JSGE
The Journal of Secondary Gifted Education
Vol. XVI, No. 2/3, Winter/Spring 2005, pp. 119–133.

Development of Authentic Assessments for the Middle School Classroom

Tonya R. Moon, Catherine M. Brighton, Carolyn M. Callahan
University of Virginia

Ann Robinson
University of Arkansas

This article discusses the rationale for, and explicates the process used in, developing differentiated authentic assessments for middle school classrooms (many of which contain gifted students) that are aligned with state academic standards. The assessments were developed based on learner-centered psychological principles and revised based on a content validation study involving a panel of 46 experts representing a variety of educational professionals. In addition to the content validation study of the assessments, interrater reliability estimates based on Kappa were calculated using student responses to the assessments in classrooms in two states. Results provide evidence that these types of assessments can provide quantifiable information about student learning, as well as inform the instructional process.

From today's understanding of cognitive science, students are not viewed as recorders of factual information, but rather as creators of their own unique knowledge structures. As such, meaningful learning is viewed as being reflective, constructive, and self-regulated (Gordon, 1992). Thus, learning that strongly emphasizes drill and practice on discrete, unconnected, or isolated factual knowledge is a tremendous disservice to students, including those who are academically talented (Moon, Brighton, & Callahan, 2003).

While use of high-stakes testing has focused teacher planning on specified, agreed-upon state-level standards, exclusive use of traditional assessments—often in the form of pencil-and-paper multiple-choice tests—have been judged to be a negative in the middle school classroom (Archbald, 1991; Dana & Tippins, 1993; Kennedy, 1996). Critics of these traditional forms of assessment argue that "standardized, multiple-choice tests have definite limitations, are overused and overinterpreted, and are unlikely to help schools achieve the reform goals" (Archbald, p.1). While best practices in the middle school include teaching conceptually and assessing student understanding of concepts, traditional standardized tests fail to do so. Cheek (1993) argued that traditional test items that examine core understanding of disciplines are often discarded because they fail to discriminate among test takers. Rather, questions that deal with peripheral details or subskills do a better job of discriminating among students and are therefore the questions selected for inclusion on traditional standardized tests.

Others maintain that traditional assessments are incompatible with the genuine knowledge, skills, and dispositions of disciplines (Cheek, 1993; Dana & Tippins, 1993; Gordon & Bonilla-Bowman, 1996). Further, Dana and Tippins have argued that these traditional assessments cannot test the extent to which a student has mastered a body of knowledge surrounding a concept, only the information tested in the selected items, nor can they provide rich information about the multifaceted thinking necessary for complex problem solving. Resnick (1987) described the imbalance between how intellective work is conducted in school and in real life: "In real life one actu-

ally engages in performances that contribute to the solution of real problems, rather than producing, on demand and in artificial situations, symbolic samples of one's repertoire of developed abilities."

Furthermore, traditional assessments in the middle school ignore the diverse needs of the learners in that setting. Traditional testing requires passive involvement with the subject material and thus is inconsistent with the developmental needs of young adolescents (Dana & Tippins, 1993). However, authentic assessments have been shown to be relevant to curricula for high-ability students (VanTassel-Baska, Bass, Ries, Poland, & Avery, 1998), as well as curricula that focus on higher level thinking (VanTassel-Baska, Zuo, Avery, & Little, 2002). In addition, authentic assessments are viewed by some in the field of gifted education as a more valid measure of student learning (Baldwin, 1994; Callahan, Tomlinson, Moon, Tomchin, & Plucker, 1995; Clausen, Middleton, & Connell, 1994). In short, traditional assessment is increasingly being viewed as insensitive to differences among learners and asynchronous with optimal learning conditions (Gordon & Bonilla-Bowman, 1996; Kennedy, 1996).

In response to these criticisms of the traditional assessment paradigm, some measurement experts have begun advocating for the use of authentic, or performance, assessment. "Performance measures have the potential for increased validity because the performance tasks are themselves demonstrations of important learning goals rather than indirect indicators of achievement" (Resnick & Resnick, 1992).

Characteristics of Authentic Assessment

Authentic assessments, often called performance-based assessments, engage students in real-world tasks and scenario-based problem solving more than traditional measures such as multiple-choice pencil-and-paper tests (Darling-Hammond, 1997). Performance-based tasks are largely open-ended and often can be answered using multiple approaches (Reed, 1993). For maximum benefit, these tasks should be relevant and meaningful to students (Henderson & Karr-Kidwell, 1998). Authentic assessment can take the form of performances, projects, writings, demonstrations, debates, simulations, presentations, or other sorts of open-ended tasks (Cheek, 1993; Dana & Tippins, 1993; Reed). While authentic assessment is highly contextual, exemplary authentic assessments always allow students to demonstrate knowledge and skills that are worth knowing (Dana & Tippins). Specifically, they:

1. are focused on content that is essential, focusing on the big ideas or concepts, rather than trivial microfacts or specialized skills;
2. are in-depth in that they lead to other problems and questions;
3. are feasible and can be done easily and safely within a school and classroom;
4. focus on the ability to produce a quality product or performance, rather than a single right answer;
5. promote the development and display of student strengths and expertise (the focus is on what the student knows);
6. have criteria that are known, understood, and negotiated between the teacher and student before the assessment begins;
7. provide multiple ways in which students can demonstrate they have met the criteria, allowing multiple points of view and multiple interpretations;
8. require scoring that focuses on the essence of the task and not what is easiest to score (p. 4).

Rationale for Differentiated Authentic Assessment in the Middle School

While many educators advocate for authentic assessment for all students, the middle school environment and the particular needs of middle school students suggest particular reasons why this approach is well suited. For example, the Carnegie Council on Adolescent Development (CCAD, 1990) calls for schools to

1. create small communities for learning where stable, close, mutually respectful relations with adults and peers are considered fundamental for intellectual development and personal growth;
2. teach a core academic program that results in students who are literate, including in the sciences, and who know how to think critically, lead a healthy life, behave ethically, and assume the responsibilities of citizenship in a pluralistic society;
3. ensure success for all students through the elimination of tracking by achievement level and promotion of cooperative learning and flexible grouping; and
4. connect schools with communities that together share responsibility for each middle grade student's success through identifying service opportunities in the community, establishing partnerships and collaborations to ensure students' access to health and social services, and opportunities for constructive after-school activities (p. 9).

This call for action from the Carnegie Council (1990)

is consistent with the implementation of authentic assessment in the middle school. One option for using authentic assessment is to allow middle school students to work on tasks of value to a particular community, yielding a truer audience for authentic feedback. Hence, this approach to assessment may use community resources to enrich the learning experience as recommended by the Carnegie Council (Kennedy, 1996).

