DOCUMENT RESUME

ED 431 030 TM 029 863

AUTHOR	Andrade, Heidi Goodrich
TITLE	Student Self-Assessment: At the Intersection of Metacognition and Authentic Assessment.
PUB DATE	1999-04-20
NOTE	62p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Classification; *Criteria; Grade 7; *Junior High School Students; Junior High Schools; *Metacognition; Protocol Analysis; *Self Evaluation (Individuals); *Sex Differences; Tables (Data)
IDENTIFIERS	*Authentic Assessment; *Scoring Rubrics

ABSTRACT

The hypothesis that task-specific criterion-referenced self-assessment can have a positive effect on students' metacognitive engagement and learning was tested. Seventh graders (n=47) were asked to invent, apply, and explain a classification system for a group of animals. Treatment subjects periodically assessed their performance in terms of a written rubric that listed the criteria for each task and gradations of quality for each criterion. Students in the control group were not asked to assess their work. Think-aloud protocols were collected and coded to provide insight into spontaneous self-assessment, the classification of self-assessment, and the influence of self-assessment on metacognitive engagement and learning. Approximately three-quarters of the students assessed themselves spontaneously. Girls in the treatment group were more metacognitive than boys, but no statistically significant differences were found for boys in treatment and control groups. Treatment students tended to outperform the control group on posttests. The rubric appeared to have a positive effect on the criteria that students used in their spontaneous self-assessments, and students who assessed their own work were remarkably willing to revise it. An appendix contains the scoring rubric given to the experimental group. (Contains 18 tables and 109 references.) (Author/SLD)

Student Self-Assessment: At the Intersection of Metacognition and Authentic Assessment

Heidi Goodrich Andrade
Harvard Graduate School of Education
Longfellow Hall—Project Zero
Cambridge, MA 02138

A paper presented at the annual meeting of the American Educational Research Association
April 20, 1999
Montreal, Canada

Abstract

Research questions. This study tests the hypothesis that task-specific, criterion-referenced self-assessment can have a positive effect on students' metacognitive engagement and learning. The study focused on four research questions:

1. Do students spontaneously self-assess when engaged in a classification task?
2. What kinds of self-assessment are students capable of under supportive conditions?
3. Does self-assessment influence metacognitive engagement in the classification task?
4. Does self-assessment influence learning about classification and arthropods?

Research design. Forty seventh-grade students were asked to invent, apply and explain a classification system for a group of animals. The treatment subjects periodically assessed their performance in terms of a written rubric that listed the criteria for each task and gradations of quality for each criterion. Students in the control group were not asked to assess their work. Think aloud protocols were collected and coded in order to answer the first three questions. Pre- and post-tests were used to determine content knowledge differences and to answer the fourth question.

Results and analysis. Approximately three-quarters of the students in the study assessed themselves spontaneously. Girls in the treatment group were more metacognitive than were girls in the control group, but no statistically significant differences were found between treatment and control boys in terms of metacognitive engagement. Statistically significant differences in pre- to post-test gains were found between both male and female students, with treatment students tending to outperform control students.
Other key findings include the positive effect of the rubric on the criteria that treatment students used in their spontaneous self-assessments, and the fact that students who assessed their own work were remarkably willing to revise it.

Objectives

Although self-evaluation, self-control and self-assessment are frequently mentioned in educational publications and in instructional and evaluative materials, there is little research on the effectiveness of self-assessment, including and especially its impact on learning and cognitive development. The Student Self-Assessment study was designed to test the hypothesis that guided self-assessment can increase metacognitive processing and learning about science. The study focused on four research questions:

1. Do students spontaneously self-assess when engaged in a classification task?
2. What kinds of self-assessment are students capable of under supportive conditions?
3. Does self-assessment influence metacognitive processing during the classification task?
4. Does self-assessment influence learning about classification and arthropods?

The basic premise throughout is that self-assessment functions in learning by increasing cognitive and metacognitive engagement and improving performance or achievement as a result. Support for this premise comes from a variety of areas of inquiry, including research on metacognition, authentic assessment, and self-regulated learning and feedback.

Literature Review

This study draws on three areas of cognitive and educational research: Metacognition, authentic assessment, and self-regulated learning and feedback. In this section, I examine the role of self-assessment in each area and draw on all three perspectives to support the hypothesis that self-assessment can serve learning by increasing cognitive and metacognitive engagement and thereby improving performance. The examination of each area focuses on four questions: 1) How is the area of inquiry defined? 2) What is the role of self-assessment in this area? 3) What form does self-assessment take in this area? and, 4) What implications does the research have for student self-assessment? I conclude the review by illustrating the common ground shared by each area of inquiry.

Self-Assessment in Metacognition

What is metacognition? The term metacognition refers to "knowledge or cognition that takes as its object or regulates any aspect of any cognitive endeavor" (Flavell, 1981, p. 37). The key components in Flavell's well-known taxonomy are metacognitive *knowledge* and metacognitive *experiences* (Flavell, 1977). Metacognitive knowledge refers to knowledge and beliefs about the workings of one's own and others' minds. It can be categorized as knowledge of person, task and strategy variables. For example, knowing that you need external memory aids to remember a list longer than six items or that another person has an unusual ability to manipulate numbers in her head is person knowledge.

Metacognitive *experiences* are cognitive or affective experiences that pertain to a cognitive enterprise. They take as their "ideational content where you are in a cognitive enterprise and what sort of progress you have made, are making, or are likely to make" (p. 107). For example, a sense of puzzlement over a paragraph, or a feeling of a gap in one's understanding of a concept are two kinds of metacognitive experiences. In older children and adults, these experiences trigger corrective moves, such as rereading the paragraph or reviewing the explanation of the concept.

Flavell (1981, 1987), Brown (1980, 1987), Scardamalia and Bereiter (1984, 1985) and others have shown that effective thinking in a variety of domains involves metacognition. Not only is metacognition an important ability for the mature thinker, but even young learners are able to reflect on and assess their own thinking in ways that significantly enhance their subject matter learning (Markman, 1981a; Palincsar & Brown, 1984, 1986, 1988; Schoenfeld, 1987). For example, research by Flower & Hayes (1981) has shown that the ability to monitor, evaluate and revise text while writing is an important part of an experienced writer's repertoire, and related to the powers of metacomprehension that children develop as they learn to write. Thus, learning to write well and developing the metacognitive skill necessary to evaluate one's own thinking go hand-in-hand.

The same conclusion has been drawn by researchers in other academic subject matters, who conclude that a key difference between high- and low-achieving students is the degree to which they monitor and evaluate their own thinking (Biemiller & Meichenbaum, 1992; Mancini, Mulcahy, Short & Cho, 1991; Nickerson, Perkins & Smith, 1985). As a result, metacognition has played a central role in many successful remediation and intervention efforts (Daiute & Kruidenier, 1985; Palincsar & Brown, 1988; Scardamalia & Bereiter, 1985; Yussen, 1983).

What is the role of self-assessment in current conceptions of metacognition? Flavell's model of metacognition places a strong emphasis on cognitive monitoring. Cognitive monitoring, as the term suggests, involves a good deal of self-assessment, in that it refers to the critical examination of one's thinking. In fact, Flavell uses the term "cognitive monitoring" interchangeably with the word "metacognition," suggesting that self-assessment plays a key role in his conception of metacognition. A simple comparison of the meanings of the words *assess* and *monitor* will illustrate this point:

> **assess**: 1: to sit beside, assist in the office of a judge.... 4: to determine the importance, size or value of.
> **monitor**: 1: to check... for quality or fidelity.... 3: to watch, observe, or check especially for a special purpose (Webster's New Collegiate Dictionary, 1980).

When used to refer to metacognitive behaviors, both words indicate making critical judgments about one's own thinking, or assessing oneself.

Ann Brown and her colleagues have proposed a taxonomy similar to Flavell's that also places a heavy emphasis on self-assessment. This taxonomy parses metacognition into *knowledge about cognition* and the *control or regulation of cognition* (Armbruster, Echols & Brown, 1982; Brown, 1978). The former component, *knowledge about cognition*, includes knowledge of four variables: Text, task, strategies, and learner characteristics. The latter component, *regulation of cognition*, refers to the coordination of those four variables when thinking. Here again, metacognition involves self-assessment through coordinating, evaluating and modifying one's approach to a task.

A further examination of the literature reveals that the notion of self-assessment through cognitive monitoring and control is ubiquitous. The act of engaging in metacognitive self-assessment has been described in many ways, including thinking about and modifying one's own thinking (Pace, 1991), self-regulation or manipulating one's ideas and approaches to solving problems (Price, 1991), controlling the processes with which one regulates cognitive behavior (Mancini, Short, Mulcahy & Andrews, 1991), planning, directing, monitoring and evaluating one's behavior (Weinert, 1987), monitoring learning or thinking about thinking (Berliner, 1990), and any active learning process involving continuous adjustments and fine-tuning of action via self-regulation (Brown, 1987), to name just a few.

This plethora of definitions and descriptions has led more than one researcher to lament the "fuzziness" of the concept (Brown, 1987; Flavell, 1981; Wellman, 1983). Nonetheless, the above collection demonstrates that it is not difficult to make a case for self-assessment as a key component of metacognition, as each example refers at least implicitly to monitoring and evaluating one's thought processes.

What form does self-assessment take in current conceptions of metacognition? Although metacognition can and does appear in some students as a natural result of cognitive development (Nisbet & Shucksmith, 1986), teachers and researchers agree that it does not appear often enough. Ann Brown (1980) notes that "in general, children fail to consider their behavior against sensible criteria, they follow instructions blindly, and they are deficient in the self-questioning skills that would enable them to determine these inadequacies." Fortunately, there is evidence that metacognition can be taught. Two approaches are presented below.

Perhaps the best known investigations into the teachability of metacognition are those by Palincsar and Brown (Brown & Palincsar, 1982; Palincsar & Brown, 1984; 1986; 1988). Their research focused on instruction in strategic activity for poor readers. The instructional procedure, called reciprocal teaching, engages small groups of students in framing questions about a passage, summarizing the passage, clarifying, and predicting. Brown (1992) explains that these four activities were selected because they are excellent comprehension-monitoring devices. For example, if a student cannot summarize what he has read, it is a good indication that understanding is not proceeding smoothly and that remedial action is necessary. Brown and Palincsar summarize the findings of their investigations this way:

> (a) Students' ability to summarize, generate questions from text, clarify, and predict all improved markedly; (b) improvements on comprehension measures were large, reliable, and durable; (c) the benefits of instruction generalized to classroom settings; and (d) there was transfer to tasks that were similar to but distinct from the training tasks (1988, p. 55).

A second example of the form self-assessment takes in work on metacognition is from a thinking skills curriculum, *Thinking Connections* (Perkins, Goodrich, Tishman & Mirman Owen, 1993). Designed to infuse the teaching of thinking into the regular curriculum, *Thinking Connections* includes one unit which focuses on metacognition. This unit, called "Mental Management," helps students develop an increased awareness of and control over their thought processes by asking themselves specific questions before and after a task.

The pre-task step, "Get Ready," has students focus their thoughts, recall the last time they did a task similar to the one they are about to engage in and remind themselves of how best to approach it, and form mental images of the task or topic. The first post-task step, "Make Connections," explicitly fosters the transfer of both content knowledge and thinking skills by having students make connections to other areas of knowledge and their own experience. The second post-task step, "Think about Thinking," has students review and assess their thinking during the task just completed by identifying what went well, which parts were difficult, and how they can improve on their thinking in the future. The strategy can be taught in a variety of ways, but the emphasis is on direct instruction and teacher modeling.

The purposes of these three steps are similar to those of Brown's work: To make students aware of their own thought processes as well as the fact that they can improve upon them through self-monitoring. Although extensive research on the effect of The Mental Management strategy on students' thinking has not yet been done, pilot testing did show that students tended to learn to think better when the strategy was used on a regular basis in the classroom.

Other researchers have experienced similar successes. Schoenfeld (1987) designed an instructional approach for an undergraduate mathematics and problem solving course that explicitly attended to the form and function of metacognitive self-monitoring and found clear evidence of marked shifts in his students' problem solving behaviors, particularly at the metacognitive level. Weinstein (1994) provides a course in strategic learning for students experiencing difficulty in college. The course places a heavy emphasis on metacognitive self-monitoring. She reports that the results are very significant: Students generally show improvements of ten percentile points or more on reading measures and on a measure of strategic learning, and they also evidence significant improvements in their grade point averages. These improvements are maintained across at least five semesters. Taken together, these studies show that self-assessment in the form of metacognitive self-monitoring and self-regulation can have a significant effect on thinking and learning.

What are the implications of this research for student self-assessment practices? Learning theory in general and research on the teaching of metacognition in particular provide insights into what is effective in instruction. The following discussion draws on this knowledge to propose a list of characteristics of instruction that promote the development of self-assessment skills.

- Awareness of the Value of Self-assessment. Brown (1978) and others (Flavell, 1981; Mancini, Mulcahy, Short & Cho, 1991; Price, 1991) point out that, unless students are aware of the value of assessing their own thinking through being metacognitive, such behaviors are unlikely to be maintained. Brown (1980) tests this claim when she differentiates between "blind" and "informed" training in a study. The former involves presenting metacognitive strategies without an explanation of the reasons for learning them. Informed training, on the other hand, ensures that students receive explicit information about the reasons for and the effectiveness of the behavior being acquired. Brown concludes that informed training, coupled with self-control and self-monitoring, is extremely important for the maintenance and generalization of skills (see also Mancini, Short, Mulcahy & Andrews, 1991). Studies by Pressley et al. (1983) corroborate this conclusion.

Schoenfeld treats the issue of awareness directly in his approach by showing his students a videotaped example of a student on what he calls a "wild goose chase"—the act of not monitoring or assessing one's own thinking and not making progress on a problem as a result. Theoretical support for awareness-raising practices such as this comes from Flavell's (1977, 1981) taxonomy of metamemory skills. His taxonomy includes one "type" of metamemory known as sensitivity, or a sense for when a situation calls for voluntary, intentional efforts to remember. Flavell notes that there is reason to believe that this sensitivity is learned, suggesting that attention to this issue should be explicit in instruction in self-assessment.

• Cueing. Teachers can play a pivotal role in fostering awareness of the when, how and why of self-assessment by alerting students to occasions when thinking metacognitively is appropriate and beneficial (Reading/Language in Secondary Schools Subcommittee of IRA, 1990). Scardamalia & Bereiter (1985) have found that children often fail to use self-regulatory strategies even when they have the necessary skills and understand that they would be beneficial, because of the processing demands of learning and using a new strategy. They developed an instructional technique known as procedural facilitation, which provides "cues or routines for switching into and out of new regulatory mechanisms... and minimize[s] the resource demands of the newly added self-regulatory mechanisms" (p. 567). They found evidence that the children's writing performance was positively affected by cueing in that they made more and better revisions than usual.

