Journal of Advanced Academics
2014, Vol. 25(1) 47–66

DOI: 10.1177/1932202X14520946
joa.sagepub.com
SAGE

Assessing Science Reasoning and Conceptual Understanding in the Primary Grades Using Standardized and Performance-Based Assessments

Kyung Hee Kim[1], Joyce VanTassel-Baska[1], Bruce A. Bracken[1], Annie Feng[2], and Tamra Stambaugh[3]

Abstract

Project Clarion, a Jacob K. Javits-funded project, focused on the scale-up of primary-grade science curricula. Curriculum units, based on an Integrated Curriculum Model (ICM), were developed for high-ability learners, but tried out with all students in Title I settings to study the efficacy of the units with all learners. The units focus on the development of students' conceptual understanding to undergird science content attainment. Teaching and learning models, such as concept formation and concept mapping were used to scaffold science learning and reasoning for appropriate curriculum differentiation. Science content mastery was measured using the Metropolitan Achievement Tests. Reasoning skills were measured using the Test of Critical Thinking. Understanding of macro-concepts and content attainment were measured by curriculum-embedded performance-based assessments. Students with the ICM outperformed students without the specialized curriculum in science content and reasoning skills, and showed greater growth in both conceptual understanding and content attainment.

Keywords

concept development, concept mapping, critical thinking, inquiry, students with low SES, performance-based assessment, primary age/level, problem-based learning, science curriculum

[1]The College of William and Mary, Williamsburg, VA
[2]National Institution of Health, Maryland, USA
[3]Vanderbilt University Nashville, Tennessee, USA

Corresponding Author:
Kyung Hee Kim, Associate Professor of Educational Psychology, Room 3122, 301 Monticello Avenue, PO Box 8795, School of Education, The College of William and Mary Williamsburg, VA 23187.
Email: kkim@wm.edu

The focus of this study was to examine from the efficacy of Project Clarion, a Jacob K. Javits-funded study situated in Title I schools. One of the goals of Project Clarion was to develop rigorous science curricula for high-ability learners in Grades K-3. The Project Clarion curriculum embedded instructional scaffolding throughout the multiple units to support concept and content acquisition for all learners.

In 1985, the American Association for the Advancement of Science (AAAS) established "Project 2061" as a long-term initiative to establish benchmarks to advance scientific, mathematical, and technological literacy for all Americans (Rutherford & Ahlgren, 1991). Nine years later, the Goals 2000: Educate America Act (1994) was signed into law, declaring that the "United States students will be first in the world in mathematics and science achievement" (Sec 102, 5A). The National Research Council (NRC, 1996) established benchmarks for students' achievement in science and defined scientific literacy for students in the National Science Education Standards as, "the knowledge and understanding of scientific concepts and processes required for personal decision making, participation in civic and cultural affairs, and economic productivity" (p. 22).

Twenty years after Project 2061 was initiated and 9 years after the NRC established standards and operationalized the definition of scientific literacy, the 2005 results from the National Assessment of Educational Progress (NAEP) indicated that students in all grade levels showed a lack of understanding of scientific concepts and reasoning (Grigg, Lauko, & Brockway, 2006). More than a decade after declaring that the United States will be the first in the world in science, U.S. students scored less than the other countries on the Trends in International Mathematics and Science Study (TIMSS; Gonzales et al., 2008). More recently, the National Center for Education Statistics (NCES) reported the results of the 2009 Program for International Student Assessment (PISA). The results of the PISA showed that U.S. students scored lower than 22 other countries on the PISA in science (NCES, 2010). This outcome suggests that existing science curriculum and instruction have failed to help U.S. students, including the highest achieving students to develop scientific literacy including understanding of science concepts and reasoning skills.