Authentic assessment may also improve teaching and learning in the middle school by preserving the integrated, complex nature of learning. In this approach, students recall learned information and utilize needed skills, but do so in the context of an authentic scenario requiring the production of new ideas in particular contexts and for particular purposes. This process of problem solving and solution finding requires and fosters a deep understanding of the discipline, as well as integration of knowledge and skills across disciplines, a basic tenet of curriculum construction in the middle school (Archbald, 1991; Tomlinson, 2001).

The National Middle School Association (NMSA) further advocates for heterogeneous groupings of students in the middle school setting, suggesting that grouping inflexibly using tracking unfairly segregates students (Jackson, & Davis, 2000; NMSA, 1995). Given this content, teachers in heterogeneously grouped middle school classrooms are placed in situations that demand appropriate, but varying degrees of challenge and support for all students. With the use of authentic assessments, students view the learning process as important and linked to skills used in the real world (Lines, 1994). The premise underlying authentic assessment is that teachers create curricular experiences targeting specific performance skills and, as a result, gain richer instructional information about students that is useful for modifying instruction for their varied needs (Darling-Hammond, 1997).

Authentic assessment may also have the potential to narrow the performance gap among various cultures and therefore be more equitable in the assessment of differnt cultural groups, another goal of the middle school movement (Egan & Gardner, 1992; Gordon & Bonilla-Bowman, 1996). The cultural performance gap seems to narrow when students are engaged in activities that provide various linguistic interpretation options, use materials familiar to the students, and build on engaging problem-solving tasks (Gardner, 1993).

While enormous amounts of money, time, and energy are placed on developing assessments for use as accountability measures, little emphasis has been placed on working with classroom teachers to develop assessments that provide reliable and valid information. In particular, there has been little emphasis on standards-based authentic assessments for purposes of documenting student learning and informing the instructional process. In order for authentic assessments to document student learning and inform instruction, attention must be given to the reliability and validity of scores obtained from their implementation.

According to Messick (1994), an important distinction in reliability and validity criteria exists between assessments used for educational accountability and authentic assessments. Authentic assessments should be evaluated by criteria that differ in emphasis, rather than kind. That is, because authentic assessments and traditional assessments emphasize different aspects of student learning, to judge each by the same criteria would be inappropriate. To understand these differences in emphasis, consider the following development process of both types of assessments. Standardized instruments used for educational accountability employ standardized procedures for administering tests, with tests and test items being secure. Authentic assessments, on the other hand, present to students upfront what is being assessed and the standards or criteria that constitute differing levels of performance (e.g., expert to novice). Because of this different emphasis, evidence of authentic assessments' reliability is provided by examining the scoring rubric, the mechanism that provides students with ways for improving performance (Messick).

The differing emphasis between standardized and authentic assessments also has implications for validity criteria. In standardized assessment contexts, evidence of validity is provided several ways, with the most frequent method being correlational. That is, a new instrument is typically correlated with an established, widely accepted instrument measuring the same construct. If the correlation is high, evidence of validity is supported. Authentic assessments for validity purposes, on the other hand, should provide evidence that students are emulating intellectual challenges faced by practicing professionals (Jamentz, 1994). Therefore, to be in accordance with theses recommendations, it was critical in this article to examine the content validity of the assessments, as well as the interrater reliability.

"Differentiated instruction" is a term used to describe a teacher's purposeful instructional responses to students' academic diversity and other differences in readiness, interests, and learning profiles pertinent to their learning (Tomlinson, 1995, 2001). The philosophy of differentiation is not only applicable to the instructional sequence, but also to the assessment process and is particularly well suited for authentic assessment—hence the term "differentiated authentic assessment."

The information and materials that follow are the result of the 5-year research effort at the University of Virginia's National Research Center on the Gifted and Talented (NRC/GT) to develop differentiated authentic assessments for middle school classrooms in the content areas of English/language arts, social studies/history, mathematics, and science. The differentiated authentic assessments were developed in alignment with several learner-centered psychological principles that are based on more than a century of research on teaching and learning (Alexander & Murphy, 1994). The specific assumptions that served as a framework for the assessments included the following:

- Learning is a process of discovering and constructing meaning from information and experience.
- The learner seeks to create meaningful, coherent representations of knowledge.
- The learner links new information with existing and future-oriented knowledge.
- Higher order strategies facilitate creative and critical thinking and the development of expertise.
- Curiosity, creativity, and higher order thinking are stimulated by relevant, authentic learning tasks of optimal difficulty and novelty for each student.
- Although basic principles of learning, motivation, and effective instruction apply to all learners, learners have different capabilities and preferences for learning mode and strategies.

The goal of this project was to design differentiated authentic assessments that promote meaningful learning in authentic situations aligned with the curriculum and instruction of middle school classrooms across the country. The purpose of this article is to describe the development of the differentiated authentic assessments and to provide information on the consistency with which classroom teachers score this type of assessment. This is one facet of educational measurement that has yet to be studied in depth (Marzano, 2002).

Development of Middle School Differentiated Authentic Assessments

Several basic principles guided the development phase of each task. First and foremost, NRC/GT staff focused on creating assessments that embodied key concepts, principles, generalizations, and processes critical to understandings in the discipline(s). Because of this focus, content standards from state and national frameworks that were reflective of understandings and applications of big ideas and core themes of the disciplines were the primary assessment targets, although processes and dispositions were included at times. For the standards that each task was designed to assess, see Appendix 1. A sample assessment can be found in Appendix 2.

Another criterion applied in the development process was that each assessment reflected current understandings or best practices in the areas of motivation, cognition, learning theory, and instruction. To actively engage students in their own learning, tasks were designed around real-life situations and required students to make connections and forge relationships between prior knowledge and skills. In addition, tasks allowed multiple pathways to solutions, allowed for a diversity of perspectives in solutions, or both.

Promotion of effective problem solving was another criterion of task development. Therefore, tasks were designed in general to require sustained work on the part of the students and at the same time allow them to have some degree of control or choice over the actions needed. In some instances, students were given the responsibility of designing and carrying out their own investigations.

Tasks were also developed to provide sufficient challenge for the range of academic diversity in the heterogeneous middle school classroom. Using the work of Tomlinson (1995, 2001), the assessments were differentiated using "The Equalizer." Beginning with the presumption that all students' tasks must relate to the same essential skills and objectives, a core on-grade-level task was designed around the specific standards to be assessed, and then modifications were made to reflect advanced understanding of the major concepts, principles, generalizations, and skills for more advanced learners or to provide the scaffolding necessary to guide struggling learners in completing the task successfully. Examples of the type of task differentiation that was done for struggling learners included more structured task context (solutions, decisions, etc.), tasks based on only single facets (applications, approaches, etc.), and less independence in planning, designing, or monitoring. In contrast, tasks for advanced students required depth and complexity of content understanding, were less structured, required integration of multiple facets of a discipline or across disciplines, and allowed for greater independence. Regardless of level, all students' tasks related to the same essential skills and objectives.