• Modeling. Modeling is a well-known instructional technique in which learners learn by observing an expert engaging in a desired behavior. Researchers in metacognition have found that instruction benefits when metacognition is modeled by thinking aloud for students (Brown, 1988; Scardamalia & Bereiter, 1985; Scardamalia, Bereiter & Steinbach, 1984). Each of the examples outlined above include explicit modeling components.

• Mediation. It is one thing to ask students to assess their own thinking, another entirely to ensure that their assessments are accurate and productive. In order for students to become competent at self-monitoring, it is necessary for the teacher to act as a mediator, assisting students in the regulation and assessment of their thinking (Price, 1991). Brown mediates by asking her students the questions they should be asking themselves and gradually giving the responsibility for question-asking over to them.

• Social context. Brown's approach to the teaching of metacognition is a group problem-solving activity, and *Thinking Connections* encourages teachers to take such an approach if they are comfortable with it. My review of the literature reveals that this is not uncommon in theory or in practice. Costa (1991), Brown (1987; 1988), Mancini, Short, Mulcahy and Andrews (1991) and others cite the considerable support to be found in the social context of collaboration among learners.

Palincsar and Brown (1988) have observed that peers are frequently in a better position to assist one another in monitoring and adjusting their comprehension of a text, presumably because they are "more likely to be experiencing the same kind of difficulty in comprehending the text than teachers, for whom comprehension occurs with relative automaticity" (p. 57). Schoenfeld finds group work valuable for several reasons. First, discussions can be analyzed for their efficiency, providing an opportunity to reflect on self-regulation and how it works. Second, sharing the burden of problem-solving means that no individual student is responsible for generating all the ideas or keeping track of all the options, freeing them to focus on decisions about the best approach to take. Third, students are "remarkably insecure.... [and] working on problems in groups is reassuring: one sees that his fellow students are also having difficulty, and that they too have to struggle to make sense of the problems that have been thrown at them" (1983, pp. 30-31, cited in Schoenfeld, 1987). Finally, Schoenfeld stresses the importance of "creating a microcosm of mathematical culture," in which "students experienced mathematics in a way that made sense, in a way similar to the way mathematicians live it" (p. 213).

The value of the social context in learning also draws broad theoretical support from the work of Mead (1934), Vygotsky (1978) and Bruner (1978). Mead writes that the development of the reflective self is impossible outside of social experience, and, according to Vygotsky, all higher order cognitive functions originate in individuals' interactions with others. Research on collaborative learning tends to support these claims that cognitive development involves the internalization of social interactions (Daiute and Dalton, 1988).

• Direct instruction. It is usually necessary to begin most instruction in self-monitoring with some direct instruction (Scardamalia, Bereiter & Steinbach, 1984), although the goal over time is to have the teacher act as intellectual coach or moderator, permitting students to manage their own thinking and learning. *Thinking Connections* stresses the role of direct instruction by suggesting that teachers explicitly teach the three questions of the Mental Management strategy.

• Transfer. The maintenance and generalizability of skills has been a major issue in the teaching of thinking (French & French, 1991; Mancini, Short, Mulcahy & Andrews, 1991; Price, 1991). Brown (1980) and Perkins (1987) have shown that, unless training encompasses planned steps to ensure the generalization of the skills being learned, it is unlikely that the actual generalization of skills will occur. *Thinking Connections* addresses the transfer problem by including a transfer step, "Make Connections," in the Mental Management strategy.

• Remedial or corrective tactics. One criticism of metacognitive strategies has to do with the fact that they do not offer students any guidance about what to do when they find their thinking is not meeting their goals. Too often, students have no idea how to correct problems in their thinking. Flavell (1981) notes that students need to develop cognitive actions or strategies for making progress as well as for monitoring progress. Instruction in metacognitive strategy use should therefore be combined with instruction in the cognitive techniques and strategies of the subject matter. For example, Schoenfeld teaches his students the heuristic strategies of the mathematician while at the same time helping them to monitor their use of those strategies.

• Duration. Derry and Murphy (1986) and Sternberg (1986) claim that a thinking skills program of less than a semester's duration does not appear to warrant serious consideration, and, in fact, a thoughtful, systematic curriculum which extends over the course of two or three years may be necessary for an effect to be significant.

In order to be effective, any approach to instruction in self-monitoring should have the qualities listed above.

Self-Assessment in Authentic Assessment

What is authentic assessment? Gardner defines assessment as "the obtaining of information about the skills and potentials of individuals, with the dual goals of providing useful feedback to the individuals and useful data to the surrounding community" (1991, p. 90). Assessment becomes authentic when it exemplifies the real-life behaviors and challenges experienced by actual practitioners in the field (Davidson et al., 1992; Hawkins et al., 1993; Wiggins, 1989b; Wolf & Pistone, 1991). On this formulation, standardized tests generally do not qualify as authentic forms of assessment (what practicing scientist, for example, ever takes one?), while portfolios, such as those used by artists, do.

According to Wiggins (1990), assessments must have certain characteristics in order to be considered authentic. An assessment must be:

- composed of tasks which we value, and at which we want students to excel—tasks worth "teaching to" and practicing. Tasks simulate, mimic, or parallel the kinds of challenges facing the worker in the field of study.
- constructed of "ill-structured" or "open-ended" challenges that require a repertoire of knowledge, as opposed to mere recall, recognition, or the "plugging in" of a ready-made algorithm or idea.
- appropriately multi-staged, leading to revised and refined products and performances.
- focused on students' abilities to produce a *quality* product or performance. Important processes and "habits of mind" are thus necessary means to the final work, and may be assessed.
- sufficiently de-mystified and known in advance to allow for thorough preparation and the possibility of self-assessment.
- adaptable to student styles and interests, whenever possible and appropriate.
- based on judgments in reference to clear, appropriate-to-the-task criteria.
- rarely limited to one-shot, one-score tests with no interaction between assessor and assessee. Often the assessment focuses on the student's response to questions or ability to justify answers and choices made.

Thus, authentic assessment not only reflects the kinds of assessment techniques employed by practitioners in the field; it must also promote learning and growth for all students.

What is the role of self-assessment in current conceptions of authentic assessment? The purpose of student self-assessment in authentic assessment mirrors the purposes of self-assessment in metacognition: To help students become critical judges of the quality of their own work and their approaches to it. Baron, for example, characterizes "enriched performance assessment tasks" as those which, among other things, "spur students to monitor themselves and to think about their progress" (1990, p. 134). Haney acknowledges the importance of designing assessments that encourage students to become "autonomous and self-regulating adults, capable of judging the success of their own endeavors" (1991, p. 154). Perrone makes a similar point when he notes that, given repeated opportunities to actively participate in the evaluation of their own work, students "have become increasingly more articulate about their progress and what they need to work on to improve their performance and enlarge their understandings" (1991, p. 166). In his discussion of student-centered assessment, Stiggins (1994) claims that "our comprehensive reexamination of achievement targets over the past decade has revealed that student self-assessment is not just an engaging activity. Rather, it turns out to be the very heart of academic competence" (p. 33).

In an extended discussion of the role of self-assessment in the arts, Wolf and Pistone (1991) note that: "No artist survives without being what the artist Ben Shahn calls 'the spontaneous imaginer and the inexorable critic.' An episode of assessment should be an occasion when students learn to read and appraise their own work" (p. 8). Teachers and students of the arts reported that the major reason for assessing student work is to teach them how to be rigorous critics of their own work.

Wolf, Bixby, Glenn and Gardner (1991) criticize the current testing system in this country for not allowing students to participate in discussions about the standards that are applied to their work, and argue that "assessment is not a matter for outside experts to design; rather, it is an episode in which students and teachers might learn, through reflection and debate, about the standards of good work and the rules of evidence" (p. 52). Wolf et al. include on their list of characteristics of useful assessments classroom practices in which teachers and students openly discuss the standards for good work and in which students reflect on the quality of their own work.

Zessoules and Gardner (1991) also highlight the role of self-assessment in authentic assessment when they list the development of "reflective habits of mind" as one of four conditions for establishing an assessment culture. As used by these authors, the word reflection refers to students' abilities to recognize and build upon their strengths as well as what challenges them in their work. They argue that reflection depends on students'

> capacity to step back from their work and consider it carefully, drawing new insights and ideas about themselves as young learners. This kind of mindfulness grows out of the capacity to judge and refine one's work and efforts before, during and after one has attempted to accomplish them: precisely the goal of reflection (p. 55).

The conception of assessment put forward by these authors challenges students to develop their capacities for self-critical judgment by carefully evaluating their own work.

What forms does self-assessment take in current conceptions of authentic assessment? The first example of the form self-assessment takes is from Arts PROPEL, a collaborative project of researchers from Harvard Project Zero, the Educational Testing Service, and the Pittsburgh Public Schools. Based on the assumption that learning in the arts occurs most fully when students reflect on as well as produce art, the PROPEL approach has students take responsibility for critiquing, refining, revising and rethinking their own work (Davidson et al., 1992; Gardner, 1991; Herman & Winters, 1994).

Students in the Ensemble Rehearsal Critique Project, for example, perform a piece of music, then write comments and suggestions for revision or practice plans on a two-part evaluation sheet (see Table 1). The first section of the sheet refers to the students' own performances. The second section, which is filled out after listening to a tape of the performance, refers to the performance of the entire ensemble. The evaluation sheets were designed this way in order to scaffold assessment of the ensemble from at least two critical perspectives—one's own and the director's. After writing their assessments, students discuss their critiques with their teacher and the rest of the class.

Table 1
Excerpt from Ensemble Rehearsal Critique Worksheet from Arts PROPEL

ENSEMBLE REHEARSAL CRITIQUE

Critique

Write down your critique of the ensemble performance specifying LOCATION (where you performed particularly well or need to improve) and MUSICAL DIMENSIONS (such as rhythm, intonation, tone, balance, articulation, phrasing, interpretation, etc. or any dimension specified by the teacher). Using words such as "because" be sure to mention any links between your own or your section's performance and the ensemble as a whole.

Location Dimension My (Section's) Performance / Ensemble's Performance

Revision

Also include remarks concerning REVISIONS OR PRACTICING STRATEGIES for yourself or the ensemble. Be sure to include the main problem in terms of its dimension and location in the piece your or the ensemble should practice on before or during the next rehearsal.

These evaluation sheets, along with journals, questionnaires, peer interviews and any teacher notes about class discussions, are collected in portfolios. The portfolios, or "process-folios," are periodically reviewed by the students and the teacher in order to involve the students in constant reflection on their activities and to help them monitor and learn from their own growth and their own setbacks (Gardner, 1991). In this way, students are actually assessing their own self-assessment skills.

In order to assign grades, PROPEL teachers formally score the work in the portfolios. Reflection skills are scored in terms of the "identification of musical elements in critical judgments," the "ability to suggest revisions or practice strategies for improving performances," and the "critical perspective(s) assumed by students while discussing the individual and ensemble performance(s)" (Davidson et al., 1992, p. 31).

Davidson et al. report that, with optimal support, evidence of the development of critical self-assessment skills does appear. Students in the Ensemble Rehearsal Critique Project become increasingly able to formulate productive and meaningful reflections on performances, to map musical terminology appropriate with their perceptions and practice strategies, to take several critical perspectives at once, to question all aspects of the ensemble when encouraged to listen carefully, and to offer suggestions for themselves and the ensemble. Davidson and Scripp (1990) summarize the effects of self-assessment in this way:

> Given cause to reflect about their own performance and the ensemble, students become more self-directive. Rather than looking at section leading, arranging music or conducting a rehearsal as added workload, students begin to see these activities as being the goal of being in the ensemble over many years.... Reflective thinking serves as the entry point in this path toward the musicianship skills of the director (p. 60).

Similar claims are made about the practice of reflection through the PROPEL approach in other artistic and academic domains, including photography, playwriting, dance, the visual arts, and mathematics (Wolf & Pistone, 1991).

A second example of student self-assessment comes from the work of Paris and Ayers (1994). These authors claim, as I do, that self-assessment contributes to authentic, learner-centered assessment practices that promote learning. Working with K - 6 teachers and administrators in Michigan, these researchers developed a portfolio approach to literacy assessment that also relies heavily on student self-evaluation and self-assessment. The portfolios employ a variety of reflection tools, including the process of selection of materials for inclusion in the portfolios, global self-evaluations, inventories, surveys, journals, self-portraits, letters, and conferences. An example of a task-specific, criterion-referenced self-assessment tool used in this project can be found in Table 2. Tools like this one are used by Paris and Ayers to promote active engagement of students in their own learning through reflection and review on a daily basis.

Table 2
Excerpt from Self-Assessment Sheet Used in Paris and Ayers' Portfolio Project

Summary of Expository Text
Name ______________________________ Date ___________________

Components of a Good Summary	Student Assessment	Teacher Assessment
I included a clear main idea statement		
I included important ideas supporting the main idea		
My summary shows that I understand the relationships between important concepts		
I used my own words rather than words copied from the text		

Paris and Ayers do not report any research-based results of their work, but they do provide some insights into the characteristics of self-assessment practices. These and others' insights are summarized in the following section.

What implications does research in authentic assessment have for self-assessment? The discussion of the role of self-assessment in authentic assessment emphasizes the need for self-assessment instruments to be criterion-referenced, task-specific, repeated and ongoing, and employed while there is still time to modify one's work. Several other characteristics that support authentic assessment in general and self-assessment in particular can be found in the literature, including:

• In Context. Most researchers agree that, to be considered authentic, assessment must be woven into the fabric of the curriculum, rather than separated out from the learning process (Herman, Aschbacher & Winters, 1992; Wiggins, 1989b). Although it may initially be necessary to structure formal self-assessment periods for students, such activities can gradually become part of the natural landscape of the classroom as students learn to self-assess automatically and regularly (Gardner, 1991).

• Clear Criteria. Most researchers also agree that assessment practices become more valid and effective when students are explicitly informed of the criteria or standards by which their work will be assessed (Herman, Aschbacher & Winters, 1992; Mabe and West, 1982; Paris and Ayers, 1994; Wiggins, 1989a, 1989b; Wolf, Bixby, Glenn & Gardner, 1991). Many researchers and teachers suggest that students themselves be involved in determining the criteria (Davidson et al., 1992; Higgins, Harris & Kuehn, 1994; Schmar, 1995; Towler & Broadfoot, 1992; Satterly, 1989). Regardless of how they are determined, however, the criteria, standards and rules of evidence must be rigorous (Wolf, Bixby, Glenn & Gardner, 1991) and must reflect curricular goals (Davidson et al., 1992).

