similar teaching and training background, and
classes with similar demographics. The teacher pairs were to be teaching the same
course (algebra I, biology, English, or citizenship/government). After the pairs
were selected, one of the two teachers in each pair was randomly assigned to re-
ceive rubric training, the experimental manipulation, and the other did not.

Initially, each of the 24 local education agencies (LEAs) in the state partici-
pated in three of the four content areas except for one, which participated in two.
Because a few teachers dropped out late in the study, ultimately 4 LEAs partici-
pated in only two content areas.

In making selections of which content areas would be represented in which
LEAs, factors such as ability levels of classes and number of schools were taken into
consideration. Some LEAs nominated only higher or average ability classes and
some nominated pairs of classes that were not matched in ability level (or we learned
by talking with teachers that their nominated classes could not necessarily be catego-
rized by ability). Because we wanted our samples to be representative of all ability
levels (about 50% of average ability and the rest evenly divided between higher and
lower), in each of the content areas we were unable to include in the study those
LEAs that did not provide an appropriate selection of student abilities. In addition,
we wanted matched classes to come from different schools that were comparable in
demographics, rather than from the same school, to avoid control group contamina-
tion. In some of the smaller LEAs, that was not possible, and we used teachers with
matched classes from the same school (one pair in biology, two in government, and
three in algebra) so that the LEA could be included in the study.

After selections and manipulations had been accomplished, but before the crite-
rion tests were administered, some teachers withdrew from the study for reasons
that ranged from scheduling conflicts, to long-term absences for surgical proce-
dures, to lack of continuing motivation. This produced a few single, unmatched
classes in each of the content areas. Although we administered the tests to the stu-
dents in these classes, we did not include them in the analyses. Of an original 71 dyads, there were 46 (65%) pairs of teachers who provided complete data. Approximately two thirds of these 92 teachers were women; approximately 85% were White, not of Hispanic origin; the remainder were African American, except for one teacher. About two thirds of the teachers had between 1 and 5 years of teaching experience. Less than 5% were first-year teachers, and less than 5% had more than 30 years of experience. The rest were about evenly distributed in 5-year ranges from 6 to 10 years through 26 to 30 years.

The compositions of the classrooms that we included in the final analyses for each content area are described in Table 1. Attrition rates for classroom pairs ranged from a low of 29% (algebra) to a high of 40% (government). The average class size ranged from a low of 32.3 (algebra) to a high of 38.1 (biology). The only major ethnic groups represented in sizeable numbers were White (not of Hispanic origin) and African American. The latter ranged from a low of 18% (biology) to a high of 23% (English and government). The percentage of women ranged from 46% (algebra) to 56% (biology). Almost all the English students were ninth graders. The algebra students were predominantly in the ninth grade (79%), with a few (15%) in the tenth grade. Similarly, ninth graders predominated in government (70%), with some (24%) in the tenth grade. The biology students were older, with 63% tenth graders and 36% ninth graders.

Materials

**CLGs.** The CLGs are intended to be the essential skills and knowledge that should be expected of Maryland high school graduates. They were developed by
content teams and reviewed by educators and the public at large to guide local curricula throughout the state. The Maryland high school CLGs for English, mathematics, science, and social studies define the body of knowledge on which Maryland’s eventual high school assessment exams will be based. They have been approved by the State Board of Education.

The CLG documents for each of the four content areas are organized into goals, expectations, and indicators of learning. Goals represent the broad areas of content that students need to master. Expectations identify topics or concepts within the goal areas. Indicators of learning address specific details of content that, when assessed, demonstrate student mastery.

A fifth CLG area, called skills for success (SFS), contains learning, thinking, communication, interpersonal, and technology skills. SFS are intended to define what high school students need to learn in addition to the knowledge and skills identified in their academic subjects. SFS attempts to represent a focus that is general rather than subject specific, applies equally well to all subjects, describes a context for the use of academic knowledge and skills, and provides tools for learning in any subject or skill area. They are to be infused into each of the high school assessments, but not assessed or scored separately.