Clear communication of student responsibilities and requirements was also a critical component of task development. In order to assess what students knew, understood, and were able to do, clear delineation of student roles and responsibilities and clearly defined performance criteria in the scoring rubric were part of each assessment task, with only subtle variations across the varied levels of the task.

Scoring Rubrics

Rubrics were designed to yield information about students' strengths and weaknesses relative to the content and processes being assessed. To provide teachers with rich, detailed instructional information, rubrics were designed for analytic scoring, where students' performances on specific task elements (domains) were assessed, with the overall performance on the assessment being the summation of the domains.

The development of each scoring rubric began with revisiting the purpose and the objectives or standards that the authentic assessment was designed to assess. After reviewing the purpose(s) of the assessment, elements of the performance to be evaluated were identified (domains). Characteristics or criteria were identified that determined each score point for each domain. These score points translated into levels of performance ranging from novice to expert. This process was repeated for each domain that was identified for the assessment.

Data Collection

Study Classrooms

All classrooms were located in states that had a state-testing program based predominantly on traditional assessments. Each of the classrooms was heterogeneously grouped and included students who were performing below grade level and on grade level, as well as students who had been identified as gifted by the respective district's identification process. Teachers volunteered to administer the assessments because the assessments were aligned with a unit of study they were teaching. In one case, the classroom teacher only administered the assessment (Fables and Folktales) to identified gifted students, and in another case the teacher made the decision to give the on-grade-level assessment to all students, not differentiating among the various ability levels of students in the class.

Fables and Folktales was completed by seventh-grade students who had been identified as gifted by the school district; Wall Street Decisions was completed by two seventh-grade, mixed-ability classrooms; You Can't Convince Me was completed by one seventh-grade, mixed-ability classroom; Creature Classifications was completed by one third-grade classroom and one seventh-grade classroom, both of which were mixed ability. Student work examples provided the data for assessing interrater reliability.

Psychometric Attributes of the Authentic Assessments

The numbers of performance assessments that were implemented by teachers and, consequently, student samples and data on outcomes of the assessment process were limited. However, the process did provide a rich source of information of how samples of students perform on practical, authentic classroom assessments. The following section describes those characteristics that could be assessed with the tasks.

Content validity. Once the development of the assessments and associated rubrics was complete, expert reviewers were solicited to participate in a content validation of the tasks. Content validation is a rational analysis based upon individual, subjective judgment (Allen & Yen, 2002). A total of 46 individuals reviewed the assessment tasks. Individuals reviewed only those tasks that were in content areas with which they were familiar. Nineteen panelists were gifted education specialists or curriculum coordinators in school districts, 18 were state department of education officials, 5 were middle school teachers, and 4 were university professors.

Content validation by this panel of experts was carried out to ascertain the degree to which each assessment addressed the learning objectives that it was intended to measure, as well as the extent to which the assessment was relevant and applicable to the world outside of school. Specifically, panelists were provided a structured framework that was used to assess the degree of relevance and representativeness of each task's content and the response process. Also as part of this process, panelists were asked to analyze critically each assessment for potential biases against students from economically disadvantaged environments, differing cultural/ethnic groups, and gender groups.

Modifications to the tasks and rubrics were made based on the assimilation of reviewers' comments, which typically involved replacing "adult lingo" with more "student-friendly" language in the rubrics. In no cases did reviewers suggest major conceptual flaws with the task or criteria defining the various levels of performance. Future reviews should include practicing professionals whose areas of expertise align with the focus of the assessments and who might provide additional suggestions regarding the applicable of the assessments to the skills required in the world of work.

Interrater reliability. In evaluating scores involving raters, it is important to know the extent to which different scorers agree (or disagree) on the values assigned to student responses. Interrater reliability is the degree to which two raters agree on the level of student performance. One way

Table 1

Independent Ratings (Teacher & NRC) for Fables and Folktales

Domain**	Kappa	% Exact of Agreement	% Adjacent of Agreement	Teacher $\overline{X}$ Rating	NRC $\overline{X}$ Rating
Purpose	.37	50	25	2.6 (.74)	1.9 (.83)
Sequencing	–*	38	38	3.0 (.00)	2.1 (.83)
Symbolism	–*	38	13	3.0 (.00)	1.9 (.99)
Word Usage	.60	75	25	2.6 (.52)	2.4 (.52)
Expressiveness	.49	63	25	2.0 (.53)	2.5 (.53)

Note. * could not be computed because domain ratings were constant
** each domain's scale range = 1 to 3
*** numbers in parentheses represent standard deviations

to compute an index of agreement between raters is with the Kappa coefficient. Kappa is the proportion of agreements after chance agreement between raters has been excluded (see Kraemer, 1982) and is used with categorical data. Using SPSS for Windows 11.5, Kappa coefficient was computed through the Crosstabs subroutine.

Kappa was computed on the five assessment tasks—Fables and Folktales, Wall Street Decisions, You Can't Convince Me, Where in the World, and Creature Classification—completed by students from six different classrooms. Because students were from different classrooms, the numbers of students completing each assessment varied. The student examples provided data for assessing interrater reliability. For Fables and Folktales and one set of students completing Wall Street Decisions, the classroom teacher and an NRC/GT staff member served as the two raters; two NRC/GT staff members were the raters for You Can't Convince Me, Creature Classification, and Where in the World, as well as one classrooms' products from Wall Street Decisions. Tables 1–9 display the reliability results for each assessment.

Fables and Folktales. This assessment task invites students to develop an original fable or folktale within the context of a storytelling festival in the year 2060. Students are assessed across six domains: purpose, sequencing, symbolism, word usage, expressiveness, and timeliness.

Eight student responses to this assessment were evaluated. Table 1 indicates that the interrater reliability of the domains ranged from 0.37 to 0.60, with exact agreement on the ratings between the teacher and NRC/GT staff ranging from 38% to 75%. The word usage domain had the greatest exact agreement (75%) and also the highest reliability coefficient (0.60). Kappa could not be computed for two domains, sequencing and symbolism, because ratings within each set of raters did not vary (i.e., no variation within teachers' ratings or no variation within NRC/GT staff ratings). Using guidelines suggested by Landis and Koch (1977), the rater reliability estimates ranged from fair (.37) to moderate (.60).