Clear criteria not only improve the validity of self-assessments, they also guide students in monitoring their own thinking and learning. Butler and Winne (1995) argue that one reason students have difficulty monitoring their work is because they do not have standards or criteria against which to judge their progress. Butler and Winne cite two approaches to supplying missing information about criteria against which to measure achievement. In one, students are taught internal criteria against which to judge their performance (e.g., Baker, 1984; Bereiter & Bird, 1985). In the second, students are induced to judge their comprehension against external information, such as feedback supplied when they attempt to answer questions (e.g., Walczyk & Hall, 1989). Both approaches have proven helpful when students address near-transfer tasks, presumably because each provides criteria that can be used to judge performance more accurately in relation to goals.

Finally, Steinberg's (1989) review of research on learner control showed that feedback that provides information about current comprehension levels and/or prescriptive advice about how to further one's learning increased persistence at learning tasks and improved performance. Criterion-based self-assessment provides such information and advice by informing students about the need to monitor their learning and by guiding them in how to improve their work.

• Task-specific. My review of the meager literature on self-assessment suggests that the literature actually refers to two different phenomena which, for the sake of clarity, I will refer to as "self-evaluation" and "self-assessment." One difference between self-evaluation and self-assessment is that the former tends to refer to global qualities of oneself as a learner, while the latter refers to one's performance on a specific task.

Self-*evaluation* can be thought of as the process of developing a broad profile of oneself as a learner (Waluconis, 1993) by examining one's own learning styles, theories of learning, personal growth, and other indicators of how one learns and the ways in which one has grown intellectually. Kisnic and Finley (1993) see the purposes of self-evaluation as "helping students make meaning, derive relevance and build coherence through their educational experience" (p. 13). This goal is often accomplished by having students write self-evaluations at different times in their academic careers, including when they begin a school year, midway through a learning experience, and/or as a summative effort at the end. For example, students may be asked to write answers to prompts such as "Looking back, I realize that I ought to change my study habits/learning style/priorities in the following way," or "I judge my weak points to be the following" (Oskarsson, 1984). Similarly, a worksheet from an assessment guidebook for teachers requires that students write about how they feel about solving math problems (D. C. Heath, 1994, p. 17).

In an extensive if somewhat dated review of research on self-evaluation in adult second language learning, Oskarsson concluded that "it is quite possible to move from self-[evaluation] in general terms, which is what most researchers in the field have been concerned with so far, to self-assessment at a more specific and detailed level" (p. 26). Self-*assessment* at a more specific and detailed level is the approach taken in this study. Rather than having students reflect globally on what has been learned or achieved (Towler & Broadfoot, 1992), I asked them to think about the quality of the processes and products of their work on a specific task, much as a teacher would do to provide feedback on a work in progress.

I have taken this approach to self-assessment because research has shown that, in comparison to global self-evaluations, task-specific self-assessment is generally more valid, promotes self-monitoring, increases persistence and improves performance. The validity of both self-evaluations and self-assessments is the most thoroughly researched of these findings. For example, in a review of 55 studies in which self-evaluations of ability were compared with measures of performance, Mabe and West (1982) conclude that the best strategies for improving the validity of self-evaluation are to have objective measures of performance and to inform the subjects that their own evaluations will be compared to those objective measures. Oskarsson reports that self-assessments that refer to "specified linguistic situations" such as one's ability to introduce a friend or ask for a telephone number more highly correlate with test results than self-evaluations that refer to general abilities such as understanding or speaking English. Thus, more valid self-assessments can be expected in reference to specific tasks.

• Opportunities for Improvement. Stamps (1989) found that self-assessment was effective and motivating only when students were able to revise their work based on their assessment of it. My own experience echoes this finding: Students quite correctly feel self-assessment is pointless unless revision is possible, and, as a result, either abandon it entirely or give only cursory attention to it when this condition is not met.

• Multidimensionality. The criteria for assessment should cover all aspects required for good performance in the task at hand, including process as well as product aspects (Hawkins et al., 1993; Towler & Broadfoot, 1992). For example, whether a student or a teacher, the judge of student work could look for evidence of inventiveness or transfer, collaboration or the intelligent use of resources, thinking skills or dispositions, and so on, depending on the requirements of the task (Wolf, Bixby, Glenn & Gardner, 1991). At the same time, however, assessment practices should avoid unmanageable complexity (Perkins, Jay & Tishman, 1993b). Self-assessment practices, therefore, must embody an elegant balance between thoroughness and simplicity.

• Sensitivity to developmental stages. The call for multidimensionality raises the question of what students are developmentally capable of in terms of self-assessment. Clearly, students become more sophisticated in their judgments as their knowledge of and control over the workings of their own minds increase (Davidson et al., 1992; Satterly, 1989). Yet, some researchers (myself included) have been struck by the sophistication with which children as young as eight can reflect on their own work (Walters, Seidel & Gardner, 1994). Any approach to self-assessment should pay special attention to the students' developmental preparedness, and neither under- nor over-estimate their abilities to judge their own work.

• Sensitivity to individual differences. Individual differences, as used in the literature on authentic assessment, means anything from intelligence profiles (Gardner, 1991) to learning styles (Hawkins et al., 1993) to motivation (Watkins, Cox, Mirman Owen & Burkhardt, 1992). These authors recommend that self-assessment practices be open-ended enough to encourage different approaches and involve significant student choice whenever possible (Wiggins, 1989a).

• Social context. The class discussions of musical performances used in the Arts PROPEL approach are one example of how even self-assessment can be a highly social experience. According to Wolf & Pistone (1991), the "dimensions of excellence," or standards by which one's work should be assessed, can grow out of public discussions such as these. Herman, Aschbacher and Winters (1992) write that public discussions may help students internalize the standards and rules they need to become effective self-assessors, and that groups facilitate learning by providing many models of effective thinking strategies and mutual constructive feedback. Although self-assessment is often done by oneself, there is no reason to think that it must be learned by oneself.

• Frequency. Research and common sense indicate that authentic assessment is longitudinal and comprised of regular and repeated observations (Gardner, 1991; Hawkins et al., 1993; Wolf, Bixby, Glenn & Gardner, 1991). In order to be effective, any approach to self-assessment should be practiced at regular intervals (Oskarsson, 1984) and result in a collection of self-assessments which can themselves be assessed.

• Assistance and Practice. Having found that reflection and self-assessment are foreign to most students, Davidson et al. (1992) write that teachers will need to use supportive formats such as worksheets, classroom discussions, questionnaires, interviews and journals to help students engage in these activities, at least at first. In addition, Satterly (1989) recommends that teachers and other experts be prepared to assist in making accurate and productive assessments because students are not always able to tell whether or not their work measures up to the standards set for it. Mabe and West and Oskarsson note that practice and experience lead to marked improvements in students' self-assessments.

• Modeling. Herman, Aschbacher and Winters (1992) write that examples of what constitutes good work are necessary aids for students in making judgments about their own work. There are at least two ways to provide such models. Hawkins et al. (1993) suggest making a library of exemplars, including critiques by master assessors, available to all students. A second way to provide models of self-assessment is to have teachers model reflection and self-assessment for their students (Davidson et al., 1992).

Self-Assessment: At the Intersection of Metacognition and Authentic Assessment

What is self-assessment? The preceding review of the literatures on both metacognition and authentic assessment make it clear that research in each area shares the common goal of teaching students to assess themselves by standing back and reflecting upon the products and processes of their work. More specifically, self-assessment is the act of monitoring and evaluating one's work and one's approach to it in terms of clearly defined criteria, for the purposes of determining whether or not one is meeting the stated goals for the task at hand.

Effective instruction in metacognition and authentic assessment also share several key characteristics. The following analysis is intended to inform instructional design both in this study and in classroom practice. Table 3 presents the key characteristics of effective instruction in metacognition and authentic assessment for comparison. It is apparent that, although the terminology is different, the meanings of many of the terms are quite similar. In fact, the first five characteristics found in both columns of Table 2.3 are almost identical. These characteristics also represent the standard core of current thinking on education in general, not just in terms of metacognition or assessment. In general, instruction is most likely to be effective when:

- Students are exposed to models and exemplars of the behaviors to be learned
- Students are scaffolded in their efforts to learn, and assisted in making accurate and constructive self-assessments
- Students support and learn from each other
- Students are in possession of the tactics and the time to improve their work, and
- Students are given ample opportunity to learn and practice the behavior.

Table 3
Key Characteristics of Effective Instruction in Metacognition and in Authentic Assessment

Metacognition	Authentic Assessment
Modeling	Modeling
Mediation	Assistance and practice
Social context	Social context
Remedial or corrective tactics	Opportunities for improvement
Duration	Frequency
Awareness	In context
Transfer	Sensitivity to individual differences
Direct instruction	Sensitivity to developmental stages
Cueing	Multidimensionality
	Clear criteria
	Task-specific

Table 3 lists several additional characteristics that are not common to both columns. The metacognition column, for instance, lists cueing as a key element of effective instruction but the authentic assessment column does not. In an earlier paper (Goodrich, 1993), I raised the question of whether or not these characteristics are idiosyncratic and appropriate for only one area, or are necessary elements of effective instruction in self-assessment. In the interest of brevity, I refer the reader to that document for a discussion of this question, and present only my conclusions here. I concluded that the characteristics of effective instruction in the self-assessment are:

1. Awareness of the value of self-assessment on the part of students
2. Clear criteria on which to base the assessments

3. A specific task or performance to be assessed
4. Models of effective self-assessment
5. Direct instruction and assistance in assessing oneself, as needed
6. The support of one's social context (peers and others)
7. A significant duration and frequency
8. Cueing as to when it is appropriate to assess one's work, as needed
9. Attention to transfer
10. The opportunity to employ remedial or corrective tactics
11. Self-assessment occurs within the context of students' work
12. Self-assessment is multidimensional and attends to both process and product
13. Sensitivity to developmental stages, and
14. Sensitivity to individual differences.

These characteristics influenced the design of the self-assessment instrument used in this study, and should guide self-assessment initiatives in the classroom.

What is the role of self-assessment in learning? I have argued that the purpose of self-assessment is to promote metacognitive engagement and thereby increase learning. Support for this claim can be found in recent research on the relationship between self-regulated learning and feedback.

Self-regulated learning and feedback. Briefly, the construct "self-regulated learning" includes and extends the construct of metacognition. Accordingly, Butler and Winne (1995) define self-regulated learning as a style of engaging with tasks in which students exercise a suite of powerful skills, including setting goals for upgrading knowledge; deliberating about strategies and selecting those that balance progress toward goals against unwanted costs; monitoring the accumulating effects of their engagement; adjusting or even abandoning initial goals; managing motivation; and occasionally even inventing tactics for making progress.

In their synthesis of research on the role of feedback in self-regulated learning, Butler and Winne provide numerous insights that support my claim that self-assessment can increase metacognition and learning. For one, they note that feedback can boost content learning: "Most studies acknowledge that cognitive processing is cued by feedback and adopt a theoretical view of feedback that suggests that if feedback cues active and elaborate processing of content (deep processing) then achievement will increase" (p. 266). For example, a study by Andre and Thieman (1988) shows that feedback that cued deeper processing of specific information enhanced learners' memory for that information on repeated questions.

Another insight provided by Butler and Winne's synthesis is that productive feedback increases self-regulated learning by providing information about guiding tactics and strategies that process domain-specific information. They cite several studies that show that in general, learning improves when feedback informs students about the need to monitor their learning and guides them in how to achieve learning objectives. For example, Bangert-Drowns et al. (1991) note that feedback is effective to the extent that it "empowers active learners with strategically useful information, thus supporting self-regulation" (p. 214).

A model of self-assessment. These researchers are converging on the same notion that I have proposed as the hypothesis for this study: Feedback functions in learning by fostering metacognitive, self-regulating behaviors that increase content learning. Butler and Winne's review makes a convincing argument that another component—deep processing—be included in this model. Deep processing is the active and elaborate processing of content (Butler & Winne, 1995). A number of experiments have appeared in the literature that can be interpreted as illustrating that more fully elaborated material results in better memory (see Anderson, 1980). Research cited above (Andre & Thieman, 1988) shows that feedback can cue deeper processing of information, so it is reasonable to expect that self-assessment can prompt not only metacognitive engagement in a task but also deeper processing of material.

The model of self-assessment now states that self-assessment embedded in an appropriately supportive instructional context (as defined earlier) increases learning by boosting metacognitive engagement and deep processing. In the remainder of this thesis I describe and discuss a study designed to test this model in general and the following four research questions in particular:

1. Do students spontaneously self-assess when engaged in a classification task? If so:
   a. To what degree do they self-assess?
   b. What criteria do they use?
   c. Are unsatisfactory self-assessments followed by revision or other corrective moves?
2. What kinds of self-assessment are students capable of on this task under supportive conditions?
   a. To what degree do they self-assess under supportive conditions?
   b. What criteria are used in addition to those provided by the researcher?
   c. Are unsatisfactory self-assessments followed by revision or other corrective moves?
3. Does self-assessment influence metacognitive engagement in the task?
4. Does self-assessment influence learning about classification and arthropods?

Question 2 refers to the self-assessments of the treatment subjects. Ideally, the "supportive conditions" provided for these students would reflect the entire list of characteristics of effective self-assessment instruction listed on page 32. The limitations of clinical research, however, have led me to define "supportive conditions" in this study as:

- Awareness of the value of self-assessment. I briefly discussed how professional athletes succeed by assessing their own performance before I asked students to begin working.
- Task-specific. Students were asked to assess their work on the arthropod classification task they were engaged in at the time of the request.
- Criterion-referenced. Students were provided with the criteria and standards for self-assessment in the comprehensible and accessible form of a rubric (see Appendix A).
- In context. Students were asked to assess themselves as they engaged in the task.

- Modeling. I briefly modeled self-assessment for each student.
- Cueing. I assisted regularly prompting students to assess themselves.
- The opportunity to employ remedial or corrective tactics. Students were asked if they wanted to try to improve their work after they assessed it.
- Attends to both process and product. The rubric referred to both the approach students took to the work as well as the quality of their final products.

Methods

Forty seventh-grade students from a public middle school in a relatively homogeneous rural/commuter community in northern Massachusetts volunteered to participate in this study. Two independent variables were measured: metacognitive processing and content knowledge. Data on students' metacognitive processing were collected by audio taping, transcribing and scoring students' think aloud protocols. Table 4 contains a summary of the coding categories. In future research the coding system will be simplified by defining code-able statements at a much coarser level of grain. For example, the six kinds of metacognitive statements—metaknowledge, metaknowledge minus, metatask, meta-experience, process and process plus—could be one metacognition category. In fact, that is the level of grain that was most useful in this study.

Table 4
Summary of Coding Categories for Think Aloud Protocols

Generating ideas: Naming a grouping or categorizing principle that can be used to group all or some of the animals.

Questions: "Lower order" questions, like "What do grasshoppers eat?" and "Are those legs?"

Metaknowledge: Statements about one's knowledge and ability.

Metaknowledge minus: A simple "I don't know," often at the end of a sentence or phrase.

Metatask: Questions and statements about the demands and nature of the task.