Student Learning

The foundation of this study rests on the following four critical understandings from concept development studies:

1. Conceptual understanding is built on categorical representations in a foundational and continuous way (Quinn & Eimas, 1997).
2. Conceptual understanding improves with self-explanation (Chi, Hutchinson, & Robins, 1989; Pine & Messer, 2000; Zeigler, 1995).
3. Changes in conceptual understanding are grounded in the generalized integration of information (Linn & Songer, 1991).
4. The thinking processes used by good learners can be emulated to design powerful educational interventions for all learners (Boyer, Bedoin, & Honore, 2001).

Research on science learning has shown that students' efficient use of thinking processes, such as analogical reasoning, metacognition, and articulation of learning, and instructional strategies, such as concept mapping and collaborative work, contribute to effective science learning, when used individually or collectively (Boyer et al., 2001; Novak, 1998; NRC, 2002; Zeigler, 1995). These findings lend support to the NRC suggestions that science instruction should

1. develop conceptual understanding, rather than fact-based understanding alone;
2. utilize concept maps to support development of conceptual understanding, embed the teaching of higher level thinking skills within the teaching of science content to support students' development of scientific reasoning; and
3. teach metacognitive strategies to develop and promote students' scientific problem-solving skills (NRC, 2002).

These recommendations also point to specific strategies that can be used to help develop scientific talent in students from a variety of backgrounds, with appropriate scaffolding and support.

The Need for Rigorous Curriculum to Nurture Science Talent

Data from longitudinal early childhood studies have demonstrated that development and implementation of intensive, high-quality, pervasive interventions can impact achievement patterns for students with low socioeconomic status (SES; Borman & Hewes, 2002; Ramey & Ramey, 1998). Furthermore, research in gifted education has shown that students with low SES can benefit from targeted interventions in content areas that are focused on higher level skill development within specific content areas, including science (Gavin et al., 2007; Little, Feng, & VanTassel-Baska, 2007; VanTassel-Baska, Avery, Hughes, & Little, 2000; VanTassel-Baska, Bass, Ries, Poland, & Avery, 1998; VanTassel-Baska, Zuo, Avery, & Little, 2002). Research outcomes from the Advanced Placement (AP) and International Baccalaureate (IB) programs have shown that students with low SES who do not have prerequisite academic skills (e.g., writing, study, and time management) necessary for success in these courses are unable to acquire the skills fast enough within the courses to be successful (Herberg-Davis & Callahan, 2008). Furthermore, the longitudinal research from the study of mathematically precocious youth has revealed that a confluence of factors contributes to the development of scientific expertise over the life span, including (a) investigative interests and a focus on finding truth through cognitive means, (b) ability in mathematics, (c) high levels of spatial ability, (d) a sustained commitment to scientific pursuits, (e) dedication to school and work within and outside the school or work environment, and (f) specific educational opportunities (Lubinski & Benbow, 2006). This research suggests that educational opportunities, in the form of a rigorous science curriculum, can be helpful to teach science literacy to all learners. It also suggests that instructional strategies that scaffold prerequisite skills must be embedded within

science curricula to nurture scientific talent and develop the scientific habits of mind necessary for success, for those students with low SES.

Instructional Features of a Rigorous Science Curriculum

A rigorous science curriculum that scaffolds learning for students who may need additional support has several distinct features. For example, a science curriculum that integrates high-level content, scientific processes, authentic products, and is concept-based has been found to enhance the science achievements of elementary gifted students with low SES (Feng, VanTassel-Baska, Quek, O'Neill, & Bai, 2005; Kim et al., 2012). In addition, inquiry-based instructional approaches have traditionally been found to be effective in teaching science. For example, problem-based science curricula and project-based learning have been found to be effective with gifted and high-ability students who have shown gains in achievement on knowledge acquisition, knowledge application, and science investigation skills at all levels of K-12 schooling (Mioduser & Betzer, 2008; VanTassel-Baska et al., 1998). Furthermore, Swanson (2006) found that problem-based curricula also produced meaningful gains in science achievement among gifted students with low SES. Moreover, Rayneri, Gerber, and Wiley (2006) found that gifted students show preferences for hands-on learning in science. Finally, VanTassel-Baska, Feng, and Brown (2008) reported that well-designed research-based curriculum units differentiated for gifted learners improved teachers' general use of differentiation strategies.