Content area rubrics. A generic rubric was developed in each of the content areas for use in instructional activities and in scoring student CRs. The same rubric was used for brief as well as extended responses. When teachers were trained in the use of rubrics, exemplars from actual student responses to both types of CRs were incorporated into the rubrics to illustrate the criteria as described in each of the score points.

Rubric development occurred at MSDE from the late summer of 1997 through early 1998 in each content area. Based primarily on Maryland’s experience with statewide testing programs, especially the Maryland Writing Test, the content specialists and the high school scoring lead decided early in the process to develop generic rather than item-specific rubrics. Driving this decision was the idea of establishing goals that could be used in instruction and that have a direct link to the scoring of the high school assessments. This rationale is consistent with the recommendations of Arter (1993) and Mehrens et al. (1998).

Teams of four to six teachers and content specialists from the LEAs brainstormed criteria for development of rubrics that would be appropriate for the high school assessments. The rubric writers used the draft rubrics in their classrooms to test the rubrics and harvest sample student responses to illustrate the qualities of the criterion that the score point descriptors contained. The rubric writers and other educators then reconvened to select, through consensus and independent scoring, appropriate exemplars and to refine the rubrics and verify that they were appropriate. From this work, 5-point scoring rubrics (0 to 4) were developed in each of the content areas. These appear in the Appendix.
Achievement tests. The study design called for tests of student achievement to be administered in May. Prior to the test item development, each content specialist had identified indicators from the CLGs that would reasonably be taught in or by April. These became the domains of the tests.

In algebra, the content included recognition, description, and extension of patterns; representing patterns and functional relations in tables, graphs, and mathematical expressions; adding, subtracting, multiplying, and dividing algebraic expressions; and description of graphs of nonlinear functions including maxima, minima, roots, limits, rate of change, and continuity.

In biology, the content included characteristics of chemical compounds and macromolecules utilized by living systems, structures of cellular and multicellular organisms, the identification and transmission of genetic traits, relations between abiotic factors and biotic diversity, and changes in environmental conditions and their effects on the dynamics of populations.

In English, the content included composition of written presentations that inform, persuade, and express personal ideas; use of prewriting, drafting, and revision strategies of effective writers and speakers; and location, retrieval, and use of information from various sources to accomplish a purpose.

In government, the content included the structure and functions of government and politics in the United States, evaluation of how the U.S. government has maintained balance between protecting rights and maintaining order, how government influences monetary policy, and the influence of demographic changes on government policies.

Three test forms were developed in each of the four content areas. Each form consisted of five SR items, two BCR items, and one ECR item. Each could be administered in a 45-min block within a class period.

Six teachers in each of the four content areas served as item writers for the study. Prior to item writing they received training in thinking skills, SFS, construction of open-ended questions, nature and use of scoring rubrics, Maryland’s rubrics, Maryland’s CLGs (presented in small groups by MSDE content area specialists), construction of SR items, and characteristics of effective versus ineffective items. These item writers and new teams of reviewers (including teachers and an instructional supervisor in each content area), a representative for SFS, a special educator, and a teacher of English as a second language revised the items for final use.

From this work, three similar test forms in each content area were constructed for administration in May. The forms were not developed to differ in any systematic way. In three of the content areas, the forms consisted entirely of nonoverlapping items. In biology, there were only four BCRs that were judged acceptable, and they were rotated such that each pair of forms had exactly one BCR in common.

The five SR items on each form were expected to require about 5 min of testing time per form. The CR items were expected to require about 40 min. Therefore, the
reliabilities for the SR subtests were expected to be, and in general were, lower than those for the CR subtests. The alpha homogeneity reliabilities of the SR subtests in algebra were .35, .37, and .52; and the CR subtests yielded alphas of .56, .62, and .73. In biology, the alpha coefficients were .16, .23, and .29 for the SR subtests; and .53, .63, and .68 for the CR subtests. After testing, one BCR was deleted from one of the English forms because scorers noted that very many students had clearly misunderstood the item. The alpha coefficients for the SR subtests in English were .34, .48, and .67; and for the CR subtests, they were .31, .69, and .69, the lowest from the form for which only two items were scored. In government, the SR subtest alpha coefficients were .14, .32, and .42; and the CR subtests yielded alphas of .64, .70, and .75.