Meta-experience: Statements about the perceived level of difficulty of the work.

Process goal-setting: Planning ahead, organizing ideas, instructions students give to themselves about how to proceed, questions they ask themselves about what to do next.

Process plus: Planning ahead/process statements that give a reason or justification for the move.

Assessment: Students' evaluations of their ideas, reasoning and categories:

- **Positive/negative/neutral**
- **Criterion-referenced:** gives reasons for accepting or rejecting a move or approach.
- **Self-check/correction:** a simple affirmation or corrective move without an explicit assessment statement.

Data on students' content knowledge were collected via the 14-item multiple choice, short answer test shown in Table 5. This test was administered both before and after students completed the classification task. Changes in their scores were compared to measure growth in content knowledge.

Table 5
Pre- and Post-test of Content Knowledge

1. List as many arthropods as you can:
2. What does the word "arthropod" mean?
3. How many *known* species of arthropods are there?
4. How many species of arthropods do scientists *estimate* there are?
5. Where do arthropods live?
6. What are the characteristics that all arthropods share?
7. What are the characteristics that differ between different arthropods?
8. In what ways is a grasshopper different from a spider? Be specific.
9. Which of the following are not arthropods?
   a. tick b. crayfish c. squid d. earthworm
10. What is the best way to classify arthropods?
11. How would you classify exercises, such as jumping jacks and chin-ups?
12. How many legs do lobsters have?
13. The only arthropods that have gills are the:
    a. centipedes b. insects c. spiders d. crustaceans
14. Tell me everything you know about arthropods that you have not yet revealed on this test (at least 2 things).

Information on three dependent variables—gender, special education classification, and scores on the California Achievement Test taken during students' sixth-grade year—was also collected and included in the analyses.

Procedures

Students were asked to think aloud as they invented, applied and explained a classification system for a group of eighteen arthropods (insects, spiders, lobsters, etc.). Students in the treatment group were asked to assess their work according to a written rubric (see Appendix A) three times—after they 1) read a page of text about arthropods, 2) created a classification system and sorted the arthropods, and 3) explained their system. Regardless of the rating they assigned themselves, they were asked if they wanted to do anything to improve their work. If they did, they were given time to revise. Students who chose to revise were asked to re-rate themselves and again given time to revise if they chose. Students in the control group were asked to think aloud while engaging in the task but were not stopped for self-assessment. Students in the control group did not see the rubric and were not asked to assess their own work. All forty students were given the pretest and post-test of content knowledge shown in Table 5.

Analysis

The analysis of the data had three main parts. The first two parts concern the degree to which students assess themselves under prompted and unprompted conditions. In order to speak to this question, descriptive statistics were calculated and the data was examined for emergent patterns or trends in the kinds of criteria students applied to their work, and in the ways in which treatment students responded to the opportunity to improve upon their work. A chi-square statistic was calculated to test for differences between the treatment and control groups in terms of criteria usage. The third research question concerns the effect of prompted self-assessment on metacognitive engagement in the classification task. Multiple regression was used to analyze the effects of experimental condition, gender, and CAT score on the number of metacognitive statements per line of text in students' think aloud protocols. Multiple regression was also used to analyze the data relevant to the fourth and last research question, which concerns the effects of experimental condition, gender, and score on the California Achievement Test on content knowledge gains.

Results

This section is organized according to the four research questions:

1. Do students spontaneously self-assess when engaged in a classification task?
2. What kinds of self-assessment are students capable of under supportive conditions?
3. Does self-assessment influence metacognitive engagement in the classification task?
4. Does self-assessment influence learning about classification and arthropods?

Question 1: Do students assess themselves spontaneously?

This question has two sub-questions, a) if students do assess themselves spontaneously, to what degree do they assess themselves? and b) what criteria do they use?

Degree of self-assessment. An examination of the number of statements coded as assessments (NASMNT) for the control subjects reveals that many students do indeed assess themselves spontaneously as they create a classification system and as they explain their systems. Fifteen of the twenty students in the control group made at least one statement coded as either a positive, negative or neutral evaluation, and three additional students made one (and only one) statement coded as a self-correction. The number of assessment statements ranged from zero to one hundred and eight. One hundred and eight was an extremely unusual data value, with the next lowest value at twenty-eight. This number reflects the length of time the student spent on the task though, not an unusually high rate of self-assessment. In order to represent the rate of self-assessment while at the same time preserving the variance in the sample, a new variable, percent assessment (PCTASMT) was calculated by dividing the number of assessment statements by the total number of lines of text in a student's think aloud protocol. Table 6 contains the descriptive statistics for both assessment variables.

Table 6
Descriptive Statistics for NASMT and PCTASMT, with and without Unusual Cases

	Mean	Standard Deviation	Median	Range
NASMT (N = 20)	13	23.45	7.5	0 - 108
PCTASMT (N = 20)	0.159	0.11	0.141	0 - 0.37

Criteria used in self-assessment. Fifteen control students made seventy-one criterion-referenced evaluative statements in all, with RJ accounting for half of the total at thirty-five statements. Because the kind of criteria is in question here and not the amount, her data will not be separated out.

Not surprisingly, almost all of the criteria mentioned by the control students were closely tied to the classification task. Over three quarters (77.5%) of the statements referred to the similarities or differences between arthropods (i.e., "I could classify them by their legs but that didn't really make any sense because they all have different legs"); another fourteen percent referred to the size of the group created by a certain approach to classification (i.e., "I couldn't do it by where they live because that would only be one [in that group]"), and the remaining nine percent was split between references to attempts to have only one arthropod in each group, the ease of learning and remembering the classification system, the comprehensiveness of the system, and whether or not the system made sense in terms of similarities and differences found between human beings. These last two criteria—comprehensiveness and a comparison with humans—were made only by RJ.

In summary, this analysis shows that many students do assess their work spontaneously. The degree to which they assess themselves varies widely, from none at all to more than once per every three lines of text in their think aloud protocols (as represented by the highest PCTASMT value, 0.37). Ninety-one percent of the criteria referred to by the control students had to do with 1) the similarities and differences between the arthropods, and 2) the size of the group created by a particular classification system. Comparisons between these criteria and those referred to by subjects in the treatment group will be made in the section that addresses Question 2.

Question 2: What kinds of self-assessment are students capable of under supportive conditions?

This question has three sub questions, a) to what degree do students assess themselves under supportive conditions? b) what criteria do they use? and (c) what do their prompted self-assessments look like?

Degree of self-assessment. The treatment group's think aloud protocols also provide evidence of spontaneous self-assessment. Sixteen of the twenty treatment subjects made at least one statement coded as a positive, negative or neutral evaluation, and the remaining four subjects made between two and five self-corrections. The number of assessment statements ranged from zero to sixty-five. Table 7 contains the descriptive statistics for the raw data, NASMT, and the transformed variable PCTASMT.

Table 7
Descriptive Statistics for NASMT and PCTASMT Scores for Treatment Subjects

	Mean	Standard Deviation	Median	Range
NASMT (N = 20)	16.75	18.03	10.5	0 - 65
PCTASMT (N = 20)	0.19	0.11	0.2	0 - 0.37

Criteria used in self-assessment. Fourteen students in the treatment group made seventy-three criterion-referenced evaluative statements in all. The number of criterion-referenced statements ranged from zero to fifteen, so no one student contributed more than twenty percent of the total number of statements.

As with the control group, the treatment subjects referred most often to the similarities and differences between the arthropods (65.8%) and to the size of the categories created by a classification system (16.4%) when evaluating their work. However, 14% of the treatment subjects' criterion-referenced evaluations referred to one of three criteria from the rubric: 1) the classification system is based on important physical characteristics, 2) each arthropod can only fit into one category, and 3) any errors in placing the arthropods in groups are corrected. The remaining 4% of the criteria used by treatment subjects referred to the need for more information and a desire to challenge oneself to come up with a "creative" classification system.

Differences in patterns of criteria usage by each group was tested by constructing the contingency table in Table 8 and calculating a chi-square statistic. Table 4.3 shows that 10 of the criteria referred to by treatment group were contained in the rubric, and 63 were not. In contrast, 2 of the criteria referred to by the control group were contained in the rubric, and 69 were not. The chi-square statistic of 5.6 ($p < .025$) indicates that the difference between the two groups in terms of usage of criteria from the rubric is not likely to be due simply to chance.

Table 8
Number of Criterion Referenced and Non-Criterion Referenced Assessment Statements as a Function of Group

	Treatment	Control
Rubric	10	2
Non-rubric	63	69

Prompted self-assessments. All of the treatment subjects readily assessed their work according to the rubric when asked to do so. Regardless of the score they gave themselves, they were asked if they wanted to try to improve upon their work. An analysis of their responses provides at least partial answers to the following questions:

1. Were the treatment subjects' assessments of their work correct?
2. How did they respond when asked if they wanted to improve their work?
3. If they could identify a way to improve their work but did not act on it, what were their reasons?
4. If they did not try to improve their work after giving themselves a three or less on the rubric, what reasons (if any) did they give?

Were the treatment subjects' assessments of their work correct? This question refers to the correctness of students' self-assessments, not to the correctness of their work. For example, if a student indicates that she thinks her system is not based on important physical characteristics and in fact it is not, her self-assessment is correct, although her work is not. This question can be answered in terms of parts or all of Criteria 2, 3 and 4 from the rubric, which read:

Criterion 2: I checked the page about arthropods to make sure my classification system is accurate

Criterion 3: I specified useful categories for the arthropods

Criterion 4: I accurately placed the arthropods in the categories

The data afford an opportunity to evaluate whether or not students were correct in their assessments of whether or not they "reread the page" (Criterion 2), whether or not they "created categories based on important physical characteristics of the arthropods" (Criterion 3) and whether or not each arthropod "only fits in one category" (Criteria 3 and 4). There is no reliable evidence of correctness or incorrectness of students' self-assessments for Criteria 1, "I read the page about arthropods carefully to make sure I understood it," or for Criteria 5, "I described accurate and complete rules for deciding which arthropods go in each category." These criteria would require an analysis of students' reading comprehension strategies and of the structure and quality of their explanations, both of which are beyond the scope of this study.

How did students respond when asked if they wanted to improve their work? Students tended to say or do one or two of five different things when asked if they wanted to improve their work after assessing it. Each statement or action is captured in the coding system discussed above and in the summary in Table 9.

Table 9
Treatment Subjects' Responses to the Opportunity to Improve their Work

Student assigned him- or herself a 4 then,	
• chose to revise work	7
• chose not to revise work	28
• felt improvement was possible but could not identify a way to improve	1
• could articulate a way to improve but chose not to act on it	4
Students assigned him- or herself a 3 or less then,	
• chose to revise work	33
• chose not to revise work	15
• felt improvement was possible but could not identify a way to improve	2
• could articulate a way to improve but chose not to act on it	6
• reconsidered, decided on higher score without revising	4

The numbers in Table 9 total one hundred because the twenty treatment subjects were asked to assess themselves on five criteria each (20 x 5 = 100). Of particular note is the fact that students were almost five times more likely to choose to revise their work when they had assigned themselves a three than when they had assigned themselves a four, the highest score possible on the rubric. Nonetheless, on seven occasions even students who had assigned themselves a four decided to improve upon their work anyway.

If students could identify a way to improve their work but did not act on it, what were their reasons? Students spontaneously gave reasons for six of the ten occasions when they chose not to improve upon their work even when they could articulate a way to do so. One reason was that the solution the student had in mind would be unwieldy or unsatisfactory: One student felt that the only way to avoid having an arthropod fit into more than one category would be to put them all into their own groups, "and then there would be so many groups it would be impossible to know them all." Two of the reasons were based on the students' beliefs that improvement was not feasible or convenient: "If I did this for hours, maybe I could improve," and "The only way I'd be able to tell [if it was correct or not] is if I looked it up or checked it with some information."

The remaining three reasons were related to the students' beliefs that their work was good enough the way it was. A girl who typically performs poorly in school told me simply, "A three is OK." Another student said that, even though she knew how she could improve her score on reading the passage about arthropods, since she could refer back to it at any time it wasn't necessary to "memorize it." Also in reference to the reading of the passage, one student said that rereading wasn't necessary because "the important parts that it said to remember, I remember."

If students did not try to improve their work after giving themselves a three or less on the rubric, what reasons did they give? Of the fifteen occasions when a student chose not to try to improve his or her work after assigning it a three or less on the rubric, five were accompanied by no reason or an unclear reason. Of the remaining ten occasions, students said that they could not think of any way to improve their work six times. Twice students felt they could not meet Criterion 1, which required that they "learn something new" from the page about arthropods, because they had just studied this subject in science class and already knew the information. Once a student said simply that his work was the best he could do. And finally, one student sensed a contradiction between the rubric and the instructions he received: The rubric required him to check his classification system against the information on the page, but the instructions asked him to create his own system. This student chose to heed the instructions, not the rubric.

In summary, the data shows that the treatment subjects also assess their work spontaneously, but that they used criteria from the rubric significantly more often than the control subjects. Treatment subjects also readily assessed themselves in terms of the rubric when asked to do so.

Question 3: Does self-assessment influence metacognitive involvement in the classification task?

This question concerns the differences between the treatment and control groups in terms of students' levels of metacognitive engagement, as represented by the percentage of statements uttered by the students during the classification and explanation parts of the procedure that were coded as metacognitive (PCTMCOG). This variable includes all of the coding categories, including the number of assessment statements made, the number of ideas generated, metaknowledge, metatask, meta-experience, and process goal-setting.

In the analysis that follows, the relationship between this outcome variable, gender (GNDR), scores on the California Achievement Test (CAT), and experimental condition (GRP) will be examined in an attempt to determine the best predictors of students' level of metacognitive engagement. The two special education students are not included in the analyses in this section because they do not have CAT scores.

Univariate Analysis
The data for the PCTMCOG variable form a relatively normal, bell-shaped distribution. Separate PCTMCOG scores for the treatment and control groups are summarized in Table 10. The means are almost the same, but the control group's standard deviation is quite a bit larger than the treatment group's.

Table 10
Descriptive Statistics for PCTMCOG Variable for Treatment and Control Groups

	Treatment	Control
Average	0.53	0.52
Standard deviation	0.15	0.23
Median	0.54	0.46
Range	0.19 - 0.78	0.22 - 1.11

Separate descriptive statistics for boys and girls on the PCTMCOG variable can be found in Table 11. The average score is nearly the same. The average CAT score for the 38 students who took the test was 752.76, with a standard deviation of 29.06. The distribution of scores closely resembles a normal bell curve. The scores for the treatment and control groups are quite similar: The average for the treatment group was 752.84 (SD = 29.33), and for the control group was 752.68 (SD = 29.59).

Table 11
Descriptive Statistics for Boys and Girls on PCTMCOG Variable

	Girls	Boys
Average	0.53	0.52
Standard deviation	0.16	0.23
Median	0.56	0.45
Range	0.22 - 0.79	0.19 - 1.11

Separate CAT statistics for boys and girls are shown in Table 12. The average CAT score for the girls is more than 26 points higher than the average for boys.