Wall Street Decisions. This assessment assesses the degree to which students understand and can apply mathematical concepts and calculations such as estimation; rate of change; and percent, decimal, and fraction conversions to make decisions about stock purchases and to explain changes in the stock market. There are three levels of the task: one designed for struggling learners, one designed for on-grade-level learners, and one designed for students above grade level in mathematical understanding. All students are assessed using the same rubric in the four domains of support for conclusions, strategy and calculations, supporting materials, justification, and presentation.

Four student responses to Task 1 of the assessment were evaluated. Table 2 indicates that the interrater reliability of the domains for Prompt 1 (struggling learners) ranged from 0.41 to 1.0, with exact agreement on the ratings between the teacher and NRC/GT staff ranging from 25% to 100%. The support for the conclusions domain had the highest exact agreement rate (100%) and the highest reliability coefficient (1.0). Based on guidelines provided by Landis and Koch (1977), estimates of rater reliability were in the substantial range (0.60–0.80) in all domains except the domains of supporting materials and presentation, which were moderate (0.40–.0.59).

Seventeen student responses were evaluated for Task 2. For this task (on-grade level learners), the interrater reliability of the domains ranged from 0.53 to 0.86, with the supporting materials and justification domains having the highest exact agreement rate (71%). As estimates of rater reliability using previously indicated guidelines, the Kappa

Table 2

Independent Ratings (Teacher & NRC) for Wall Street Decisions: Task 1*

Domain**	Kappa	% Exact of Agreement	% Adjacent of Agreement	Teacher $\overline{X}$ Rating	NRC $\overline{X}$ Rating
Support for Conclusions	1.0	100	–	3.3 (1.5)	3.3 (1.5)
Strategy and Calculations	0.78	75	25	2.8 (1.3)	2.5 (1.0)
Supporting Materials	0.47	25	25	3.5 (1.0)	3.0 (1.4)
Justification	0.87	25	50	3.3 (1.5)	2.3 (.96)
Presentation	0.41	50	25	3.0 (1.4)	3.5 (.58)

Note. *designed for struggling learners
** each domain's scale range = 1 to 4
*** numbers in parentheses represent standard deviations

Table 3

Independent Ratings (Teacher & NRC) for Wall Street Decisions: Task 2*

Domain**	Kappa	% Exact of Agreement	% Adjacent of Agreement	Teacher $\overline{X}$ Rating	NRC $\overline{X}$ Rating
Support for Conclusions	0.69	57	29	3.1 (1.1)	2.6 (1.1)
Strategy and Calculations	0.71	57	29	2.6 (1.1)	2.0 (1.0)
Supporting Materials	0.86	71	14	2.7 (1.4)	2.7 (1.1)
Justification	0.71	71	29	2.6 (1.1)	2.0 (1.0)
Presentation	0.53	29	71	2.3 (1.4)	2.6 (1.0)

Note. *designed for on-grade level learners
** each domain's scale range = 1 to 4
*** numbers in parentheses represents standard deviations

coefficients were considered moderate to substantial for all of the domains (see Table 3).

The Kappa coefficients for the domains in Task 3 (above-grade learners) could not be computed because of the lack of variability within the teacher ratings and within the NRC/GT staff ratings (i.e., domain ratings were constant within the set of teacher ratings and NRC/GT staff ratings; see Table 4). Three students responded to this particular level of the assessment task.

You Can't Convince Me. The purpose of this assessment is to engage students in thinking about, discussing, and identifying the essential elements of persuasive rhetoric. In addition, students are given the opportunity to practice communicating in a clear, concise manner to a specific audience and in a specific format. Students also engage in the process of preliminary instrument design as they create a rubric to be used by judges in evaluating persuasive speeches. Students are assessed in the domains of essential elements, checklist, clarity of descriptors, presentation, and peer evaluation (optional).

Table 5 indicates that the interrater reliability of the domains ranged from 0.42 to 0.81, with exact agreement on the ratings between the teacher and NRC/GT staff

Table 4

Independent Ratings (Teacher & NRC) for Wall Street Decisions: Task 3

Domain*	% Exact of Agreement	% Adjacent of Agreement	Teacher $\overline{X}$ Rating	NRC $\overline{X}$ Rating
Support for Conclusions	33	33	4.0 (.00)	3.0 (1.0)
Strategy and Calculations	66		4.0 (.00)	3.0 (1.7)
Supporting Materials		66	4.0 (.00)	2.3 (1.2)
Justification		66	4.0 (.00)	2.7 (.58)
Presentation**				

Note. * each domain's scale range = 1 to 4
** did not have presentation ratings from teachers or NRC/GT staff
*** numbers in parentheses represents standard deviations

Table 5

Independent Ratings (Teacher and NRC) for You Can't Convince Me: Task 1

Domain*	Kappa	% Exact of Agreement	% Adjacent of Agreement	Teacher $\overline{X}$ Rating	NRC $\overline{X}$ Rating
Essential Elements	.81	89	11	2.6 (.73)	2.7 (.71)
Checklist	.59	22	56	2.4 (.53)	1.7 (.71)
Clarity of Descriptions	.78	22	56	2.4 (.53)	1.4 (.53)
Presentation	.42	33	44	2.1 (.78)	1.8 (.44)

Note. * each domain's scale range = 1 to 3
** numbers in parentheses represent standard deviations

ranging from a low of 22% (checklist and clarity of descriptions domains) to a high of 89% (essential elements domain). As estimates of rater reliability using previously indicated guidelines, the Kappa coefficients were considered moderate to substantial for all of the domains.

Creature Classifications. The purpose of this assessment is to assess students' proficiency in developing classification systems for biological organisms. Students are assessed in the area's appearance, bug-selection decisions, thoroughness, and ease of use/quality of classification.

Fifteen students responded to the assessment. Table 6 indicates that the interrater reliability of the assessment domains ranged from 0.55 to 0.95, with the exact agreement rate ranging from 40% (appearance) to 93% (bug selection and ease of use domains). Landis and Koch's (1977) guidelines for rater reliability indicated that two of the domains, bug selection and ease of use, were almost perfect (0.80–1.0), while the other two domains, appearance and thoroughness, were moderate (0.40–0.59) and substantial (0.60–0.80), respectively.

Where in the World? This assessment task is designed to measure students' understanding of key cultural elements of countries and regions around the world. Students are assessed in the areas of thoroughness, validity of choices, appeal of display, and supporting materials.