These reliability estimates are lower than would be acceptable for decision making about students. However, for our study of effects between groups, they were judged reasonable because statistical power was developed through multiple replications, both within and across classrooms pairs, of the primary study design. It should be noted that the low reliabilities of our short criterion measures implies that the observed effect sizes should be considered underestimates because measurement error contributes to their denominators.

Procedures

The experimental manipulations occurred during March of 1998 at a local community college for 2 days. The first day of workshop training was attended by all teachers who participated in the study (described in the “Participants” section). The CLGs and appropriate instructional strategies were covered. All teachers were told they would be administering tests provided by MSDE, and they were told the specific content from the CLGs that the tests would cover (see Achievement tests, previously mentioned). A total of 71 pairs of teachers participated in the initial training.

One teacher from each pair had been randomly assigned by coin flip to receive the second day-long workshop on the use of rubrics as instructional tools. These teachers became the experimental group. Only the experimental group attended the second day of the training. Because we did not have available a workshop that we felt would not introduce irrelevant and confounding differences between the two groups, the teachers who were not selected were excused for the day of rubric training and became the control group.

CLGs training. Teachers from both groups were brought together at a local community college to receive training that emphasized the indicators that would be tested within their respective content area CLGs and that suggested instructional strategies focusing on the implications of the CLGs for instruction. This was done for three reasons. First, we did not want instruction in general teaching techniques
or familiarity with the CLGs to differ systematically between the two groups. Providing an introductory training session also allowed the later training session for the experimental group to focus on rubrics and their uses as opposed to either instruction in general or the CLGs. Third, all participants received at least some attention through this training. The generic rubrics were not introduced at this session. Although the general emphasis was the same in all content areas, each content area approached this training from a slightly different perspective.

In algebra, the training session focused on functions and algebra. Two content specialists worked with the participants to describe and then to list related implications for instruction. After discussion, the indicators were reviewed with the participants to verify that their understandings were accurate. For each content indicator, sample instructional activities that addressed patterns and functional relations in mathematics were provided. In addition, test items that were appropriate to assessing those indicators were suggested. Many of the instructional activities and assessment items came from textbooks widely used across the state. Instructional activities were selected to provide different representations of the same material using the language of mathematics and appropriate technology.

The biology training session was focused primarily on lesson design. It began by emphasizing that all lessons should be planned and that all test items should be developed making a deliberate attempt to combine science skills and process indicators with actual biology content. The content specialist asked the participants to examine the language of the indicators and to discuss the biology content limits. Participants were introduced to a “5–E” format of lesson planning (engagement, exploration, explanation, extension, and evaluation) and the value of connecting science goals, expectations, and indicators to real-world applications as well as to hands-on, inquiry-based instruction.

In English, the training session focused just on the indicators that would be tested. The content specialist discussed with participants the pertinent goals, expectations, and indicators to provide an overview and clarification of the content they should cover. Participants then brainstormed instructional activities appropriate for those content indicators and how, in general, to link instructional strategies to specific indicators of learning.

Participants in the government CLGs session focused on political systems. The content and SFS specialists prepared model lessons for the social studies teachers combining SFS indicators in thinking skills and communication skills with government expectations. In small groups, participants were then asked to select a pair of indicators (one from SFS and one from government) and to prepare their own model lesson combining the two. Model lessons were shared with the large group, and critiqued feedback was encouraged.

**Rubric training.** The goals for rubric training were to familiarize the teachers with the generic rubrics and to suggest to teachers ways to use the rubrics
The teachers who received the rubrics workshop were asked to incorporate the scoring rubrics into their daily lesson plans, utilizing them for instruction with their students.

The day began with a large group presentation. It included an overview of the characteristics of rubrics, how they can be used as a vehicle for conceptualizing student achievement for a teacher, and how they may be used in instruction to help show students how levels of their own achievement can be demonstrated, recognized, and improved.