Table 12
Descriptive Statistics for Boys' and Girls' CAT Scores, n = 38

	Girls	Boys
Average	764.43	738.35
Standard deviation	24.73	28.09
Median	761	737
Range	714 - 828	689 - 781

Bivariate Analysis

t-tests. Two-tailed t-tests were used to analyze the relationships between experimental condition (GRP) and CAT scores, GRP and PCTMCOG, gender (GNDR) and PCTMCOG, and GNDR and CAT scores. The t-tests showed that the treatment and control groups were equivalent in terms of mean CAT scores (t-statistic = -0.017, p = .987) and mean PCTMCOG scores (t-statistic = -0.096, p = .924), suggesting that the GRP
variable may not lend much explanatory power to a fitted regression model.

A t-test showed a highly significant difference between boys and girls on the achievement test, with girls outscoring boys on average (t-statistic = -3.0, p = .005). No significant difference between boys and girls was found for PCTMCOG (t-statistic = -0.27, p = .79).

Correlation. I examined a plot of PCTMCOG versus CAT for any sign of a relationship between them. The plot contains a sizable amount of scatter, but still suggests a weak positive correlation between the two variables. This was confirmed by the estimated correlation of .24. CAT, which will be the third predictor in the regression model, also appears to have little potential in terms of explanatory power.

In summary, the t-tests and estimated correlation coefficient reported in this section suggest that neither gender, CAT scores nor experimental condition are strongly related to PCTMCOG. A regression model was fit to examine the relationship between PCTMCOG and these three predictors more carefully.

Regression Analysis

Simple linear regression. Table 13 summarizes the results of the simple linear regression analyses. On average, for every 100 point difference in the score on the CAT, a student's PCTMCOG score differed by only two-tenths of a percent. Students in the control group averaged just over two percentage points higher than students in the treatment group in terms of the percent of PCTMCOG statements uttered, and girls scored just over two percentage points higher than boys. The residuals for each model appeared to be approximately randomly scattered, with no unusual patterns. The regression assumptions, therefore, have not been violated.

Table 13
Simple Regression Models for GRP, CAT and GNDR

	ß	se(ß)	t	R Square statistic	p
GRP	-0.021	0.061	-0.352	0.003	.728
CAT	0.002	0.001	1.458	0.056	.153
GNDR	0.021	0.060	0.351	0.003	.728

As suggested by the t-tests and correlation analyses, none of the three predictors alone explain much of the variance in PCTMCOG. Experimental condition and gender each explain only .003% of the variance, and CAT scores explain less than 6%. None of these relationships were statistically significant, however.

Multiple linear regression. In order to examine the combined power of the three predictors, a multiple regression model was fit. Table 14 shows the hierarchical construction of the model. The first model is a simple, uncontrolled regression of PCTMCOG on GNDR, identical to the model in the table above. In the second model, the CAT variable was added to GNDR. The slope coefficient and standard error for GNDR did not change very much with this addition, indicating that GNDR and CAT each contribute independent information. The R-square statistic increased by .056, but the model still explains less than 6% of the variance in PCTMCOG.

Model 3 predicts PCTMCOG on the basis of GNDR, CAT and GRP. The R-square statistic only increases by .004, which means that, in combination, the three predictors only explain 6.3% of the variance in PCTMCOG. A test of the influence of the two unusual data values discussed earlier resulted in a model with a slightly larger R-square statistic (R-square statistic = .16) but it did not reach statistical significance (p = .12).

-- insert Table 14 here --

A test for interactions revealed one statistically significant interaction between GNDR and GRP (t-statistic = 2.68, p = .01). The effect of gender on PCTMCOG scores, therefore, differs by group. Figure 1 shows that, on average, girls in the treatment group have higher PCTMCOG scores than girls in the control group, and the opposite relationship exists for boys.

The R-square statistic for the regression model that includes the GNDR by GRP interaction almost quadruples (R-square statistic = .23) and the model reaches statistical significance at the relaxed .10 level (p = .06). The residuals for the interaction were random, as were the residuals for the final regression model including the interaction.

Table 14

Multiple Regression Models for Metacognition (PCTMCOG) Data (n = 38)

Model	GNDR ß1	t	CAT ß2	t	GRP ß3	t	GNDR* GRP ß4	t	R-sqr	F (df)	p
1	0.02	0.35							.003	0.12 (1, 36)	.73
2	0.02	-0.33	0.002	1.43					.06	1.09 (2, 35)	.35
3	-0.02	-0.35	0.002	1.43	-0.02	-0.38			.06	0.76 (3,34)	.53
4	-0.18	-2.09*	0.002	1.6	-0.19	-2.27*	0.11	2.68**	0.23	2.47 (4, 33)	.06

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

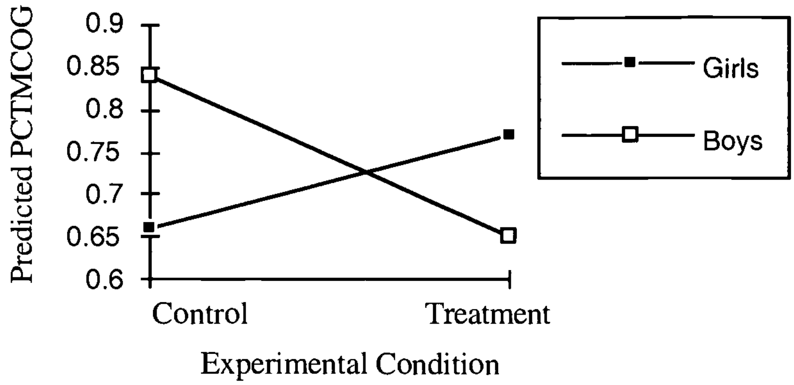

Figure 1. Predicted PCTMCOG for Girls and Boys in the Treatment and Control Groups, Holding CAT Score at its Mean

In summary, these analyses show that 23% of the variance in PCTMCOG can be attributed to a combination of gender, CAT scores, experimental condition and an interaction between gender and condition.

Question 4: Does self-assessment influence students' learning about classification and arthropods?

This question concerns the differences between the treatment and control groups in terms of content learning, as represented by the difference between pretest and post-test scores (DIFF). In the analysis that follows, the relationship between this outcome variable (DIFF), gender (GNDR), scores on the California Achievement Test (CAT), and experimental condition (GRP) will be examined in an attempt to determine the best predictors of content learning.

Univariate Analysis

The distribution of the DIFF data approximates a bell-shaped distribution. The range is from a one-point decrease to a twenty-three point increase in total score, with an average of 9.88 and a standard deviation of 4.92.

Separate DIFF values for the treatment and control groups are summarized in Table 15. The average pre- to post-test difference for the treatment group is more than four points higher than the average difference for the control group, and the standard deviation is less than two points higher.

Separate descriptive statistics for boys and girls on the DIFF variable can be found in Table 16. The means are less than one point apart, and standard deviations just over one point apart, suggesting there is little difference between boys and girls in terms of this variable.

Table 15
Descriptive Statistics for DIFF Variable for Treatment and Control Subjects, n = 38

	Treatment	Control
Average	11.91	7.55
Standard deviation	5.21	3.41
Median	11	7
Range	3 - 23	-1 - 12

Table 16
Descriptive Statistics for DIFF Variable for Boys and Girls, n = 38

	Girls	Boys
Average	9.61	10.24
Standard deviation	5.44	4.24
Median	9	10
Range	-1 - 23	4 - 19

Bivariate Analysis

t-tests. Two-tailed t-tests were used to analyze the relationships between experimental condition (GRP) and DIFF and gender (GNDR) and DIFF. The t-tests indicate no significant difference between boys and girls on the DIFF variable (t-statistic = 0.39, p = .7), but a highly significant difference between the treatment and control groups (t-statistic = -3.36, p = .002). This suggests that GNDR will not contribute much explanatory power to a regression model, but that there may be an effect of treatment.

Correlation. I examined the plot of DIFF versus CAT for any sign of a relationship between the outcome and predictor. Surprisingly, the plot resembles a random distribution of data. A test of correlation showed a very weak positive correlation of 0.057, suggesting that CAT has little potential in terms of explanatory power.

Regression Analysis

Simple linear regression. Table 17 summarizes the results of the simple linear regression analyses. On average, students in the treatment group scored 4.65 points higher than students in the control group. A 100 point difference in CAT score was associated with a 0.9 point difference in DIFF, and girls tended to score 0.63 points lower than boys. An examination of the residuals for each model revealed that the regression assumptions have not been violated.

As suggested by the t-tests, there is a highly significant relationship between GRP and DIFF. GRP explains 23% of the variance of DIFF and the relationship is statistically significant at p = .002. Neither of the other two predictors explain a statistically significant amount of variance.

Table 17
Simple Regression Models for DIFF on GRP, CAT and GNDR

	ß	se(ß)	t-statistic	R Square statistic	p
GRP	4.32	1.42	7.79	.20	.002
CAT	0.009	0.028	0.34	.003	.74
GNDR	-0.63	1.59	-0.39	.004	.7

Multiple Regression. In order to examine the combined power of the three predictors, a multiple regression model was fit. Table 18 shows the hierarchical construction of the model. The first model is a simple, uncontrolled regression of DIFF on GNDR, identical to the model in the table above. The second model adds CAT to the equation. The resulting R-square statistic is not statistically significant at .009, p = .85.

Model 3 is the final multiple regression model for predicting DIFF on the basis of GNDR, CAT and GRP. The R-square statistic jumped to .21 and is statistically significant at the .05 level. In combination then, GRP, CAT and GNDR account for 21% of the variance in DIFF. The greatest proportion by far belongs to GRP. The slope coefficient for GRP indicates that, holding CAT and GNDR constant, treatment subjects scored 4.29 points higher in terms of DIFF on average than did control subjects.

-- insert Table 18 here --

The residuals for the final regression model are randomly distributed. A test for interactions found no statistically significant interactions between the predictors used in this model.

In summary, these analyses have shown that 21% of the pre- to post-test difference in scores can be attributed to the combined effect of students' gender, CAT scores and experimental condition. The greatest proportion of the explanatory power of the regression model belongs to the GRP variable, indicating that the treatment had an effect on content learning.

Discussion

This discussion has five parts. In Part One, I discuss the findings related to my first research question: Do student assess themselves spontaneously? In Part Two, I

Table 18

<u>Multiple Regression Models for DIFF Data (n = 38)</u>

	GNDR			CAT			GRP					
Model	ß1	se(ß1)	t	ß2	se(ß2)	t	ß3	se(ß3)	t	R-sqr	F (df)	p
1	-0.63	1.59	-0.39	-	-	-	-	-	-	.004	0.16 (1,38)	.7
2	-0.85	1.82	-0.47	0.02	0.03	0.51	-	-	-	0.009	0.17 (2,35)	.85
3	-0.55	1.65	-0.33	0.01	0.03	0.48	4.29	1.46	2.93*	0.21	2.99 (3,34)	.04

* $p < .05$

discuss the quantity and quality of student self-assessment under supportive conditions as defined in this study. In Part Three, I propose an answer to the question of whether self-assessment influences students' metacognitive engagement. Part Four is a discussion of the effect of self-assessment on subject matter learning. Part Five highlights key findings of the study, discusses implications for educational practice and suggests directions for future research.

Part One: Do students assess themselves spontaneously?

This study showed that most but not all students can and do assess themselves spontaneously as they create and/or explain a classification system. The fact that three-quarters of the students in the control group assessed their work and/or their approach to it at least once is an encouraging finding because it suggests that self-assessment is not an unrealistic expectation for seventh graders. At the same time, the degree of self-assessment suggests that there is room for improvement: five of the twenty students did not assess themselves at all, and three students made only two or three evaluative statements. Only one student, who made 108 evaluative statements which accounted for nearly 26% of her lines of text, could potentially be said to be performing at ceiling. The performance of each of the other students can conceivably be improved through instruction and practice.

The criteria employed by the control students were closely tied to the particular classification task at hand and mostly of a "lower order" nature. By "lower order" I mean referring to only the most basic demands of the task (i.e., grouping the animals according to their similarities), as opposed to "higher order" considerations such as the comprehensiveness, elegance or creativity of the classification system created. Over 94% of the criteria used by the control group can be characterized as lower order. It may be that a primary purpose of self-assessment—at least self-assessment according to a rubric, as done in this study—is to alert students to higher order criteria that they may otherwise overlook.

The results from the treatment group also show that students can and do assess their own work spontaneously: Sixteen of the twenty students made at least one evaluative statement during classification and/or explanation without prompting (that is, without being asked to rate themselves according to the rubric—statements made in response to the rubric were not counted in this part of the analysis). Again, the major portion (82%) of the criteria students relied on in their assessments were of a lower order nature, referring to similarities and differences between arthropods and the number of animals in each group. However, 14% of the criteria were more sophisticated in that they referred to the general rules of classification contained in the rubric: i.e., the classification system must be based on important physical characteristics, and no one animal should be able to fit into more than one group. The remaining 4% can be considered higher order criteria because they referred to the need for more information in order to do the task well, and the desire to produce a "creative" classification system. This finding also suggests that criterion-referenced self-assessment may be effective in promoting the use of sophisticated criteria and standards when students judge their work.

Part Two: What kinds of self-assessment are students capable of on this task under supportive conditions?

The analysis of the treatment students' prompted self-assessments in terms of the rubric shed light on several interesting questions concerning patterns of criteria usage, the correctness of students' self-evaluations, and students' responses to opportunities to improve their work.

Patterns of criteria usage. The statistically significant difference between the treatment and control groups in terms of the kinds of criteria used (rubric versus non-rubric) shows that self-assessment according to written criteria can influence students' spontaneous evaluative thoughts. A study of longer duration could be expected to have practical as well as statistically significant effects on students' thinking.

Correctness of student self-evaluations. Correctness was difficult to judge because of ambiguities in the rubric and in the pictures of the arthropods, but even my relatively generous definition of "correct" resulted in only about two-thirds (68.3%) of the assessments were correct. This number falls within the range of correlations found in research on the validity of self-ratings of ability as compared to objective tests or teacher ratings. It is, however, a very wide range: In his review, Oskarsson (1984) cites correlations between .39 and .92. It appears that a variety of task variables and person variables explain the differences. For example, self-ratings of second language oral proficiency were more valid than self-ratings of written proficiency, self-ratings of concrete tasks were more valid than ratings of general skill level, and good students tended to underrate themselves, while less proficient students tended to overrate their ability.