Forty-one students' responses to the assessment were evaluated. Interrater reliability of the domains (see Table 7) ranged from 0.10 (supporting materials) to 0.72 (thoroughness), with the exact agreement rate ranging from a low of 57% (validity of choices and appeal of display domains) to a high of 83% (supporting materials domain). The previously established guidelines for judging the Kappa coefficient as an indicator of rater reliability indicate that the supporting materials domain reliability was only

Table 6

Independent Ratings (NRC) for Creature Classifications

Domain*	Kappa	% Exact of Agreement	% Adjacent of Agreement	Teacher $\overline{X}$ Rating	NRC $\overline{X}$ Rating
Appearance	.55	40	47	2.5 (.65)	2.2 (.68)
Bug Selection	.95	93	7	2.9 (.26)	3.0 (.00)
Thoroughness	.61	73	20	2.0 (.85)	2.2 (.77)
Ease of Use	.95	93	7	1.9 (.83)	1.9 (.88)

Note. * each domain's scale range = 1 to 3
** numbers in parentheses represent standard deviations

Table 7

Independent Ratings (NRC) for Where in the World: Task 2

Domain*	Kappa	% Exact of Agreement	% Adjacent of Agreement	Teacher $\overline{X}_1$ Rating	NRC $\overline{X}_2$ Rating
Thoroughness	.72	74	26	1.9 (.77)	1.9 (.83)
Validity of Choices	.43	57	40	2.8 (.53)	2.5 (.59)
Appeal of Display	.67	57	36	1.8 (.72)	2.1 (.78)
Supporting Materials	.10	83	17	2.0 (.31)	2.0 (.27)

Note. * each domain's scale range = 1 to 3
** numbers in parentheses represent standard deviations

slight (0.0–0.2), the validity of the choices domain was moderate (0.40–0.59), and the validity of the thoroughness and appeal of display domain was substantial (0.60–0.80).

Findings

Teachers' and Students' Responses to Authentic Assessments

Collecting reliability and validity evidence on the authentic assessments is only useful to the degree that a teacher would implement the assessments in his or her classrooms. Teachers and students involved in classrooms where authentic assessments were implemented were asked to reflect on their experiences with using or doing the assessments.

Middle school teachers and students generally expressed positive responses about the differentiated assessments in the middle school. As one teacher put it, "Most of [the students] I'd say for the most part seemed to enjoy it and seemed to get something out of it. Two or three of them did above and beyond, did beautiful, beautiful work. I was very, very thrilled" (Arnold interview, Y3, #1, p. 1).

Assigning Assessment Outside of Class Time

For many teachers, using differentiated assessments was a new approach and required teachers to reconceptualize the classroom. For many, the first step was to assign the work to be completed outside of class, rather than to change instructional and classroom routines. Teachers frequently introduced the assessments during class, but required the bulk of the work to be done outside of class time. Joan Borden, a seventh-grade teacher at one middle school, described the introduction of the assessment task, Creature Classification:

> I took the rubric and we spent one class period—in fact, actually, it was two [class periods]—step-by-step telling them what was expected. I explained to them that everybody was working for a 3. That's the one I emphasized. We mentioned the 2, and I told them since it would be failing, we wouldn't even discuss that. They could read that

> on their own, but as I went through it with each class, I emphasized the 3. (Borden interview, Y3, #5, p. 1)

Following this initial introduction, Ms. Borden largely left students to complete the tasks independently. "For science, we have no time in class . . . we had to do it all by ourselves, and I had to go to the library and get about 500 books" (student interview, Y3, #3, p. 5).

Teachers in other subject areas followed suit. While the bulk of work was completed outside of school, eighth-grade math teacher Wendy Arnold described how she incorporated skills and concepts from Wall Street Decisions into other math instruction:

> I kind of took it [assessment task] a little piece at a time every day, and we just built on that. The rubric was given to them when I gave them the pack of what they're supposed to do. We went through that where they knew what was going to be expected, where they could organize their little checklist and all this kind of thing from the rubric. So we worked on it [assessment task] and did some pieces just about every day, but they put it all together themselves. (Arnold interview, Y3, #1, pp. 2–3)

Use of Rubrics to Guide Project Completion

Students used the rubrics to guide the completion of their work in varying ways: to guide their initial planning, during the process, and at the end of the project to check accuracy and completeness. Many students explained how they used the rubric accompanying the tasks at the beginning of the project, finding the detailed criteria helpful in their initial planning. "I was looking at all of this stuff and me and my dog were sitting there and we put this in to try to get to expert . . . we try to put a little bit of everything in it" (student interview, Y3, #1, p. 4). Another student verified the helpful nature of the rubrics to guide their work processes: "[I looked at it while I was doing the project]. To look and see what we were supposed to do on it. Yes, ma'am, it was real helpful" (student interview, Y3, #2, p. 8).

Others took a different approach, using the rubrics periodically while completing the assessment. One student explained how the rubrics guided his thinking through the process as his understanding of the task developed over time: "The more I read [about the stock market], I realized it had nothing to do with [the specific task requirements], and so I picked out what I thought was the best for each company and then I put it down here" (student interview, Y3, #3, p. 9). The specificity of the rubric and the key objectives of the task assisted the student in identifying the essential elements and discarding other, less relevant information.

Other students used rubrics most significantly at the conclusion of the project. The rubric allowed students the opportunity to complete the assessment and then use the rubric to determine whether all required elements were present, sufficient, and in the correct format. In essence, some students used the rubric as a final checklist.

> The first time I went through, [I realized] that I needed to add a little bit more of supporting materials. At first, I didn't put in the [mathematical computations] on [the appropriate sheets], and I had to do calculations and estimations and stuff. (student interview, Y3, #1, p. 4)

Although students varied in their use of the rubrics to guide their project completion, all seemed to agree that the rubrics were helpful. Students liked the teachers' clear explanations of product expectations characteristic of the rubric. As one student put it, "It [rubric] was more detailed, like on this, it said 20 or more. . . . I mean, this one said exactly what I needed to hear . . . and I just needed to read it once to know what I was doing" (student interview, Y3, #3, p. 11).

Teachers acknowledged the students' positive reaction to the rubrics.

> Most of them [students] liked the [rubrics] because it gave them definite guidelines. They're used to rubrics; this wasn't the first time they've seen a rubric. They like to know exactly what they needed to have and where. Some of the kids wanted more clarification, exactly what this, that, and the other. Most of the kids really liked it. They like to see things cut and dried, and black and white, where they know exactly what they need to do. (Arnold interview, Y3, #1, p. 2)

Although students clearly appreciated clarity and specificity in their teachers' explanation of project expectations, they also appreciated the opportunity to interpret some elements of the task creatively.