The rest of the day was spent in content groups led by content specialists. In each group, the generic rubric was distributed and discussed. They were then given anchor papers that had been developed along with the rubric, and discrepancies between the score points and teacher judgments about the anchor papers were discussed. The teachers then scored training sets of student responses and discussed the application of the rubric and exemplars to assigning an accurate score. Finally, some teaching activities that incorporate rubrics were discussed. These included using examples to help students conceptualize levels of achievement according to the rubric, analyzing with students the language of the rubric and thinking with them about criteria they should use to assign score ratings to examples, peer review in which students revise each others’ work, pairing students whose scores differ by only one point to discuss the differences, and the use of feedback forms that are tied to the rubric. Other uses were brainstormed in each group.

**Test administration and scoring.** Content area exams were administered in May in algebra, biology, English, and government to both groups of students. Each teacher administered the three forms to random thirds of their classes. The papers were randomly ordered prior to forwarding to the teachers, and the teachers were instructed to distribute the exams to their students in the order they were received.

SR items were scanned at MSDE. Teams of content area teachers and specialists, many of whom were involved in the rubric development project, selected sample student answers to typify the upper and lower limits of each rubric score point. Through this process, called rangefinding, anchor responses for each point on the CR rubric were chosen to be consistent with the exemplars used in rubric training.

Scoring guides were produced during rangefinding. Each scoring guide for an item consisted of a series of student responses chosen to illustrate the score points on the rubric. Each response was accompanied by a brief description of why it was rated at that score. As part of training, each scorer participated in practice scoring by rating and discussing training sets prepared from actual student responses.

Qualified teams of readers, hired and trained by MSDE’s commercial scoring contractor, used the scoring guides to score student responses for the CR items with participation of the MSDE content and scoring specialists during the summer.
Scoring was blind; no information identified papers as coming from the experimental group or the control group. Two readers, whose scores were averaged for the item score for that paper, scored each CR response. If the readers differed by more than one score point, their scores were ignored and a new and highly experienced reader assigned the score for that item. Exact agreement for the initial two readers was reached for 74% of the papers in algebra and adjacent agreement for another 25%. The exact agreement rate for biology was 72%, and the adjacent agreement rate was 27%. In English, the exact and adjacent agreement rates were 60% and 38%, respectively. The exact agreement rate for government was 64%, and the adjacent agreement rate was 35%. The generic rubric in each content area had five score values, and the scores on any individual item ranged from 0 to 4. To reflect the longer response time, the score for the ECR was doubled in computing a final CR score for each student. Thus, SR scores could vary from 0 to 5 and CR scores could vary from 0 to 16.

RESULTS

These data could have been analyzed using repeated measures analysis of variance separately for each of the contents using the class means, either for forms separately or summed across forms. That approach was not taken because it would have been impossible to compare the item formats or the contents directly. Instead, each of the teacher pairs was conceptualized as an individual study that yielded six criterion measures, three of each item type, that were used to estimate six effect sizes comparing the student scores of the teacher who had with those of the teacher who did not have the rubric training. The technique of meta-analysis (Hedges & Olkin, 1985) was used to synthesize and model the effect sizes from the several teacher pairs according to the dimensions of the study (content and item format).

Each of the individual effect sizes was the mean of the students whose teacher received rubric training minus the mean of the students whose teacher did not, with the difference divided by the pooled within-groups standard deviation and then adjusted for bias due to finite sample size. The tabled bias-correction factors in Hedges and Olkin (1985, p. 80) were used to apply the adjustment.

Any pair of teachers produced an effect size for each of two item types (CR and SR) for each of three forms (administered to a random third of each class of students), for a total of 6 effect sizes. In English and biology, we had available 6 effect sizes from each of 11 pairs of teachers, yielding 66 in each content, and totaling 132 from those two contents. From algebra and government there were 6 effect sizes for each of 12 teacher pairs, totaling 72 effect sizes for each content, and 144 in both. The grand total of effect sizes was 276.

The existence of several effect sizes from each teacher pair, and the completion of two formats by each student, introduces dependencies into the analysis that vio-
late an assumption of meta-analysis that the effect sizes are independent. Although it is common to ignore lack of independence in meta-analyses, one suggestion for a way to compensate for dependencies is to apply Bonferroni-like control over the familywise Type 1 error rate (Gleser & Olkin, 1994; Schafer, 1999). In this study, the teacher pairs are independent, but there are six effect sizes from each pair. Bonferroni-like control was applied using the adjusted alpha of .05/6 = .008, and 99.2% confidence intervals for effect sizes are provided. In the subsequent analyses, however, no statistical significance decision would have been different had the .05 alpha level been used instead.