The conditions of this study predicted validity measures on the high end of the range, however, because self-assessment referred to a concrete task with clear criteria. Sixty-eight percent correct is not particularly high. Several explanations are possible. The simplest is that the ambiguities in the rubric and pictures confused students, and my expanded definition of "correct" did not account for every error. This is a very real possibility, as I explained in the Results section. Another explanation is that students did not understand key terms and phrases that were not necessarily unclear but were unfamiliar to them. Although "important physical characteristics" may seem clear enough, it is possible and even probable that some students misinterpreted it, or at least interpreted it differently than I did. One definition likely to have been used by students includes habitat as well as number of legs, wings and so on. There is evidence that this definition was employed by a number of students.

The obvious implication for future research and practice is to ensure that students understand the terms and concepts in the rubric. Such misunderstandings are less likely in instructional practice because the rubric would be used to assess the content being taught, but clarifying terms and concepts used in self-assessment is still an important concern.

Both of the above explanations for the relatively mediocre number of correct self-assessments are based on the clarity of the information contained in the rubric. In their review of the role of feedback in self-regulated learning, Butler and Winne (1995) point out that other kinds of explanations are needed:

> ...considering feedback merely in terms of the information it contains is too simplistic. Rather, learners interpret such information according to reasonably stable and relatively potent systems of beliefs concerning subject areas, learning processes, and the products of learning. These beliefs influence students' perceptions of cues, their generation of internal feedback, and their processing of externally provided feedback. In the last case, beliefs filter and may even distort the message that feedback is intended to carry. Moreover, characteristics of information in elaborated feedback... influence how a learner will use feedback (p. 254).

Butler and Winne base this claim in part on a framework developed by Chinn and Brewer (1993) to explain the nature of and means for changing students' entrenched views and misconceptions about scientific principles. Chinn and Brewer have identified four factors that influence conceptual change: (a) the nature of a student's prior knowledge, (b) characteristics of a new model or theory meant to replace the student's inadequate or misconceived one, (c) aspects of anomalous information presented to the student in order to signal that his or her current conceptual structure is inaccurate, and (d) the depth of processing the student engages in when considering the anomalous data.

I have found this framework useful in thinking about the correctness of students' self-assessments. For example, the two explanations I gave above for the relatively low number of correct self-assessments already take into account students' prior knowledge (e.g., what they think "important physical characteristics" means) and at least some characteristics of the model or theory meant to replace a given student's inadequate one (the clarity of the information provided by the rubric). However, it is also necessary to think about the aspects of anomalous information presented to the student in order to signal that his current conceptual structure is inaccurate, and the depth of processing the student engages in when considering the anomalous data.

I believe that some portion of the incorrect self-assessments generated by the students in my sample can be explained by an interaction between these two forces. That is, in order for anomalous information to signal to a student that his thinking is inaccurate, he must engage in relatively deep processing (i.e., active and elaborate processing) of the data at hand. This is because the student himself is determining whether or not the information is anomalous. Take, for example, the student whose classification system is based on a combination of habitat and physical characteristics. When faced with the task of rating himself in terms of a criteria that states, "I created categories based on important physical characteristics of the arthropods," he can either stick with his belief that habitat is a physical characteristic, or he can stand back and question whether in fact his assumption is true. The latter option requires deeper processing of information. In the context of self-assessment this deep processing is of particular importance because the rubric itself does not explicitly present information as "anomalous" or indicate correctness or incorrectness. It simply provides a criterion for the student to consider. If he chooses not to engage in the deep processing required to consider the criterion carefully, his chances of being incorrect in his assessment increase.

In general then, self-assessment done well requires a well-developed understanding of the words and concepts being assessed as well as deep processing of information in order to reveal misconceptions and incorrect work. Students who do not understand the content or are not motivated or able to engage in deep processing may be at a disadvantage. Educators using self-assessment techniques should be prepared to provide assistance to such students.

Responses to the opportunity to improve. The second question addressed by my analysis of the treatment subjects' self-assessments concerns how they responded when asked if they wanted to improve their work. This question has less to do with self-assessment *per se* than with one of the basic characteristics of authentic and self-assessment in practice: Assessment must happen while there is still time to revise. That is, rather than occurring at the end of a unit or project, when it serves primarily to tell teachers what students do and do not know, authentic assessment occurs repeatedly over the course of the unit or project and serves to indicate to the students their strengths and areas in need of improvement. The following analysis suggests that students tend to respond positively to this aspect of self-assessment.

In 40 of the 100 times that students were asked if they wanted to try to improve their work they said yes. Seven of the 40 were from students who had already given themselves the highest possible rating on the rubric. Considering the circumstances of the study and the results of prior studies of students' poor revision habits (Nold, 1982; Scardamalia & Bereiter, 1983), these strike me as remarkably high numbers and quite encouraging. In a study of novice and expert writers, for example, Scardamalia, Bereiter and Steinbach (1984) note that students' texts are typically devoid of substantive revision, suggesting a failure to rethink first-made decisions.

Take first the circumstances of the study: Students were told that their work did not count toward a grade and that, since I just wanted to study their thinking and not the classification system they produced, there were "no right answers" (a white lie, I admit, but I thought at the time that it was necessary to set students at ease). Hence there was no motivation to do well on the classification task in order to get a better grade in school. So why were students motivated to attempt to improve their work?

It is possible that students wanted to please the researcher, a classic research complication. It is also possible that students enjoyed getting out of class, wanted to make it last as long as possible and saw revision as a way to do so. It is also possible that students enjoyed the task enough to want to continue it. This possibility seems less likely than the others, although one student did say that she enjoyed talking out loud and having me listen to her.

Taken together, these three explanations may account for some portion of the motivation to revise demonstrated by the students, but the numbers nonetheless fly in the face of prior research and common experience, which shows that students typically do not revise their work. My research suggests that this phenomenon may be due less to an inability or lack of motivation to revise than to the absence of a cue to self-assess and of clear criteria to guide improvement. In fact, one student told me that having the rubric made the task easier "because I knew what I had to do." My own experience with children and adults echoes this statement: Revision is more likely when learners have some indication of what it takes to improve. The rubric used by the treatment subjects provided just such guidance.

This study was not designed to determine whether students in the treatment group were more likely to revise than those in the control group, but an informal look at the data reveals eight instances of revision during or after treatment students explained their classification systems, as compared to two instances in the control group. Both the formal and informal analyses of students' responses to the opportunity to improve their work suggest that an investigation of the influence of explicit cues and criteria on revision behaviors could be quite valuable.

Although the number of times students in the treatment group chose to revise is pleasantly surprising, there were nonetheless sixty occasions when students chose not to revise their work. Twenty-seven of the 60 were from students who had given themselves a rating of 3 or less on the rubric, which should have suggested to them that there was room for improvement. In 10 instances students could identify a way to improve but chose not to act on it. The next question has to be, why did students chose *not* to revise?

Students were not systematically interviewed about their reasons for choosing not to revise in this study, but many articulated their reasoning spontaneously. Of the 27 occasions when students who gave themselves a rating of 3 or less chose not to revise, only seven were accompanied by a strong reason: two girls said they could not meet the demand that they "learn something new" from the page of information about arthropods because they already knew it all; one boy said he was following my directions rather than the adhering to the rubric; and four students reread the rubric and decided that they had rated themselves incorrectly and their work actually deserved a higher rating. On the other 20 occasions students either said they could not think of a way to improve (9), could think of a way to improve but didn't want to follow through on it (6), or gave no reason or an un-interpretable reason (5).

The 20 times that students gave no reason or a weak reason for not attempting to improve their work can be explained in many of the same terms that correctness and revision were: Motivation and deep processing. In the case of non-revision, however, there may have been a lack of motivation to improve, and/or a failure to process the information deeply enough to realize a potentially fruitful approach to revision.

Other explanations are also worth considering. I present them in the spirit of "model building" as done in multiple regression analysis. Each explanation may contain some small portion of explanatory power, and in combination they may explain a lot.

One possible explanation grows out of evidence that suggests that the students did not consider the rubric an "authority" on quality, at least as compared to the pre- and post-tests and to the verbal instructions I gave them regarding the nature of the classification task. Take, for example, the boy who gave himself a rating of 1 because he did not reread the page about arthropods to see that his classifications system was consistent with the information there: He saw no need to attempt to improve his score because the classification system "wasn't really based on actual fact, it was based on how it would be to me." He went on to explain that I had asked him to "make up" a classification system and that was what he did. He knew the official, textbook classification system for arthropods because he had just studied it in science class, but he followed my directions as he heard them and made up a new one. That being the case, the official, textbook information contained in the page about arthropods was irrelevant, as was the criterion in the rubric that required he check his work against it.

Questions posed to other students following this boy's comments revealed that his interpretation of the task was widespread: Many students believed that I wanted them to "make something up" rather than rely on their knowledge of arthropod classification. As a result, the criteria contained in the rubric may have seemed to contradict the instructions I had given and students may have chosen to ignore them or felt that it was impossible to meet them by revising their work. My "white lie" about no right answers on the classification task has come back to haunt me: At least some students apparently took me at my word and disregarded the criteria in the rubric.

There is additional evidence suggesting that the rubric lacked authority in the students' minds. At the end of thirteen of the twenty sessions with treatment subjects, I asked the students how they thought they did on the task and why. Four of the thirteen students based their summative evaluations on how well they thought they did on the post-tests. This isn't surprising, since most students are accustomed to judging themselves and being judged in terms of test scores. Rather, it reinforces my suspicion that students did not always see the rubric as an authority. If that was sometimes the case, it may help explain why students chose not to revise their work when they gave themselves a 3 or less on the rubric: It wasn't, in their minds, a serious judge of quality.

The last explanation to be added to my "model" explaining why students sometimes chose not to improve their work concerns Butler and Winne's (1995) claim that student beliefs must be accounted for in any consideration of the role of feedback in self-regulated learning. A body of literature that is relevant here concerns children's theories of learning (Dweck, 1975; Dweck & Bempechat, 1983). Dweck and her colleagues have shown that children tend to hold one of two theories about the nature of intelligence. An "entity" theory holds that intelligence is trait-like and immutable—you either have it or you don't. This theory is associated with a sense of helplessness and lack of control over one's abilities. In contrast, an "incremental" theory of intelligence holds that intelligence can be developed through effort.

The theory of intelligence a child holds influences her selection of achievement goals and learning behaviors. Children who hold incremental theories of intelligence tend to be concerned with building their competence and learning from experiences, while those who hold entity theories are concerned with "looking smart," even if it means avoiding challenging tasks. Self-regulated learning behaviors follow from an incremental theory of intelligence, as does the motivation to engage in deep processing. More superficial processing and work avoidance stems from an entity theory (Nolen, 1988). Research into the relationship between epistemological beliefs and affect during self-regulation support these findings. For example, Carver and Scheier (1990) have shown that students who believe in quick learning may withdraw from tasks in which they progress more slowly than anticipated.

The distinction between incremental and entity theories of intelligence and their effects on behavior may help explain why some students in this study simply chose not to attempt to revise their work. It is possible that some of their choices were rooted in the belief that they could not improve and might even embarrass themselves by trying and failing. The special education student who said simply, "A 3 is OK" when asked if she wanted to improve her work is a likely candidate for this explanation. Her difficulties in school may have resulted in the belief that she had no control over her ability, a sense of helplessness, and a tendency to avoid challenging work. Choosing not to attempt to improve her work (or at least her rating of it) would be a natural and predictable response.

It is also possible that students' beliefs influenced the goals they selected when given the classification task to complete and that, in turn, the selected goals helped determine the tactics and strategies in which students engaged. Studies by Schutz (1993) and Winne (1991) have shown that multiple variables simultaneously affect students' selection of goals and the relations that feedback has to those goals. Most importantly, students judge their own performance in terms of the goals they have selected. If, for example, a boy in this study selected a rating of 3 as his goal, he would likely chose not to revise his work once that goal was obtained, regardless of the implied expectations of the rubric or the researcher.

The results from this study provide evidence that the relationship between self-assessment and revision is complex but often fruitful. Future research on this relationship should address issues of motivation, authority, beliefs about learning and the types of goals that students select as a result of those beliefs.

Part Three: Does self-assessment influence students' metacognitive engagement in the classification task?

Analyses showed the treatment and control groups were statistically equivalent in terms of metacognitive processing, indicating that self-assessment did not increase metacognitive involvement in the classification task overall. However, the interaction between gender and treatment was highly significant. This interaction shows that criterion-referenced self-assessment had a positive effect on girls' metacognitive engagement in the classification task, but a negative effect on boys'. In broad stroke, this finding is consistent with research on sex differences in responsivity to feedback and in achievement motivation and learned helplessness, which has generally shown that girls and boys differ both in their attributions of success and failure, and in their response to evaluative feedback (Dweck & Bush, 1976; Dweck, Davidson, Nelson & Enna, 1978). However, the patterns found in this study do not match those seen in Dweck's research. Briefly, research by Dweck and others (Deci & Ryan, 1980; Hollander & Marcia, 1970) have shown that girls are more likely than boys to be extrinsically motivated and to attribute failure to ability rather than to motivation or the agent of evaluation. As a result of these attributions, girls' performance following negative adult feedback deteriorates more than boys' performance.

This study suggests that self-generated feedback has a very different effect than negative adult feedback has on girls' performance, as reflected by the fact that the girls' scores on the metacognition variable appears to have been enhanced by self-assessment. Some interesting contradictions in the research literature suggest that this may not be peculiar to my research. A study by Roberts and Nolen-Hoeksema (1989) found no evidence that women's greater responsivity to evaluative feedback led to performance decrements, suggesting that women's maladaptive responsiveness to feedback is not absolute. Also of interest are earlier studies by Bronfenbrenner (1967, 1970), which found that when peers instead of adults delivered failure feedback, the pattern of attribution and response reversed: Boys attributed the failure to a lack of ability and showed impaired problem solving while girls more often viewed the peer feedback as indicative of effort and showed improved performance.

Noting that the more traditional finding of greater helplessness among girls was evident only when the evaluators were adults, Dweck et al. (1978) have taken these findings to mean "that boys and girls have not learned one meaning for failure and one response to it. Rather, they have learned to interpret and respond differently to feedback from different agents" (p. 269). This seems a reasonable conclusion to draw, and relevant to the gender differences found in this study. Although this research does not allow me to examine students' attributions of success or failure, the girls' improved performance on the metacognition variable does suggest that they attributed their scores on the rubric to effort and responded by being more metacognitive and self-regulating, while the boys attributed their scores to ability and responded by engaging in less self-regulation. Such a response reflects an intrinsic, rather than an extrinsic motivation on the part of girls (Boggiano & Barrett, 1985), leading me to further speculate that self-assessment fosters an intrinsic orientation.

The above explanation for the differences between boys and girls on metacognition scores are largely speculative, however. This study was not designed to provide evidence of students' attributions and orientations, and informal analyses of indirect evidence did not reveal gender differences at all, much less explain them. For example, I compared 11 girls' and 8 boys' responses to my follow-up question, "How do you think you did on this task and why?" There were no clear differences in effort versus ability attributions. I also compared treatment boys and girls in terms of their willingness to improve upon their work, a possible indicator of an effort attribution. Again, no clear patterns emerged: The 11 girls revised on 21 occasions, and the 9 boys revised on 20 occasions. Research that provides data on students' attributions of success and failure and their perceptions of who is assessing whom is needed.