> Yeah, I like [rubrics to be] specific because if I have it specific, I know exactly what I'm going to do, but if it's a little open, I can have a little creativity in there, and do a little more things, and

still get what she's asking for. (student interview, Y3, #3, p. 12)

Potential for Future Use

Although students and teachers agreed upon the positive response to differentiated authentic assessments, teachers were mixed about their future use of the new assessment approach. As a result of involvement in the authentic assessment project, Joan Borden began to shift her instructional and assessment behaviors.

> Next year . . . and I'm thinking maybe this summer about trying to make it a unit, that's a maybe, and see if I can go back and incorporate the text and all of this stuff that we're held to the fire with and let everything I do revolve around entomology, but that's a kind of pie in the sky idea right now, and it would just depend on . . . if I really had . . . I just have to sit down and look at what I could incorporate using the insects. I think it's a possibility, but I just have to go through. (Borden interview, Y3, #5, p. 4)

Other teachers resisted the idea of a significant change in their instructional and assessment behaviors to better attend to student academic diversity through differentiated assessments, citing irreconcilable differences with state testing formats.

> [Would I use it again?] If time were available. I'll tell you, I really had to push to get it in. They have us so crammed with all this [state testing] stuff and they keep changing years on us with what we're supposed to do and how and everything, that it's tough. I enjoyed doing it with the kids, and I can see a lot of areas where it's worthwhile, but my problem right now is they have us so hogtied. (Arnold interview, Y3, #1, p. 2)

Conclusions and Implications

On a national level we have a history for demanding that assessments provide quantifiable information about student learning that is both reliable and valid. However, as a nation we have failed in working with teachers to develop classroom assessments that provide high-quality information about student learning so that the instructional process is better informed. To date, guidelines with psychometric standards for classroom assessments where teachers make judgments about student learning do not exist.

While a review of the literature revealed no studies on the reliability of classroom assessments, in general the interrater reliability coefficients were similar to those found in studies on classroom observations of student performance. In general, the Kappa coefficients ranged from 0.55 to 0.95, indicating that ratings between two independent raters were fairly consistent with one another, despite the lack of training. This range of coefficients also suggests that the assessments elicit student responses that are reflective of the performance criteria in the scoring rubrics. In addition, the coefficient ranges suggest that the criteria are clearly delineated. For any coefficients that fell into a less-than-acceptable range, the domain descriptions need to be more clearly defined.

The results of this study begin to provide evidence that differentiated authentic assessments for classroom purposes can be developed to provide consistent information about student learning. In addition, the results suggest that these assessments can be used in middle school classrooms to assess students' obtainment of academic learning standards. Oftentimes, particularly in high-stakes accountability environments, the focus of classroom instruction is on test preparation (Moon et al., 2003; Moon, Callahan, & Tomlinson, 2002), rather than helping students gain understanding through the construction of their own knowledge and making interconnections among facts and concepts within and across disciplines. This view of learning is reflected in many contemporary instructional methods used in today's classrooms: writing across the curriculum, hands-on approaches, problem solving and reasoning emphases, and cooperative learning.

Although the number of students responding to the assessments was small, this study does begin to provide evidence to suggest that, with proper development and implementation, teachers can successfully use differentiated authentic assessments, the type advocated by gifted education professionals, in their classrooms to measure academic standards identified for the content areas.

References

Alexander, P. A., & Murphy, P. K. (1994). *The research base for APA's learner-centered psychological principles*. Invited symposium at the annual meeting of the American Educational Research Association, New Orleans, LA.

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory.* Prospect heights, IL: Waveland Press.

Archbald, D. A. (1991). *Authentic assessment: What it means and how it can help schools* (Report No. TM 020 999). Madison, WI: National Center for Effective Schools Research and Development. (ERIC Document Reproduction Service No. ED366624)

Baldwin, A. Y. (1994). The seven plus story: Developing hidden talent among students in socioeconomically disadvantaged environments. *Gifted Child Quarterly, 38*, 80–84.

Callahan, C. M., Tomlinson, C. A., Moon, T. R., Tomchin, E. M., & Plucker, J. A. (1995). *Project START: Using a multiple intelligences model in identifying and promoting talent in high-risk students* (Research Monograph No. 95136). Storrs: National Research Center on the Gifted and Talented, University of Connecticut.

Carnegie Council on Adolescent Development. (1990). *Turning points: Preparing American youth for the 21st century.* New York: Carnegie Corporation.

Cheek, D. W. (1993). Plain talk about alternative assessment. *Middle School Journal, 25*(2), 6–10.

Clausen, D. R., Middleton, J. A., & Connell, T. J. (1994). Assessing artistic and problem solving performance in minority and non-minority students using a nontraditional multidimensional approach. *Gifted Child Quarterly, 38*, 27–32.

Dana, T. M., & Tippins, D. J. (1993). Considering alternative assessment for middle level learners. *Middle School Journal, 25*(2), 3–5.

Darling-Hammond, L. (1997). *The right to learn: A blueprint for creating schools that work.* San Francisco: Jossey-Bass.

Egan, K., & Gardner, H. (1992). An exchange: The unschooled mind: How children think and how schools should teach. *Teachers College Record, 94*, 397–407.

Gardner, H. (1993). *Multiple intelligences: The theory in practice.* New York: BasicBooks.

Gordon, E. W. (1992). *Implications of diversity in human characteristics for authentic assessment* (Report No. CSE-TR-341). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing, University of California at Los Angeles.

Gordon, E. W., & Bonilla-Bowman, C. (1996). Can performance-based assessment contribute to the achievement of educational equity? In J. B. Baron & D. P. Wolf (Eds.), *Performance-based student achievement: Challenges and possibilities* (pp. 32–51). Chicago: National Society for the Study of Education.

Henderson, P., & Karr-Kidwell, P. J. (1998). *Authentic assessment: An extensive literary review and recommendations for administrators* (Report No. TM 028 235). (ERIC Document Reproduction Service No. ED 418140)

Jackson, A. W., & Davis, G. A. (2000). *Turning points 2000: Educating adolescents in the 21st century*. New York: Teachers College Press.

Jamentz, K. (1994). Making sure that assessment improves practice. *Educational Leadership, 51*(6), 55–57.

Kennedy, M. (1996). A teacher's manifesto: Designing learning which cures rather than causes academic risk: Part 2. *Journal of At-Risk Issues, 2*(2), 16–27.

Kraemer, H. C. (1982). Kappa coefficient. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences*. New York: Wiley.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

Lines, C. (1994). Authentic assessment at the middle level. *Middle School Journal, 25*(4), 39–41.

Marzano, R. J. (2002). A comparison of selected methods of scoring classroom assessments. *Applied Measurement in Education, 15*, 249–267.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Moon, T. R., Brighton, C. M., & Callahan, C. M. (2003). The influences of state testing programs on elementary teachers and students. *Roeper Review*, *25*, 45–60.