The first step of the meta-analysis yielded an overall estimate of effect size by pooling the 276 effect size estimates into one average. In this study, the overall effect size average was .1166. With a standard error of .0342, the overall average effect size of .1166 was significantly different from zero at the .008 level of alpha.

The meta-analysis also yielded a chi-square statistic that evaluates whether there is evidence of heterogeneity of effect sizes about the overall average. The chi-square for heterogeneity was statistically significant, $\chi^2(275, N = 3,191) = 515.309, p < .001$.

Because the effect sizes varied significantly about their average, the effect sizes for the content areas were grouped, and effect size averages were found separately. It was found that the four content area effect sizes differed significantly from each other, $\chi^2(3, N = 3,191) = 20.766, p < .001$. The pooled heterogeneity chi-square was also statistically significant, indicating that there was further variation that remains to be explained around the content area average effect sizes, $\chi^2(272, N = 3,191) = 494.543, p < .001$.

Because the heterogeneity chi-square was significant, further modeling to differentiate the two formats, SR and CR, was undertaken within each content. These results are discussed later. Following separation of the effect sizes into the eight groups by content and format, the combined heterogeneity chi-square remained statistically significant, $\chi^2(271, N = 3,191) = 487.862, p < .001$. However, no further modeling was undertaken because we had exhausted the variables available to us to predict effect size. The results for each of the combinations of content area by item type appear in Table 2.

Within algebra, the SR effect size was .2656, and the standard error was .0948. The CR effect size was .3082, with a standard error of .0953. Because neither confidence interval for algebra spans zero, the effect sizes for both item types were significantly different from zero. The difference between the effect sizes for the formats was not statistically significant, $\chi^2(1, N = 775) = .110, p = .739$.

Within biology, the SR effect size was .0498, with a standard error of .0954. The CR effect size was .4615, and its standard error was .0978. The SR confidence interval spans zero, but the CR interval does not. Therefore, only the CR effect size was significantly different from zero. The narrower 95% interval for the SR effect size was −.1373 to .2369, which also spans zero, showing that the lack of statistical
significance was not due to our use of a Bonferroni-like adjusted alpha. The difference between the effect sizes for the formats was statistically significant, $\chi^2 (1, N = 838) = 9.081, p < .003$.

Within English, the SR effect size was $-.0686$, with a standard error of $.1031$; and the CR effect size was $.0054$, with a standard error of $.1038$. Both confidence intervals span zero, and so neither was significantly different from zero. The 95% interval for the SR effect size was $-.3406$ to $.2034$; and for the CR, effect size was $-.2684$ to $.2793$. Thus, neither would have been significantly different from zero had the unadjusted alpha of 0.05 been applied. The difference between the effect sizes for the formats was not statistically significant, $\chi^2 (1, N = 760) = .257, p = .612$.

The SR effect size within government was $-.1343$, and the standard error was $.0928$. For government, the CR effect size was $.0358$, with a standard error of $.0920$. Because both confidence intervals span zero, neither was significantly different from zero. Had adjusted alpha not been applied, the 95% intervals would have been $-.3162$ to $.0476$ for the SR effect size and $-.1445$ to $.2161$ for the CR effect size, neither being significantly different from zero. The difference between the formats was not statistically significant, $\chi^2 (1, N = 818) = 1.705, p = .192$.