Targeted interventions that include well-designed, rigorous curricula provided to gifted students with low SES early in their educational careers are critical for such learners' future success. The Integrated Curriculum Model (ICM) was used as the basis to develop the science curriculum in this study (VanTassel-Baska, 1986; VanTassel-Baska & Little, 2003, 2011). The ICM provided the framework for the integration of content knowledge with concept development and higher level scientific research processes within the context of a problem-based scenario. In addition, differentiation strategies graphic organizers, and other scaffolding efforts were infused throughout the lessons to meet the needs of all learners and provide opportunities for students with low SES to acquire the prerequisite skills needed to excel in advanced science courses in high school.

Assessing Science Learning With Young Gifted Students With Low SES

The research base for using performance-based assessments in various fields to measure curricular goals, such as those previously identified in this article, has been built on since the early 1990s. For example, research in gifted education indicates that curriculum-embedded performance-based assessments (PBAs) have been found to be valid and reliable measures of student learning, including science acquisition (Adams & Callahan, 1995; Moon, Brighton, Callahan, & Robinson, 2005). PBAs have been used successfully to measure complex reasoning, higher level thinking, and content learning in science. Furthermore, PBAs have been used successfully with populations of students who historically have had difficulty with selected-response, standardized

measures of achievement, including preschool children (Schappe, 2005), gifted special needs high school students (Cooper, Baum, & Neu, 2004), and gifted students with low SES (Tali Tal & Miedijensky, 2005; VanTassel-Baska, Feng, & de Brux, 2007; VanTassel-Baska, Feng, & Evans, 2007; VanTassel-Baska, Johnson, & Avery, 2002). In science, curriculum-embedded PBAs have proven to be effective assessments of science literacy, content understanding, scientific reasoning, and higher level thinking and performance across grade levels (Fowler, 1990; Liu, Lee, Hofstetter, & Linn, 2008; Spektor-Levy, Eylon, & Scherz, 2009). Pre- and post-treatment concept maps (Novak, 1998) have provided a valid and reliable approach for assessing changes in conceptual understanding in science (Nafiz, 2008). Given the complexity of the knowledge and skills required for students to demonstrate scientific literacy, multiple measures of performance are often needed for students to show what they know and what they are able to do in science. Therefore, when designing a complex and rigorous curriculum for gifted students, that has multiple learning outcomes related to higher level thinking, advanced science content, and conceptual understanding, multiple measures of performance should be embedded. Furthermore, external outcome measures should also be used as cross-validation measures to provide additional assessment of whether the curriculum accomplished its intended science literacy goals.

Purpose of the Study

Project Clarion was a 5-year, curriculum scale-up study funded under a Jacob K. Javits grant by the U.S. Department of Education (VanTassel-Baska & Bracken, 2004). The ICM and PBAs were designed according to the research-based best practices identified in the literature. The purpose of the project was to scale up a rigorous science curriculum previously found effective with elementary gifted learners in Grades 3 to 5 (VanTassel-Baska et al., 1998). The scaled-up curriculum was implemented with primary-grade students in Title I schools, using multiple dependent measures that assessed science content attainment and conceptual understanding. The ICM was developed and pilot tested in the first two years of the study and then implemented for 2 years. Data were collected for each year of implementation. The specific research questions addressed in this study included the following:

1. Do students with the ICM increase science content knowledge and reasoning skills?
2. Do students with the ICM increase *content* and *concept* mastery in science as measured by pre–post PBAs? If so, is there a difference in the increase in content and concept mastery by gender, ethnicity, grade, and/or ability level?