Part Four: Does self-assessment influence students' learning about classification and arthropods?

The multiple regression analysis shows that experimental condition had a highly significant effect on pre- to post-test gains. Thus, there is support for my hypothesis that self-assessment can increase content learning.

The explanation for these results is somewhat less straightforward. Recall that the reasoning behind my hypothesis was that self-assessment would increase cognitive monitoring, which would prompt deeper processing of the material, which would in turn result in more learning. The results for the girls in the study appear to support this model: Girls who were asked to assess themselves were more metacognitive and learned more than the girls who were not. However, the study did not provide evidence of increased monitoring by boys in the treatment group, and there was no effect of gender on learning, suggesting that there is no overall link between metacognition and learning. This was confirmed by a test of correlation, which showed almost no relationship between the two outcomes ($r = -.06$). Because there is no clear relationship between metacognition and learning, an alternative explanation for these results must be sought.

The results might be better represented by a model that does not include metacognition. It is possible that self-assessment caused the boys to engage in deeper processing of the material and resulted in increased learning even without increasing self-monitoring. The research on self-regulated learning and feedback supports this explanation. Butler and Winne (1995) claim that, when the target of instruction is domain knowledge, cognitive processing is cued by feedback and, if feedback cues active and elaborate processing of content, then achievement will increase. Thus, the self-assessment done by boys in this study could have caused them to think more about the material at hand and remember more of it on the post-test, even if it did not increase their metacognitive engagement.

This distinction between metacognition and deep processing may also explain how the girls in the treatment group could be significantly more metacognitive than the boys but not perform significantly better than boys on the post-test: Metacognitive engagement may have less of an effect on content learning than deep processing. The differences between the role of these two constructs in self-assessment strikes me as a potentially fruitful avenue of investigation.

Dweck's research may also shed some light on this issue. The girls in my study may have been more eager than the boys were to please the researcher, who was obviously interested in thinking and self-assessment. They may have picked up on subtle and not-so-subtle cues and responded by making more metacognitive and evaluative statements, but not by thinking more carefully about the task or the material than the boys. I anticipated this possibility, and collected evidence of what students thought the study was about at the end of each session. I asked what they would tell some one who asked what they did during the session. If they consistently said they had evaluated their own work or thought aloud, I would have been concerned. Rather, the vast majority of students said they answered questions and classified arthropods, leading me to conclude that the students' perceptions of what I wanted from them would not have a negative effect on the study. A more precise measure of subjects' perceptions of what is expected from them may be advisable in future research.

It is also possible that self-assessment caused all of the students to spend more time with the material, thereby boosting their familiarity with it. Research on the effects of allowing students to control the sequence of lessons or the presentation of feedback during computer-assisted instruction (Steinberg, 1989) has generally found that learners who are granted full control often exit instruction before mastering the material whereas learners provided with feedback are more likely to persist. In this study, being stopped for self-assessment and being asked if they wanted to revise their work often meant that students spent more time working with and thinking about the material. Without such prompts, the control group was likely to stop working sooner than the treatment group.

Limited time on task might account for some portion of the control group's lower pre- to post-test gains, but a time on task explanation alone is insufficient because it does not explain the cognitive mechanisms involved in learning. That is, what is it about increased time on task that leads to more learning? Steinberg's research illustrates that feedback was instrumental in increasing persistence, and Butler and Winne report that feedback promotes deep processing. In combination, these studies suggest that self-assessment caused treatment subjects to engage in deeper processing of the material than control subjects, and the result was more learning.

Part Five: Key Research Findings and Implications for Future Research and Practice

The results of this study can be summarized as follows: There is a positive relationship between self-assessment and metacognition for girls, a positive relationship between self-assessment and learning for both boys and girls, and no clear link between metacognition and learning. Thus, the hypothesis that task-specific, criterion-referenced self-assessment can increase metacognitive engagement was partially supported by the fact that girls who assessed themselves were more metacognitive than girls who did not, and the hypothesis that self-assessment can increase content learning was fully supported by the fact that students who assessed themselves showed higher pre- to post-test gains than students who did not. My model of self-assessment was not supported because metacognition and learning were not consistently related. Other key findings include the positive effect of the rubric on the criteria that treatment students used in their spontaneous self-assessments, and the fact that students who assessed their own work were remarkably willing to revise it.

Although the cognitive mechanisms underlying self-assessment are still in question, its effects on thinking and learning are clear: Self-assessment can increase metacognition in some students and learning in many. This is likely to be welcome news to educators who struggle against a growing curriculum and burgeoning class sizes to provide adequate feedback and instructional support for their students.

Although the applicability of the results of this study to other racial and socio-economic populations is unknown, several implications for both educational practice and future research suggest themselves. For one, it is crucial to ensure that students understand the terms and concepts with which they are expected to assess themselves. The need to create such understandings provides rich teaching opportunities in classrooms, and will often overlap with or precisely match teachers' learning objectives for their students. For this reason, supporting student self-assessment is not likely to be much of a burden in terms of class time.

Research and practice also should take into consideration the need for motivation to improve one's work after assessing it, the possibility of the failure to process information deeply enough to realize a potentially fruitful approach to revision, the complications caused when the task or prompt and the criteria for the task appear to contradict each other, the probability that students will not always see the criteria used in self-assessment as an authority, and students' theories of learning and the types of goals they select as a result of those theories.

Finally, these results have implications for a model of self-assessment that characterizes it as a process of increasing metacognition and self-regulation and, in turn, increasing learning. This study suggests that deep processing of information, not metacognition, is the key to learning, a finding that contradicts prior research on metacognition and begs further investigation. This is perhaps the most important and compelling issue raised by this research.

Appendix A: Scoring Rubric Used by Treatment Subjects

Criteria for Arthropod Classification

1) I read the page about arthropods carefully to make sure I understood it

4 I read carefully and checked my comprehension of the page as I read. I feel that I have a good understanding of the information presented and that I have learned something new.

3 I read carefully and feel I understand the passage.

2 I read the passage but did not check my understanding of it. I feel there are probably gaps in my memory and understanding of the information.

1 I skimmed the page and do not remember or understand most of it.

2) I checked the page about arthropods to make sure my classification system is accurate

4 I reread the page about arthropods to make sure my system is consistent with the information on the page. If I left something out or found errors, I corrected my work in a way that improved its quality.

3 I reread the page to make sure my work is accurate. I corrected errors if I found them.

2 I reread the page but not thoroughly. I missed some errors and failed to correct others.

1 I didn't reread the page to make sure my classification system is accurate. I made little effort to find errors or correct my work.

3) I specified useful categories for the arthropods

4 I created categories based on important physical characteristics of the arthropods. Each arthropod can only fit in one category.

3 I created categories that make me think about important characteristics of arthropods.

2 I created categories that allow me to classify but don't really make me think about the important characteristics of arthropods.

1 I created categories that use only unimportant characteristics of arthropods.

4) I accurately placed the arthropods in the categories

4 I placed each arthropod in the correct category and checked to see that it only fits in that one category.

3 I placed each arthropod in the correct category.

2 I made some mistakes when I placed arthropods in categories.

1 I made many errors when placing arthropods in categories.

5) I described accurate and complete rules for deciding which arthropods go in each category

4 I clearly and completely described the rules for deciding which arthropods go in each category; I described these rules in a way that would allow some one else to put the arthropods in the same categories I did.

3 I clearly described rules for deciding which arthropods go in each category.

2 I described the rules for deciding which arthropods go in each category, but I left things out and created confusion, or I included information about the categories that does not really help put the arthropods in correct categories.

1 I listed rules, but they do not describe the categories.

Adapted from Marzano, Pickering, & McTighe (1993).

Appendix B: Overview of Task Read to All Subjects

Introduction

Imagine a modern day Noah, trying to assemble and organize all the animals of the Earth. In order to keep things orderly, he needs to give each animal a place beside its relatives. He first tries to arrange them according to size, but he's not happy with the results. Placing the ostrich between the kangaroo and the tuna fish just won't do, nor would it seem right to place the hummingbird between the tarantula and the field mouse. They just don't seem to be closely related.

Twentieth-century Noah then decides to separate all the animals into large categories, like animals that live on land and animals that live in water. That doesn't work out so well either, since some shelled animals live in the ocean while others climb trees, and some six-legged animals, like mosquitoes, spend the first part of their lives wriggling in the water and the last part flying in the air.

Noah has a tough job, sorting all these animals in a way that makes sense. We can assume he gives up and just sticks everybody anywhere. The job of sorting animals in a reasonable way is picked up by real zoologists—people who study animals. About 150 years ago, somebody finally classified all animals into two groups: those with backbones (vertebrates, like humans and birds), and those without backbones (invertebrates, like worms and crickets). Zoologists have been arguing about how to divide animals within each group into smaller related groups ever since.

Take, for example, the invertebrate (no backbone) group called Arthropoda. The phylum Arthropoda contains nearly one million known species of animals, including lobsters, spiders, beetles, bees, grasshoppers and centipedes, to name just a few. How should a zoologist group or classify all these species?

Your task is to make up a classification system that makes sense to you. All the information you need is in the passage on the next page. Read it over, ask any questions you have, then use this information to create a classification system, or a reasonable way to group the animals you are given. Be prepared to give your reasons for classifying the animals the way you do.

Appendix C: Passage about Arthropods Read by All Subjects

Arthropods

The phylum Arthropoda consists of more species than all the other phyla put together. Zoologists have identified over 900,000 species, including insects, spiders, ticks, fleas, moths, butterflies, scorpions, crabs, lobsters, crayfish and shrimp, but they estimate that about 6 million species actually exist. If all of the insects—just one part of the arthropod phylum—could all be collected in one place, they would weigh more than all the other land animals combined!

Arthropods are found in all environments, from land to water, from pools of crude oil to the Himalayan Mountains. The word *arthropod* means "jointed leg." All arthropods share certain characteristics: Their bodies are made of several sections, they have several pairs of jointed legs, and they have an external skeleton or shell. The "exoskeleton" acts like a suit of armor to protect the soft body organs. It also keeps the animals from losing too much water and drying out.

Just as there are shared characteristics among arthropods, there are also differences. One difference is the number of legs. For example, lobsters have 5 pairs of legs, spiders have 4 pairs, and centipedes have one pair on most sections of their bodies. Another difference is the number of visible body sections. Crustaceans, such as the lobster, and arachnids, such as the spider, have two visible body sections. Insects, such as ants, have three. Centipedes and millipedes have many body sections.

Besides legs, most arthropods have other structures attached to their bodies. Most of these other structures are jointed, so they can bend and move. All arthropods have jointed mouthparts for feeding. Spiders have two pairs of jointed mouthparts. They sometimes look like legs but they are actually used to hold and chew food. Centipedes, millipedes, crustaceans such as lobsters and shrimp, and insects may have antennae. Spiders have no antennae but do have poison fangs. Most ants have one or two pairs of wings.

Arthropods also differ in how they get oxygen. Most crustaceans live in water. They have gills beneath their exoskeleton. Other arthropods are mostly land dwellers. They have a network of small tubes that run through their bodies. These tubes make up the respiratory system.

References

Anderson, J. R. (1980). *Cognitive psychology and its implications.* San Francisco: W. H. Fremman & Co.

Andre, T., & Thieman, A. (1988). Level of adjunct question, type of feedback, and learning concepts by reading. *Contemporary Educational Psychology, 13,* 296-307.

Armbruster, B., Echols, C., & Brown, A. (1982). The role of metacognition in reading to learn: A developmental perspective. Volta Review, 84(5), 45-56.

Baker, L. (1984). Children's effective use of multiple standards for evaluating their comprehension. *Journal of Educational Psychology, 76,* 588-597.

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

Baron, J. B. (1990). Performance assessment: Blurring the edges among assessment, curriculum, and instruction. In A. Champagne, B. Lovitts & B. Calinger (Eds.), *Assessment in the service of instruction* (pp. 127-148). Washington, DC: American Association for the Advancement of Science.

Bereiter, C., & Bird, M. (1985). Use of thinking aloud in identification and teaching of reading comprehension strategies. *Cognition and Instruction, 2,* 131-156.

Berliner, D. (1990). Helping kids learn how to learn. Instructor, 99(5), 16-17.

Biemiller, A. & Meichenbaum, D. (1992). The nature and nurture of the self-directed learner. Educational Leadership, 75-80.

Boggiano, A. K. & Barrett, M. (1985). Performance and motivational deficits of helplessness: The role of motivational orientations. *Journal of Personality and Social Psychology, 49,* 1753-1761.

Bransford. J. D. (1979). *Human cognition: Learning, understanding and remembering.* Belmont, CA: Wadsworth Publishing Company.

Bronfenbrenner, U. (1967). Response to pressure from peers versus adults among Soviet and American school children. *International Journal of Psychology, 2,* 199-207.

Bronfenbrenner, U. (1970). Reactions to social pressure from adults versus peers among Soviet day school and boarding school pupils in the perspective of an American sample. *Journal of Personality and Social Psychology, 15,* 179-189.

Brown, A.L. (1978). Knowing when, where, and how to remember: A problem of metacognition. In R. Glaser (Ed.), Advances in instructional psychology, Vol. 1. Hillsdale, NJ: Erlbaum.

Brown, A.L. (1980). Metacognitive development and reading. In R. Spiro, B. Bruce & W. Brewer (Eds.), Theoretical issues in reading comprehension. Hillsdale, NJ: Erlbaum.

Brown, A. L. (1987). Metacognition, executive control, self-regulation and other more mysterious mechanisms. In F. Weinert & R. Kluwe (Eds.), Metacognition, motivation, and understanding (pp. 65-116). Hillsdale, NJ: Erlbaum.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex intervention in classroom settings. The Journal of the Learning Sciences, 2(2), 141-178.

Brown A. L. & Palincsar, A. S. (1982). Inducing strategic learning from texts by means of informed, self-control training. *Topics in Learning and Learning Disabilities*, *2*(1), 1-17.

Bruner, J. S. (1978). *Human growth and development.* Oxford: Clarendon Press.

Butler, D. & Winne, P. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 65(3), 245-281.

Carver, C. & Scheier, M. (1990). Origins and functions of positive and negative affect: A control-process view. *Psychological Review*, *97*, 19 - 35.

CAT/5 Spring Norms Book (1993). Monterey, CA: CTB, a division of the Macmillan/McGraw-Hill School Publishing Company.

Chinn, C. & Brewer, W. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research*, **63**, 1 -49.

Costa, A. L. (1991). The school as a home for the mind: A climate for thinking. In R. Mulcahy, R. Short & J. Andrews (Eds.), *Enhancing learning and thinking*. New York: Preager.