### DISCUSSION

The results of this study provide some empirical support for instructional value in teacher knowledge of rubrics. That support was strongest for CR items in biology but also appeared for both SR and CR formats in algebra. There was no evidence found to support instructional value of teacher knowledge of rubrics in English or in

<table>
<thead>
<tr>
<th>Content</th>
<th>Format</th>
<th>Average Effect Size</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algebra</td>
<td>SR</td>
<td>.2656</td>
<td>.0155</td>
<td>.5157</td>
</tr>
<tr>
<td></td>
<td>CR</td>
<td>.3082</td>
<td>.0567</td>
<td>.5597</td>
</tr>
<tr>
<td>Biology</td>
<td>SR</td>
<td>.0498</td>
<td>-.2020</td>
<td>.3016</td>
</tr>
<tr>
<td></td>
<td>CR</td>
<td>.4615</td>
<td>.2035</td>
<td>.7195</td>
</tr>
<tr>
<td>English</td>
<td>SR</td>
<td>-.0686</td>
<td>-.3406</td>
<td>.2034</td>
</tr>
<tr>
<td></td>
<td>CR</td>
<td>.0054</td>
<td>-.2684</td>
<td>.2793</td>
</tr>
<tr>
<td>Government</td>
<td>SR</td>
<td>-.1343</td>
<td>-.3792</td>
<td>.0476</td>
</tr>
<tr>
<td></td>
<td>CR</td>
<td>.0358</td>
<td>-.2069</td>
<td>.2785</td>
</tr>
</tbody>
</table>

Note. SR = selected response (multiple-choice) items; CR = constructed-response (essay) items. *The confidence coefficient is 99.2% for all intervals.
government, but neither was there support for a position that teacher knowledge of rubrics has detrimental effects on student achievement.

Because there were differences noted between the content areas as well as the item formats an evaluation of the results should treat them separately. In algebra, the overall effect size was approximately 0.3. Although the effect size for the CR items was greater than that for the SR items, they were both in the same general range. Indeed, it can be argued from the algebra results that knowledge of the rubric to be used in evaluation of student work produces a learning advantage that generalizes across assessment formats.

The results for biology suggest a positive effect size of 0.5 for CR items, but no effect was supported for SR items. The magnitude of the difference for CR items was the largest found. It is possible that the use of a generic rubric is very different from the way biology teachers typically grade student CRs, which may be more focused on presence or absence of specific elements that are idiosyncratic to the task. According to that hypothesis, the CR assessments used in the study would likely have been very different from the more checklist-style criteria used by the teachers in the nonrubric condition; that is, the rubric training may have induced a change in the teaching approach of the teachers who received it, which affected students’ reactions to CR items, which demanded higher order thought consistent with the biology rubric.

No support for rubric effects was found for any item type in either English or government. One possible explanation is that teachers in these areas were resistant to instructional change in response to the rubric training. For various reasons, that is a reasonable hypothesis in these two content areas given the context of our study.

As noted earlier, the Maryland Writing Test has been a graduation requirement in Maryland for many years, and it is scored using rubrics that, in the judgment of English specialists at MSDE, are not very different from the generic rubric used here with the English teachers. It may be argued that English teachers in Maryland already use generic rubrics, either explicitly or implicitly, and perhaps even instructionally, in their grading; therefore, the training experience may not have contained much, if any, new information for those teachers.

It has been mentioned that a multiple-choice test of citizenship skills had been a requirement for graduation for students studying government up to and including the year that this study was conducted. That test is based on a highly specific curriculum that does not stress the higher order thinking embodied in the CLGs in government. It is possible that at least some government teachers felt that emphasizing an expanded set of content goals and higher order thinking processes would be a disservice to their students who needed to be successful on the citizenship skills test. At least some of these teachers may therefore have ignored, with respect to their immediate classroom activities, both the CLGs and the rubric understandings provided in this study.
The exact agreement scoring rates for English and government were lower than for algebra and biology. This suggests differences in the clarity of the rubrics that may have affected their value in instruction. Because the clearest results in favor of the rubric group were found in algebra, it is worthwhile to examine the rubric for algebra in comparison with the others. There are five elements that appear throughout each of the four nonzero score values. These are application, representation, explanation, justification, and analysis. Each of these may be capable of being described to a student more easily than the terminology used in the other content area rubrics, such as understanding, extending meaning, tone, and insight. It is possible that the latter terms, whose meanings may require greater elaboration and perhaps cause greater confusion, are less useful instructionally than those in the mathematics rubric. Degree of abstraction in terminology could be varied systematically in future research as well as effects of the use of analytic scoring and other rubric modifications.