Method

Sampling Procedures

In the rural district where this study was conducted, schools were randomly assigned to the experimental and comparison treatments. In the urban and suburban participating

districts, teachers were randomly assigned to each condition, once districts and schools had agreed to participate in the project. Students were assigned to heterogeneous classrooms by principals at the beginning of each academic year, independent of the study. Thus, the project used intact student groups for curriculum implementation. All students took the Naglieri Nonverbal Ability Test (NNAT; Naglieri, 1997) as a baseline measure of cognitive functioning. The results revealed insignificant differences between the experimental and comparison groups at the onset of the study.

Participants

As of the 2008-2009 school year, during Year IV of Project Clarion, a total of 3,462 students across three Title I school districts in the state of Virginia had participated in the project (only about a third of the students made it to the end of the study due to students who changed schools, graduated out of the program, and so on). The students resided in rural (28%), urban (41%), and suburban school districts (31%). Within each school district, two Title I–designated schools participated. Title I schools are composed of a majority of low SES students. In the rural district, one school was randomly assigned as the experimental school and another school was randomly assigned as the comparison school. Within the experimental building, all classrooms of kindergarten, first-, second-, and third-grade students participated in the study. In the urban and suburban school districts, teachers were randomly assigned to either the comparison group or the experimental group in two schools with Title I status.

Experimental group. At the beginning of this study, teachers were randomly assigned to either the experimental or comparison group. However, due to foibles of school practices and policies, random assignment conditions could not be maintained for the final year of the project. Thus, experimental group students for this study included those who

1. received instruction in the ICM for 2 years during Grades 1 to 3,
2. participated in the PBAs for 2 years, and
3. took the first-year and the second-year follow-up post-achievement tests.

A total of 250 students in the experimental group met the assessment and intervention and duration requirements. Of the experimental students, 70 (28%) students were in first grade, 88 (35%) students in second grade, and 92 (37%) students in third grade. There were 136 (54%) girls and 114 (46%) boys. Ethnic and racial distribution included 173 (69%) Caucasian students, 37 (15%) African American students, 26 (10%) Hispanic American students, 4 (2%) Asian American students, 2 (1%) Native American students, and 1 (<1%) were classified as Other. Ethnicity and race data were missing for 7 students.

Comparison group. Comparison group students for this study were those students who

1. did not receive instruction in the ICM during the initial phase of the curriculum implementation,
2. took either the first-year or the second-year follow-up posttest-dependent measures.

These criteria resulted in 1,401 students in the comparison group, among whom 707 (51%) were boys and 692 (49%) girls. Gender data were missing for two students. Ethnic and racial distribution included 509 (36%) Caucasian students, 277 (20%) Hispanic American students, 265 (19%) Asian American students, 240 (17%) African American students, 6 (<1%) Native American students, and 19 (1%) were classified as Other. Ethnicity and race data were missing for 85 students.

Instruments

The study employed multiple measures to assess student learning and to check curriculum implementation fidelity. A nonverbal ability test, the NNAT (Naglieri, 1997) was used at the beginning of the study to assess baseline cognitive differences across control and experimental groups. Curriculum-embedded PBAs were administered as a pretest prior to curriculum implementation and as a posttest at the end of implementation. A norm-referenced, standardized test of science achievement, the *Metropolitan Achievement Test, Eighth Edition* (MAT-8; Harcourt Educational Measurement, 2001) was administered as a standardized measure of science. A critical thinking measure, the Test of Critical Thinking (TCT; Bracken et al., 2004) was used as a posttest only. Teachers were provided with training on administration of the assessments and administered the tests to their own students. Observation data on teacher implementation fidelity were collected once each year.