Moon, T. R., Callahan, C. M., & Tomlinson, C. A. (2002). Effects of state testing programs on elementary schools with high concentrations of student poverty-Good news or bad news? *Current Issues in Education, 6*(8). Retrieved June 23, 2005, from http://cie.asu.edu/volume6/number8

National Middle School Association (NMSA). (1995). *This we believe: Developmentally responsive middle schools.* Columbus, OH: Author.

Reed, L. (1993). Achieving the aims and purposes of schooling through authentic assessment. *Middle School Journal, 25*(2), 11–13.

Resnick, L. (1987). The 1987 presidential address: Learning in school and out. *Educational Researcher, 16*(9), 13–20.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. Gifford & M. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37–75). Norwell, MA: Kluwer.

Tomlinson, C. A. (1995). *How to differentiate instruction in mixed-ability classrooms*. Alexandria, VA: Association for Supervision and Curriculum Development.

Tomlinson, C. A. (2001). *How to differentiate instruction in mixed-ability classrooms* (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

VanTassel-Baska, J., Bass, G., Reis, R., Poland, D., & Avery, L. (1998). A national pilot study of science curriculum effectiveness for high-ability students. *Gifted Child Quarterly, 42,* 25–36.

VanTassel-Baska, J., Zuo, L., Avery, L. D., & Little, C. A. (2002). Curriculum study of gifted-student learning in the language arts. *Gifted Child Quarterly, 46*, 30–44.

Author Note

Funding for this research was supported under the Javits Act Program (Grant No. R206R000001) as administered by the Office of Educational Research and Improvement, U.S. Department of Education. Grantees undertaking such projects are encouraged to express freely their professional judgment. This research, therefore, does not necessarily represent positions or policies of the government, and no official endorsement should be inferred.

Appendix 1
Assessment Standards

Assessment Name	Standards to be Assessed
Folktales and Fables	I. Students will demonstrate their ability to: • create a story with a message/purpose. • sequence an orally presented story in a way that is easy for the listener to follow. • use symbolism effectively in their storytelling. • select and use colorful nouns, verbs, adjectives, and adverbs appropriately. • vary the tone and volume of their voice to add drama to their storytelling. • complete a project in a timely manner. II. Students will: • use verbal communication skills such as word choice, pitch, feeling, tone, and voice. • organize and synthesize information for use in written and oral presentations. • elaborate on a central idea in an organized manner.
Wall Street Decision	I. Students will demonstrate their ability to: • use mathematical logic to make an appropriate decision given many equally appealing choices. • choose appropriate strategies to solve problems. • apply strategies appropriately. • perform accurate mathematical calculations, transformations, and conversions. • use graphs, tables, and/or charts to organize and display relevant information. • describe their problem-solving and decision-making process so that others can easily understand them. • present information in a legible and appealing format. II. Students will: • identify representations of a given percent an describe in writing the equivalence relationship between fractions, decimals, and percents. • solve problems that involve addition, subtraction, and multiplication • use estimation strategies to solve multi-step practical problems involving whole numbers, decimals, and/or fractions.

- compare, order, and determine equivalent relationships among fractions, decimals, and percents.
- solve consumer application problems.
- solve practical problems involving whole numbers, integers, and rational numbers, including percents. Problems will be of varying complexity involving real life data.

You Can't Convince Me

I. Students will demonstrate their ability to:

- identify elements of persuasive rhetoric.
- analyze the elements of persuasive rhetoric in order to choose the most "critical" elements.
- communicate appropriately to a chosen audience.
- organize ideas in a clear and concise manner.
- work collaboratively in pairs.

II. Students will:

- use a variety of planning strategies to generate and organize ideas.
- select vocabulary and information to enhance the central idea.
- give and seek information in conversations and group discussions.
- identify persuasive messages in non-print media.
- apply knowledge of the characteristics of various literary forms.
- identify persuasive techniques.

Creature Classifications

Students will demonstrate their ability to:

- access scientific data and/or information.
- describe biological creatures in multiple ways.
- classify organisms in useful ways.
- visually present information about scientific organisms in a manner that appeals to a specific audience.
- appropriately cite sources of information.

Where in the World?

Students will demonstrate their ability to:

- engage in a logical process of research, analysis, and questioning that leads them to valid, thorough information about a concept or idea.
- choose the most relevant information about a region to communicate a big idea or them to a specific audience.
- visually present information about cultural regions in a manner that is appealing to a specific audience.
- choose cultural regions or countries that emulate specific characteristics.

Appendix 2
Sample Authentic Assessment
Fables and Folktales

A good storyteller grabs the imagination of his or her audience and holds the listeners captive with the tales he or she is telling. You have learned about fables and different types of folktales: Trickster Tales, How-and-Why Stories, Tales of Enchantment, and so forth. Now it is your turn to weave your own magic.

The Situation: The year is 2060. You have lived a long life and learned much along the way. A teacher at a local middle school has invited you to participate in the annual storytelling festival hosted by the school. You must create your own fable or folktale to share with the students.

In the process of developing your story, you will need to ask yourself a number of questions, including the following:

- What type of story do I want to tell?
- What message/moral/explanation/advice do I want my story to give to the listeners?
- How will I use symbolism to connect my story to uni-

versal themes that transcend time and/or place?

- Do I want to modernize or revise an old story or create a brand-new one?
- Who will my characters be and what will they be like?
- What will my story be about?
- How will my story unfold? What will happen first? What will happen next?
- What storytelling techniques will I use in sharing my story with others?

As you decide on answers to these questions, record your ideas on the planning page provided (front and back). Once this form is complete, have the teacher look it over and initial it when he or she is satisfied that you are ready to put the pieces together into a well-crafted story.

As you develop the story itself, think about how you can make the words you use, the details that you include, and the expressiveness of your voice make the tale you tell more interesting and/or exciting.

The storytelling festival is scheduled for _____.

Come prepared—Your work will be evaluated using a score sheet like the one below.