To interpret the magnitudes of the effects found, it should be recalled that the effect size in this study represents the advantage, in standard deviation units, of a 1-day educational experience about rubrics for teachers. Interpretation of the magnitude of the effect size index was discussed by Cohen (1988), who gave guidelines of 0.2 for small effects, 0.5 for medium effects, and 0.8 for large effects. A small effect size of 0.2 is the sort of magnitude Cohen felt would occur in a new field of inquiry. Lack of understandings that are needed in developing strong treatments was offered as one of the reasons effect sizes would be so small in a beginning behavioral science area. The typical effect sizes observed by Black and Wiliam (1998) in their recent synthesis of formative assessment interventions were between 0.4 and 0.7, which might be characterized as in Cohen’s medium range.

This study focused only on understandings about rubrics, which may be thought as only an element of instructional adaptations of assessments. It did not manipulate knowledge of the objectives, for example. Indeed, all teachers received the same presentation on the goals of instruction (the CLGs). Furthermore, the nature of these rubrics as generic, although common in some fields such as writing, is unusual in other disciplines. The developers of the treatments intended in this study to enhance knowledge and uses of the rubrics by teachers had little research to guide them. Given the lack of direction from research on the statement and use of rubrics and the low reliabilities in this study that resulted from short criterion assessments, small effect sizes as low as 0.2 should not be surprising. Yet, the observed magnitudes of those effects that were statistically significant were in the 0.3 to 0.5 range. These considerations suggest that the results are consistent with the findings of Black and Wiliam (1998).

It is possible that improved rubrics or stronger treatments could be designed so that rubric training might be found effective in all content areas. Indeed, the rubrics used in this study were drafts that have since been revised. It is also possible that the positive effects of rubric training are already common within the repertoire of
teachers in certain disciplines. However, the results of this study are encouraging the use of rubrics as one way to promote more effective instruction by teachers. To do that, enhanced understandings are needed about the ways teachers may use rubrics in classrooms to prepare for instruction (Mehrens et al., 1998) and to communicate achievement expectations (Wiggins, 1998) as well as ways they are perceived and used by students. It may then be possible to design more effective ways to enable teachers to become skilled in their instructional uses of rubrics. For example, it may be that the advantages of understandings about rubrics become stronger over longer periods of time, which would provide opportunities for teachers to learn how to use rubrics and to internalize those uses. It would also be worthwhile to investigate whether analytic as opposed to holistic rubrics would be of greater instructional value, especially in content areas such as English and government where little support was found in this study.

Although this study has demonstrated, in at least some content areas, positive effects of focused teacher understandings about rubrics, whether those effects remain at the same levels across different types of rubric, such as objective specific as opposed to generic, remains for additional study. It will also be useful to study whether rubric training effects can be found to generalize to other research contexts. It is particularly important to study generalizability of effects across assessment formats, as was found here in algebra, because that would suggest that the advantages of rubrics in the learning process are demonstrable irrespective of the way in which student achievement is documented. Further research should also be undertaken to explore the mechanisms by which instructional change due to rubrics is effected as well as the nature of the resulting student behaviors that lead to improvements in achievement.

ACKNOWLEDGMENTS

This work was supported, in part, by Grant R305F60143–96 from the U.S. Department of Education’s Office of Educational Research and Improvement. We thank the Maryland State Department of Education content specialists whose efforts were essential to the success of this project: Elaine Crawford and Cindy Hannon (algebra), Gary Hedges (biology), Mary Jo Comer (English), Diane Johnson (government), and Susan Oskin (skills for success).

REFERENCES


**APPENDIX**

**Generic Rubrics**

**Algebra Rubric**

4. A response at this level analyzes the full range of the problem correctly. It represents all of the information appropriately, and applies mathematical concepts, which are essentially complete and correct, to solve the problem. This response thoroughly explains the process(es) used to solve the problem and justifies clearly the conclusion(s) in the context of the problem. This response must have the correct answer.

3. A response at this level analyzes the problem correctly. It represents most of the information appropriately, and applies mathematical concepts to solve the problem correctly. It may contain minor flaws. This response explains the process(es) used and shows some justification of the conclusion in the context of the problem.
2. A response at this level analyzes most of the problem correctly. It represents some of the information appropriately, and applies mathematical concepts to solve the problem. It may contain major flaws, or be incomplete. This response shows little or no attempt to explain the process used or to justify the conclusion.