NNAT. The NNAT is a 38-item nonverbal matrix analogy test of spatial reasoning. Verbal directions are minimal. The test can be administered in either group or individual format. Forms are available for multiple grade levels. Time for administration is 30 min. Naglieri (1997) reported adequate raw score internal consistency, with coefficients between .81 and .89, and sufficient evidence of validity for screening and research purposes. The NNAT has been widely used for the identification of gifted students because of its sensitivity toward inclusion of low SES and minority students. The NNAT was administered to first-, second-, and third-grade students. Study staff administered the NNAT within the assigned experimental and comparison classrooms in the fall of the first year of this study. Based on the NNAT scores, the ability levels of students in the experimental group were categorized as high-, average-, and low-ability. High-ability students (21%) are those who scored above 115 (i.e., above +1 standard deviation), average-ability students (50%) are those who scored between 85 and 114 (i.e., between ±1 standard deviations), and low-ability students are those who scored less than 85 (i.e., less than −1 standard deviation). Students were neither denied nor granted access into participating in the curriculum based on their NNAT score.

NNAT internal consistency reliability for participating grades in this study ranged between .80 and .85. The NNAT was an appropriate measure for grouping students into the three general ability categories of interest.

MAT-8. The MAT-8 was chosen as an external measure of science achievement because it is well-known, widely used, and has been validated as a group-administered achievement test. On-grade level MAT-8 science measures were administered as pre- and post-assessments to experimental and comparison students in Grades 1 to 3. Each test required about an hour to administer in each class. The KR-20 and KR-21 coefficients reported in the examiner's manual are all greater than .80. Students of all three ability levels completed this assessment. Classroom teachers administered this test, following test administration procedures as identified in the administration manual. Assessed MAT internal consistency reliability for participating grades in this study ranged between .79 and .89.

TCT. The TCT (Bracken et al., 2004) was administered to experimental and comparison students in Grade 3. The TCT was based theoretically on Paul's (1992) model of critical thinking and the Delphi Report (Facione, 1990). Bracken et al. (2004) reported adequate internal consistency for research purposes, between .83 and .89. The TCT assesses children's critical reasoning skills within seven life domains including social, affect, academic, competence, family, physical, and spiritual. Only third-grade students were administered the TCT before graduating out of the program. Like the MAT, the TCT was administered by classroom teachers. TCT internal consistency estimates for the participating third-grade students in this study was .80.

PBAs. Curriculum-embedded PBAs were developed to assess learning derived from the curriculum intervention during the project implementation. Two instruments were designed or adapted from other instruments for the study. Each assessment focused on a different dimension of learning:

1. Conceptual understanding (i.e., *Concept*). To assess understanding of a macro-*concept* (systems or change), students responded to a series of open-ended questions which required students to provide examples of a macro-concept, a description of the features of one example, and generalizations about the targeted macro-concept, comparable with the requirements for concept formation (see Taba, 1966).
2. Science content attainment (i.e., *Content*). To assess science *content* attainment, students were asked to draw a concept map about the unit topic using a well-established model for concept mapping in science (Novak, 1998).

The pilot test for the PBAs focused on gathering validity and reliability evidence to determine the degree to which the assessments could be used to draw inferences about student understanding of the macro-concept (*Concept*) and science content (*Content*). Multiple studies were conducted to gather supporting validity evidence. First, experts

in the field of gifted education, curriculum development, and assessment conducted a content validity review of the PBAs. Expert reviewers rated each assessment on three dimensions (i.e., content relevance, clarity, and format) using a Likert-type scale of 1 to 3, with 1 being the lowest and 3 being the highest fit, and an overall mean score was calculated. The inter-rater reliability among expert reviewers for content relevance and clarity was .79. The reviewers also provided qualitative feedback, which was used to improve the assessments.

Second, a rubric was developed or adapted to score each PBA and exemplar identified to aid in scoring. Scale internal consistency ranged from .63 to .76 for the PBAs. Inter-rater reliability was assessed by having project staff and graduate students attain a threshold accuracy of 90% or higher on scoring 10 student responses independently after training. Results ranged from agreement proportions of .86 to .89 for each of the scorers. Random checking was also employed throughout the scoring process by a trained staff member. Classroom teachers administered the PBAs to experimental students at the beginning of each curriculum unit and on completion of the unit.