Daiute, C., & Dalton, (1988). "Let's brighten it up a bit:" Communication and cognition in writing. In B. Rafoth & D. Rubin (Eds.), *The social construction of written communication*. Norwood, NJ: Ablex.

Daiute, C., & Kruidenier, J. (1985). A self-questioning strategy to increase young writers' revising processes. Applied Psycholinguistics, 6, 307-318.

Davidson, L., & Scripp, L. (1990). Tracing reflective thinking in the performance ensemble. *The Quarterly, 1*(1&2), 49-62.

Davidson, L., Myford, C., Plasket, D., Scripp, L., Swinton, S., Torff, B., & Waanders, J. (1992). *Arts PROPEL: A handbook for music*. Cambridge, MA: Harvard Project Zero.

Derry, S., & Murphy, D. (1986). Designing systems that train learning ability: From theory to practice. *Review of Educational Research, 56*(1), 1-39.

Deci, E., & Ryan, R. (1980). The empirical exploration of intrinsic motivational processes. In L. Berkowitz (Ed.), *Advances in experimental social psychology: Vol. 13*. New York: Academic Press.

Dweck, C. (1975). The role of expectations and attributions in the alleviation of learned helplessness. *Journal of Personality and Social Psychology, 31*, 674-685.

Dweck, C. & Bempechat, J. (1983). Children's theories of intelligence: Consequences for learning. In S. Paris, G. Olson, and H. Stevenson (Eds.), *Learning and motivation in the classroom.* Hillsdale, NJ: Lawrence Erlbaum.

Dweck, C. & Bush, E. (1976). Sex differences in learned helplessness: 1. Differential debilitation with peer and adult evaluators. *Developmental Psychology, 12*, 147-156.

Dweck, C., Davidson, W., Nelson, S., & Enna, B. (1978). Sex differences in learned helplessness: 2. Contingencies of evaluative feedback in the classroom and 3. An experimental analysis. *Developmental Psychology, 14*(3), 268-276.

Ericsson, K. A. and Simon, H. A. (1984). *Protocol analysis: Verbal reports as data.* Cambridge, MA: The MIT Press.

Flavell, J. H. (1977). *Cognitive development*. Englewood Cliffs, NJ: Prentice-Hall.

Flavell, J. H. (1981). Cognitive monitoring. In P. Dickson (Ed.), *Children's oral communication skills*. New York: Academic Press.

Flavell, J. H. (1987). Speculations about the nature and development of metacognition. In F. Weinert & R. Kluwe (Eds.), Metacognition, motivation, and understanding (pp. 21-30). Hillsdale, NJ: Erlbaum.

Flower, L. A., & Hayes, J. R. (1981). The pregnant pause: An inquiry into the nature of planning. Research in the Teaching of English, 15(3), 229-243.

French, C., & French, F. (1991). Evaluating programs that claim to teach thinking skills: Critical issues and options. In R. Mulcahy, R. Short & J. Andrews (Eds.), *Enhancing learning and thinking*. New York: Preager.

Gardner, H. (1991). Assessment in context: The alternative to standardized testing. In B. R. Gifford and M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction.* Boston: Kluwer.

Goodrich, H. (1993). Self-assessment: An intersection between authentic assessment and metacognition . Qualifying paper. Cambridge, MA: Harvard Graduate School of Education.

Haney, W. (1991). We must take care: Fitting assessments to functions. In V. Perrone (Ed.), *Expanding student assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Hawkins, J., Frederikson, J., Collins, A., Bennet, D., & Collins, E. (1993). Assessment and technology: New ways to assess student learning. *Technology and Education, 36*(5).

Heath, D. C. (1994). *Heath mathematics connections: Assessments level 4*. Lexington, MA: D. C. Heath.

Herman, J., Aschbacher, P., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Herman, J., & Winters, L. (1994). Portfolio research: A slim collection. *Educational Leadership, 52*(2), 48-55.

Higgins, K., Harris, N., & Kuehn, L. (1994). Placing assessment into the hands of young children: A study of student-generated criteria and self-assessment. *Educational Assessment, 2*(4), 309-324.

Hollander, E., & Marcia, J. (1970). Parental determinants of peer orientation and self-orientation among preadolescents. *Developmental Psychology, 2,* 292-302.

Kusnic, E., & Finley, M. (1993). Student self-evaluation: An introduction and rationale. In J. MacGregor (Ed.), *Student self-evaluation: Fostering reflective learning*. San Francisco: Jossey-Bass.

Mabe, P., & West, S. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67*(3), 280-296.

MacGregor, J. (1993). *Student self-evaluation: Fostering reflective learning*. San Francisco: Jossey-Bass Publishers.

Mancini, G., Mulcahy, R., Short, R., & Cho, S. (1991). Teaching cognitive strategies in the classroom: A content-based instructional model. In R. Mulcahy, R. Short & J. Andrews (Eds.), Enhancing learning and thinking (pp. 67-96). New York: Preager.

Mancini, G., Short, R., Mulcahy, R. & Andrews, J. (1991). Social cognition and social competence: Instructional issues. In R. Mulcahy, R. Short & J. Andrews (Eds.), Enhancing learning and thinking (pp.193-206). New York: Preager.

Markman, E. (1981a). Comprehension monitoring: Developmental and educational issues. In W. P. Dickson (Ed.), Children's oral communication skills. New York: Academic Press.

Marzano, R., Pickering, D. & McTighe, J. (1993). *Assessing student outcomes: Performance assessment using the dimensions of learning model*. Alexandria, VA: ASCD.

Massachusetts Department of Education (1995). *Mathematics curriculum content chapter: Achieving mathematical power*. Malden, MA: Massachusetts DOE.

Mead, G. H. (1934). *Mind, self and society*. Chicago: University of Chicago Press.

Nickerson, R., Perkins, D., & Smith, E. (1985). The teaching of thinking. Hillsdale, NJ: Erlbaum.

Nisbet, J. & Shucksmith, J. (1986). Learning strategies. Boston: Routledge & Kegan Paul.

Nolen, S. (1988). Reasons for studying: Motivation orientations and study strategies. *Cognition and Instruction*, **5**(4), 269 - 287.

Nold, E. W. (1982). Revising. In C. Frederiksen, M. Whiteman, and J. Dominic (Eds.), *Writing: The nature, development and teaching of written communication*. Hillsdale, NJ: Erlbaum.

Oskarsson, M. (1984). *Self-assessment of foreign language skills: A survey of research and development work*. Strasbourg: Council for Cultural Co-operation.

Pace, S. F. (1991). Government processes for including thinking skills in the curriculum. In R. Mulcahy, R. Short & J. Andrews (Eds.), Enhancing learning and thinking (pp. 21-34). New York: Preager.

Palincsar, A., & Brown, A. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. Cognition and Instruction, 1(2), 117-175.

Palincsar, A., & Brown, A. (1986). Interactive teaching to promote independent learning from text. The Reading Teacher, 39(8), 771-777.

Palincsar, A., & Brown, A. (1988). Teaching and practicing thinking skills to promote comprehension in the context of group problem solving. Remedial and Special Education, 9(1), 53-59.

Paris, S. G., & Ayres, L. R. (1994). *Becoming reflective students and teachers with portfolios and authentic assessment.* Washington, DC: American Psychological Association.

Perkins, D. (1981). *The mind's best work.* Cambridge, MA: Harvard University Press.

Perkins, D. (1987). Thinking frames: An integrative perspective on teaching cognitive skills. In J. Baron & R. Sternberg (Eds.), *Teaching thinking skills: Theory and practice.* New York: Freeman.

Perkins, D. (1992). *Smart schools: From training memories to educating minds.* New York: The Free Press.

Perkins, D., Goodrich, H., Tishman, S., & Mirman Owen, J. (1993). *Thinking connections: Learning to think and thinking to learn.* Reading, MA: Addison-Wesley.

Perkins, D., Jay, E., & Tishman, S. (1993). Beyond abilities: A dispositional theory of thinking. *The Merrill Palmer Quarterly, 39*(1), 1-21.

Perrone, V. (1991). *Expanding student assessment.* Alexandria, VA; Association for Supervision and Curriculum Development.

Pressley, M., & Levin, J. (Eds.) (1983). *Cognitive strategy research: Psychological foundations.* New York: Springer-Verlag.

Price, M. A. (1991). Teaching thinking to preschoolers. In R. Mulcahy, R. Short & J. Andrews (Eds.), Enhancing learning and thinking (pp. 53-66). New York: Preager.

Reading/Language in Secondary Schools Subcommittee of International Reading Assn. (1990). Thinkers and readers: Secondary perspectives. *Journal of Reading, 33*(6), 460-463.

Reichel, A. G. (1994). Performance assessment: Five practical approaches. *Science and Children*, October, 21-25.

Roberts, T. & Nolen-Hoeksema, S. (1989). Sex differences in reactions to evaluative feedback. *Sex Roles, 21*(11/12), 725-746.

Satterly, D. (1989). *Assessment in schools*. New York: Basil Blackwell.

Scardamalia, M., & Bereiter, C. (1983). The development of evaluative, diagnostic and remedial capabilities in children's composing. In M. Martlew (Ed.), *The psychology of written language: A developmental approach.* London: John Wiley.

Scardamalia, M., & Bereiter, C. (1985). Fostering the development of self-regulation in children's knowledge processing. In S. F. Chipman, J. W. Segal, & R. Glaser (Eds.), Thinking and learning skills, Vol. 2: Research and open questions (pp. 563-577). Hillsdale, NJ: Lawrence Erlbaum.

Scardamalia, M., Bereiter, C. & Steinbach, R. (1984). Teachability of reflective processes in written composition. Cognitive Science, 8,173-190.

Schmar, E. (1995). Student self-assessment. In T. Azwell & E. Schmar (Eds.), *Report card on report cards: Alternatives to consider*. Portsmouth, NH: Heinemann.

Schoenfeld, A. H. (1983). Problem solving in the mathematics curriculum: A report, recommendations, and an annotated bibliography. Washington, DC: Mathematical Association of America.

Schoenfeld, A. H. (1987). What's all the fuss about metacognition? In A. Schoenfeld (Ed.), Cognitive science and mathematics education. Hillsdale, NJ: Erlbaum.

Schutz, P. (1993). Additional influences on response certitude and feedback requests. *Contemporary Educational Psychology*, **18**, 427 - 441.

Stamps, A. (1989). A new method for teaching architectural studios: Use of objective feedback to help design students refine their capacity for self-evaluation. *Perceptual and Motor Skills, 69*, 827-832.

Steinberg, E. (1989). Cognition and learner control: A literature review, 1977-1988. *Journal of Computer Based Instruction, 16*, 117-121.

Sternberg, R. (1986). Cognition and instruction: Why the marriage sometimes ends in divorce. In R. Dillon & F. Sternberg (Eds.), *Cognition and instruction*. Orlando, FL: Academic.

Stiggins, R. J. (1994). *Student-centered classroom assessment*. New York: Merrill.

Towler, L., & Broadfoot, P. (1992). Self-assessment in the Primary school. *Educational Review*, 44(2).

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Walczyk, J., & Hall, V. (1989). Effects of examples and embedded questions on the accuracy of comprehension self-assessments. *Journal of Educational Psychology, 81*, 435-437.

Walters, J., Seidel, S., and Gardner, H. (1994). Children as reflective practitioners: Bringing metacognition to the classroom. In C. Collins-Block, and J. Mangieri (Eds.) *Creating powerful thinking in teachers and students: Diverse perspectives*. Orlando, FL: Holt, Rinehart & Winston.

Waluconis, C. J. (1993). Self-evaluation: Settings and uses. In J. MacGregor (Ed.), *Student self-evaluation: Fostering reflective learning*. San Francisco: Jossey-Bass.

Watkins, J., Cox, P., Mirman Owen, J., & Burkhart, G. (1992). Designing assessments for learner- and learning-centered schools. Draft of document in preparation.

Watts, K. H. (1996). Bridges freeze before roads. In T. Guskey (Ed.), *ASCD yearbook 1996: Communicating student learning*. Alexandria, VA: Association for Supervision and Curriculum Development.

Weinert, F. E. (1987). Introduction and overview: Metacognition and motivation as determinants of effective learning and understanding. In F. Weinert & R. Kluwe (Eds.), Metacognition, motivation, and understanding (pp. 1-16). Hillsdale, NJ: Erlbaum.

Weinstein, C. (1994). Strategic learning/strategic teaching: Flip sides of a coin. In P. Pintrich, D. Brown, & C. Weinstein (Eds.), Student motivation, cognition and learning. Hillsdale, NJ: Erlbaum.

Wellman, H. M. (1983). Metamemory revisited. In M. Chi (Ed.), Trends in memory development. Basil, Switzerland: S. Karger.

Wiggins, G. (1989b). Teaching to the (authentic) test. *Educational Leadership, 46*(7), 41-47.

Wiggins, G. (1989a). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*(9), 703-713.

Wiggins, G. (1990). Workshop on authentic assessment. Unpublished materials. Rochester, NY: Consultants on Learning, Assessment, and School Structure.

Winne, P. (1991). Motivation and teaching. In H. Waxman and J. Walberg (Eds.), *Effective teaching: Current research.* Berkeley, CA: McCutchan.

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17*, 31-74.

Wolf, D., & Pistone, N. (1991). *Taking full measure: Rethinking assessment through the arts*. New York: College Board Publications.

Yussen, S. (1983). The role of metacognition in contemporary theories of cognitive development. In D. Forrest-Pressley, G. MacKinnon, & T. Waller (Eds.), Metacognition, cognition and human performance. Vol. 1: Theoretical perspectives (pp. 253-283). New York: Academic Press.

Zessoules, R., & Gardner, H. (1991). Authentic assessment: Beyond the buzzword and into the classroom. In V. Perrone (Ed.), *Expanding student assessment*. Alexandria, VA: ASCD.

TM029863

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Student Self-Assessment: At the Intersection of Metacognition and Authentic Assessment	
Author(s): Heidi Goodrich Andrade	
Corporate Source: Harvard Graduate School of Education	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the b of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	Level 2A	Level 2B
[X]	[ ]	[ ]
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this docum as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its syst contractors requires permission from the copyright holder. Exception is made for non-profit reproductio n by libraries and other service agenc to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature: Heidi Andrade	Printed Name/Position/Title: Heidi G. Andrade, Ed.D.
Organization/Address: Harvard Grad School of Educ., Longfellow Hall - Project Zero, Cambridge MA 02138	Telephone: 617-496-1813 FAX: 617-496-4288
	E-Mail Address: heidi@pz.harvard.edu Date: 5/11/99

(ov

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:	
Address:	
Price:	

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:	
Address:	

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

The Catholic University of America
ERIC Clearinghouse on Assessment and Evaluation
210 O'Boyle Hall
Washington, DC 20064
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

(Rev. 9/97)