1. A response at this level shows some attempt to solve the problem but analyzes the problem incorrectly. It represents little or no information appropriately and makes some attempt to apply mathematical concepts to solve the problem. This response makes little or no attempt to explain the process used or to justify the conclusion. It may have the correct answer with no supporting information or have inappropriate mathematical concepts.

0. A response at this level shows no evidence of mathematical thinking or no response is given.

*Justify conclusion* means the student will use mathematical principles (definitions, postulates, theorems) to support the reasoning used to solve the problem.

*Explain* the processes used means the student will use the language of mathematics to communicate how the student arrived at the answer.

A *major flaw* is an error that affects solving the problem.

A *minor flaw* is an error that does not affect solving the problem.

**Biology Rubric**

4. There is evidence in this response that the student, using analysis, has developed a full and complete understanding of the question or problem. The student has synthesized information to provide a correct answer and the supporting evidence demonstrates an integration of ideas. Information provided in the response demonstrates that the student has extended scientific concepts beyond the question or problem. The response is enhanced through the use of accurate terminology to explain scientific principles.

3. There is evidence in this response that the student, using analysis, has developed a good understanding of the question or problem. The student has synthesized information to provide a correct answer and the supporting evidence is complete. The response uses mostly accurate terminology to explain scientific principles.

2. There is evidence in this response that the student has a basic understanding of the question or problem. The student provides a correct answer, but the supporting evidence is only moderately effective. The response uses limited accurate terminology to explain scientific principles.

1. There is evidence in this response that the student has some understanding of the question or problem. The student may provide a correct answer, but
the supporting evidence is only minimally effective. The response makes little use of accurate terminology to explain scientific principles.

0. There is evidence that the student has no understanding of the question or problem. The student provides an incorrect answer or does not answer the question. The response does not make use of scientific terminology.

English Rubric

4. The response demonstrates a thoughtful creation, examination, and extension of meaning expressed in a distinctive voice and a deliberate tone. Through development, it fulfills a purpose and provides relevant support and complete elaboration. It organization skillfully follows an established structure that enhances the purpose throughout. Control of language demonstrates conscious selection of words and careful attention to audience understanding and interest and to conventions.

3. The response demonstrates the creation, examination, and extension of meaning and maintains a consistent voice and tone. Through development, it establishes a purpose and uses support and elaboration. Its organization adequately follows a structure and supports the purpose with minor inconsistencies. Choice of language demonstrates attention to audience understanding and interest and to conventions with few, if any, errors which interfere with meaning.

2. This response shows some creation, examination, and extension of meaning. Although development may be incomplete, the response addresses a purpose and uses support and elaboration. Its organization may lack consistency and therefore may not fully support the purpose. Use of language suggests some attention to audience understanding and interest and to conventions. The response may contain errors which interfere with meaning.

1. The response shows an attempt to create, examine, and extend meaning. Its development may include either a purpose that is not evident or may be misunderstood or contains support or elaboration that is inadequate. Its organization may be unstructured and confusing. Use of language may show minimal attention to audience understanding and interest and to conventions. The response may contain errors which interfere with meaning.

0. Other.

Government Rubric

4. This response shows understanding of the historical development and/or current status of principles, institutions, or processes of political systems. The response is insightful, integrates knowledge, and demonstrates powerful application.
3. This response shows some understanding of the historical development
and/or current status of principles, institutions, or processes of political sys-
tems. The response is thorough with accurate documentation and appropri-
ate application.

2. This response shows knowledge of the historical development and/or cur-
rent status of principles, institutions, or processes of political systems. The
response is acceptable with some support or key ideas. The response makes
some use of higher order thinking skills.

1. This response shows minimal knowledge of the historical development
and/or current status of principles, institutions, or processes of political sys-
tems. The response is related to the question but is inadequate, with signifi-
cant misconceptions and/or absence of key ideas.

0. This response is unrelated to the question or does not provide an answer to
the question.
Copyright of Applied Measurement in Education is the property of Lawrence Erlbaum Associates and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.