The ICM Design and Development

The units in the curriculum intervention followed a core framework based on the ICM (VanTassel-Baska, 1986). The ICM was developed for students enrolled in kindergarten through third grade. Major goals of each unit focused on developing an understanding of macro-concepts, such as *systems* or *change*, as well as developing scientific reasoning and investigative skills, and learning a science concept deeply and well. Unit content was aligned with national and state standards, which provided the opportunity for schools to use the units as a fundamental component of their curriculum (VanTassel-Baska et al., 1998). Multiple teaching models were used to systematically integrate science content and vocabulary with scientific investigative and reasoning skills including a vocabulary development model, a concept development model, the Wheel of Scientific Investigation and Reasoning, and concept mapping (Frayer, Frederick, & Klausmeier, 1969; Novak, 1998; Taba, 1962).

The ICM was developed during the first year of the study, pilot tested, and revised during the second year of the study. As a part of unit revisions, PBAs were refined, piloted, and made ready for use during the third and fourth years of the study. In addition, the ICM and the PBAs were validated during pilot testing. Content experts conducted reviews of each curriculum unit using the criteria of standards alignment, content validity, and instructional soundness. Furthermore, during each year of implementation, teachers provided feedback on the units, indicating areas of strength and needed improvements. Units were revised in a 3-year iterative process.

Professional development of teachers. Topics for initial training of teachers included concept development, the teaching models used in the units, science content, and training on the assessments. During the first year of the study, training was provided by the principal investigators and project director. During subsequent years, project ambassadors provided training. Training for each year of the project included 2 days of

Table 1. Means and Standard Deviations for First- and Second-Year Follow-Up MAT Scores and Second-Year Follow-Up TCT Scores for the Experimental (n = 250) and Comparison Groups (n = 1,401).

Group		Experimental (E)			Comparison (C)			E − C	
Test	Time	*n*	*M*	*SD*	*n*	*M*	*SD*	Δ	*d*
MAT	Follow-up I	165	22.10	3.74	1,075	21.16	4.89	.94**	0.22
	Follow-up II	57	24.30	5.53	221	22.46	5.63	1.84*	0.33
TCT	Follow-up II	85	16.14	5.67	318	14.18	5.45	1.96**	0.35

Note. MAT = Metropolitan Achievement Test; TCT = Test of Critical Thinking.
*p < .05. **p < .01

instruction for all teachers of the experimental group. Follow-up professional development included information on grouping and differentiating instruction for an additional 2 days. Teachers were also provided with time for collaborative planning. Finally, a professional development "ambassador" was assigned to each school district to answer questions, facilitate unit implementation, and provide follow-up professional development as needed. Project ambassadors were highly experienced and skilled personnel in the field of gifted education. During the final year of the project, comparison teachers were provided with an opportunity to receive 2 days of training on the ICM.

Comparison group curriculum. Comparison students received the districts' adopted science curriculum during the implementation of Project Clarion. District-based science instruction was very limited at the K-1 levels in the comparison classrooms. Typically, instruction consisted of a few experiments, implemented once a week. In Grades 2 to 3, science instruction was provided for approximately 120 min per week. Instruction consisted of conducting experiments and studying topics delineated in the core textbook.