	Wondrous Wordsmith (3 Points)	**Skillful Storyteller (2 points)**	**Tale-Teller in Training (1 point)**
Purpose Score: _____	The story you tell clearly and powerfully leads your listener to understand and appreciate the main idea/message.	The listener is able to understand the purpose of your story.	The main point of your story is unclear. The listeners are left unsure of the message you are trying to get across.
Sequencing Score: _____	You effortlessly lead your listener along your story's path—from the introduction of the characters to the final resolution of conflict.	There are minor inconsistencies or gaps in the sequencing of your story. Still, listeners are able to understand and follow the basic series of events.	The listener is unable to follow your story. The sequence of events you use is illogical or overly cumbersome.
Symbolism Score: _____	Characters and events in your story are clearly symbolic of people and happenings across time and/or generations.	You use symbolism to represent people or happenings, but the symbolism does not easily transfer or connect to other times and/or generations.	There was little or no symbolism apparent in your story, or the symbolism does not transfer to other times and/or generations.
Word Usage Score: _____	You use vivid and powerful nouns, verbs, adjectives, and adverbs when telling your story. Your listener can visualize in detail what happens.	You use nouns, verbs, adjectives, and adverbs appropriately to express your ideas. Your listener is able to picture events or people in your story.	You do not make appropriate use of nouns, verbs, adjectives, and adverbs. Your listener is unable to visualize people or events in your story.
Expressiveness Score: _____	Your story comes to vibrant life as you vary the tone and volume of your voice to match what is happening in your story.	Your voice is clear as you tell your story, but you do not vary your tone of voice and/or volume in a way that captivates and holds the listener's attention.	It is difficult to hear you as you tell the story. You do not vary your volume or tone of voice.
Timeliness Score: _____	You are prepared and present your story at the festival as scheduled.	You are not prepared to present your story at the scheduled time, but you present it within 1 or 2 days.	You are not prepared to present your story at the scheduled time or within 2 days of the festival.

About the Authors

Catherine M. Brighton is an assistant professor at the Curry School of Education, University of Virginia, a principal investigator for the National Research Center on the Gifted and Talented, and director of the Institutes on Academic Diversity at the University of Virginia. She earned her master's and doctoral degrees at the University of Virginia in educational psychology/gifted education. She is the president of the Virginia Association for the Gifted, a former classroom teacher, teacher of the gifted, curriculum coordinator, and elementary assistant principal.

Leah B. Bucknavage is a doctoral candidate in the School Psychology Program at The Pennsylvania State University. Her research interests lie primarily in the area of early childhood education and services to this population. She is especially interested in parent interventions and how they can be used to improve school outcomes. She can be contacted at lbb133@psu.edu.

Carolyn M. Callahan is a department chair and professor in the Curry School of Education, University of Virginia, and is also associate director of the National Research Center on the Gifted and Talented. She has authored numerous articles, book chapters, and monographs on the topics of creativity, the identification of gifted students, program evaluation, and the issues gifted females face. She is past president of The Association for the Gifted and the National Association for Gifted Children and currently serves as an at-large member on the board of NAGC. She serves on the editorial boards of *Gifted Child Quarterly*, *Journal for the Education of the Gifted*, and *Roeper Review*.

David W. Chan is a professor in the Department of Educational Psychology and program supervisor of the Programs for the Gifted and Talented at the Chinese University of Hong Kong. His research interests include stress and coping, psychopathology and health, creativity and intelligence, and gifted education and talent development. He can be contacted at davidchan@cuhk.edu.hk.

Ho-Seong Choe is a professor in the Department of Education, Kyungnam University. His research interests include curriculua for gifted students and evaluation of gifted education programs.

Andrew Johnson is professor of holistic education in the Department of Educational Studies: Special Populations at Minnesota State University, Mankato. He specializes in gifted education, holistic education, literacy instruction, and spiritual intelligence. His most recent books include *Making Connections with Elementary and Middle School Social Studies* (Sage) and *The Inner Curriculum: Classroom Activities to Develop Emotional Intelligence* (Royal Fireworks Press). He can be contacted at andrew.johnson@mnsu.edu.

Kyung Hee Kim is an assistant professor in the Department of Teacher Education at Eastern Michigan University. She obtained her Ph.D. in gifted and creative education from the Department of Educational Psychology at the University of Georgia. She has worked at the Center for Test Scoring and Reporting Services as a research assistant and has also worked at the Torrance Center for Creativity and Talent Development as the international program coordinator at the University of Georgia. Her research interests include the nature of creativity, characteristics of creatively gifted students, assessing creativity, and nurturing creativity, particularly concerning environmental and cultural interactions.

Eun Ah Lee is a postdoctoral researcher at the School of Natural Resources at The Ohio State University. She has her M.Ed and Ph.D in science education from Seoul National University in Korea. She has worked for several institutes, including the Korean Educational Development Institute. Her research interests are earth system education, the nature of science, and scientific creativity.

Tonya Moon is an associate professor in the Curry School of Education at the University of Virginia and is also a principal investigator of the National Research Center on the Gifted and Talented. She has written numerous articles and monographs on the topics of assessment, the identification of gifted students, effects of high-stakes testing, and technical issues associated with state-mandated assessment programs. She is past president of the Virginia Educational Research Association and is on the executive board for the AREA Research on Giftedness and Talent SIG. She serves on the editorial boards of *Gifted Child Quarterly*, *Journal of the Education of Women and*

Minorities in Science, and *Journal of Middle Level Research Quarterly*.

Soo-Kyong Park is an earth science teacher at the Busan Science High School. Her research interests lie in the constructivist-oriented science instruction for gifted students.

Kyung-Hee Park is a researcher at the Korean Educational Development Institute. Her research interests include giftedness and professional development for gifted education teachers.

Jane Piirto is Trustees' Professor of Graduate Education at Ashland University, Ohio. She is the author of *Understanding Creativity*, *"My Teeming Brain": Understanding Creative Writers*, *Understanding Those Who Create* (two editions), and *Talented Children and Adults: Their Development and Education* (the third edition of which is currently in production). She has published many articles and is a frequent public speaker. She is an award-winning poet and novelist. She has also directed 15 Ohio Summer Honors Institutes for talented adolescents.

Ann Robinson is a professor of education and founding director of the Center for Gifted Education at the University of Arkansas at Little Rock. She has taught both high school and college English and served as a gifted and talented facilitator. Her research interests include professional development practices, the analysis of the knowledge base in gifted education, and teacher conceptualizations of the linkage between curriculum and assessment. She is a former editor of *Gifted Child Quarterly* and has served on the executive board of the National Association for Gifted Children. She is president of the Arkansas Association of Gifted Education Administrators, a professional advocacy organization.

Hae-Ae So is a researcher at the National Center on Gifted and Talented Education at the Korean Educational Development Institute. Her interests include science education and gifted education.

Frank C. Worrell is an associate professor and professor in charge of the doctoral School Psychology Program at the University of California, Berkeley, as well as the faculty director of the Academic Talent Development Program. His research interests include psychosocial variables in gifted and talented populations, as well as identification of students for GATE programs. He serves on the editorial boards of *Gifted Child Quarterly, The Journal of Secondary Gifted Education, Roeper Review,* and *School Psychology Quarterly.*