Results

Research Question 1: Differences in the MAT and the TCT Between Experimental and Comparison Groups

Table 1 shows means and standard deviations for first- and second-year follow-up MAT scores and second-year follow-up TCT scores for experimental (n = 250) and comparison (n = 1,401) groups. The results of the independent-samples t test for the first-year follow-up MAT indicated that the experimental (n = 165) group scored significantly higher than the comparison (n = 1,075) group, even after the Bonferroni correction (adjusted $\alpha = .05 / 3 = .017$), $t(1238) = 2.36$, $p = .005$, Cohen's $d = 0.22$ (small effect). Subsequently, the results for the second-year follow-up MAT also indicated that the experimental (n = 57) group scored significantly higher than the

comparison (n = 221) group, $t(276) = 2.21$, $p = .028$, Cohen's $d = 0.33$ (small to medium effect). This was insignificant after the Bonferroni correction, which might be due to the smaller sample size for the second-year follow-up MAT. Finally, the results for the second-year follow-up TCT also indicated that the experimental ($n = 85$) group scored statistically significantly higher than the comparison (n = 318) group, even after the Bonferroni correction, $t(401) = 2.92$, $p = .004$, Cohen's $d = 0.35$ (small to medium effect).

Research Question 2: Growth in Concept *and* Content *by Latent Growth Curve (LGC) Modeling*

LGC modeling. LGC modeling was conducted using AMOS (version 17) software (Arbuckle, 2008) to answer the second research question. LGC is an advanced application of structural equation modeling (SEM) used to analyze longitudinal data. Unlike typical SEM, LGC model does not specify latent variables that represent underlying theoretical construct that are represented by observable indicators. Instead, in LGC there are two latent variables of intercept and slope: an intercept parameter, which represents each individual's score at baseline; and a slope parameter that represents each individual's rate of growth across the timeline. LGC modeling has several advantages over SEM (Bollen & Curran, 2006; Byrne, 2010). LGC modeling enables researchers to explicitly answer questions about patterns in data observed across multiple points in time (Bollen & Curran, 2006). Specifically, it enables, in one analysis, to determine whether significant overall growth occurs within individuals over time, whether significant variability exists in between-individuals' growth, and whether a relationship exists between baseline values and growth over time. Another advantage is that LGC modeling enables researchers to model and test both potential covariates of growth and growth trajectory (e.g., linear, quadratic) in the same analysis (Byrne, 2010). A final advantage is that it enables the estimation of a measurement model that separates measurement error from true score growth.

LGC analysis requires multivariate normative data, an adequate sample size, and data from the same measure on at least three separate occasions. Multivariate normality is a common assumption in SEM (Kline, 1998). In this study, we examined the values of univariate skewness and kurtosis to judge whether each variable was approximately normally distributed. No values of the skewness and kurtosis were greater than |1.0|. In addition, data were screened for outliers. Two outliers were found, which were removed from the analyses. There were missing values before the analyses, but all of the missing values were deleted listwise so that 236 students were used for the analyses. When the multivariate normality assumption is met, the maximum likelihood (ML) parameter estimates are asymptotically efficient (see Salvalei, 2008). Thus, in this study, because the data met the multivariate normality assumption, ML estimation was used for all of the analyses.

LGC estimation typically follows a two-step procedure (Bollen & Curran, 2006), in which an unconditional LGC model is tested first, and then a conditional LGC model is tested. Thus, for the unconditional LGC, we tested two models, representing

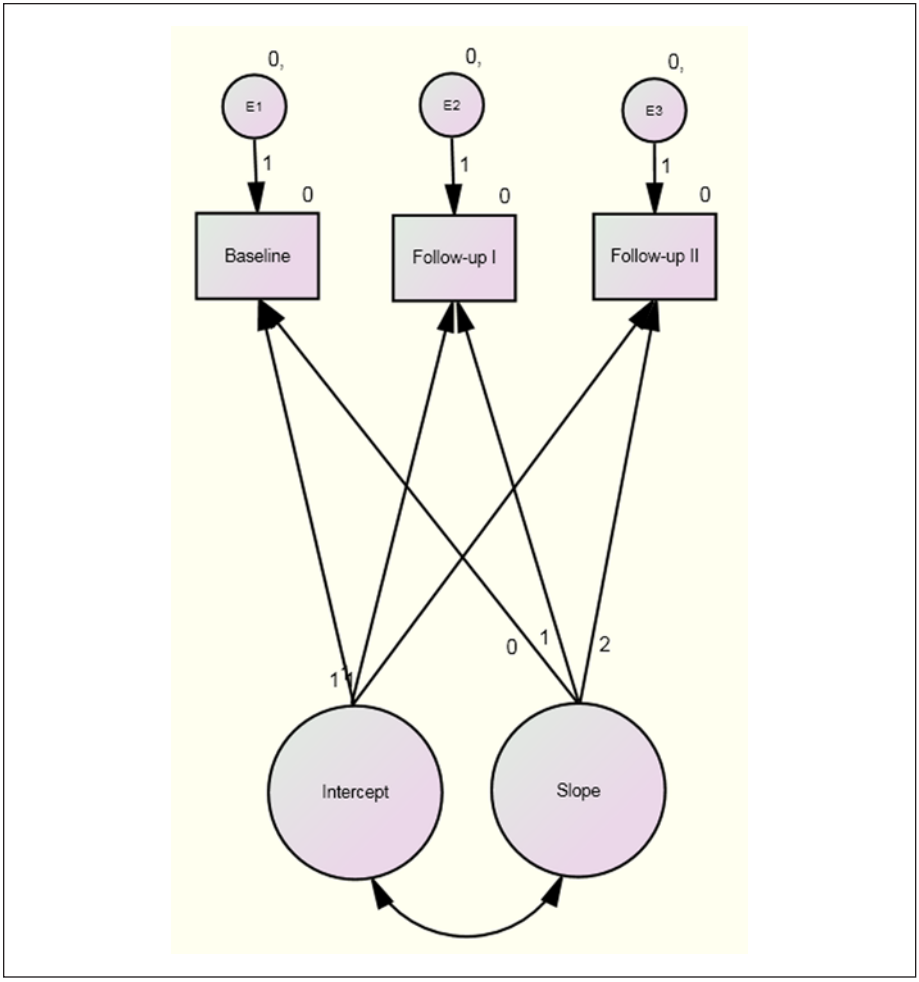

Figure 1. Unconditional latent growth curve model of the experimental group students' learning growth over 2 years.

possible forms of growth (i.e., linear change and quadratic change) to determine the overall shape of the individual change trajectories. A latent slope factor indicates the within-individual growth process. Linear growth is reflected by fixing the paths flowing from the latent slope to the PBA scores at (a) baseline, (b) a-year-follow-up, and (c) 2-year-follow-up to: 0, 1, and 2, respectively. A linear growth model fit the data better than a quadratic model. Figure 1 shows the theoretical LGC model of learning growth.

Traditionally, a fixed slope and intercept are assumed for the entire sample in a fixed effects model. However, in LGC models, the intercept is the initial level and the slope indicates the rate of growth, which is modeled as random effects. Thus, the

Table 2. Means and Standard Deviations for Baseline and First- and Second-Year Follow-Up Concept and Content Scores for the Experimental Group (n = 250).

PBA	Time	*M*	*SD*
Concept	Baseline	9.92	2.79
	Follow-up I	10.84	3.12
	Follow-up II	13.82	2.75
Content	Baseline	2.71	1.63
	Follow-up I	5.73	1.25
	Follow-up II	6.49	4.45

Note. PBA = performance-based assessment.

scores at baseline are assumed to be the initial value of the growth process that continuously develops over 2 years. A latent concept is modeled to gauge the initial level of knowledge. The initial level is modeled by constraining the paths from the latent intercept to the scores at three different times to 1.0 because if the initial level of knowledge matters, this effect should be constant during the whole growth process.

The mean of the slope in the LGC models indicates the average effect of learning growth; when mean values of the slope differ significantly from zero, within-individual variation in the linear growth process is indicated. The variance parameter of the intercept indicates the amount of heterogeneity of the initial concept and content. The mean of the intercept representing the average initial concept and content variables were fixed to zero to identify the model. Thus, the degree of between-individual variability is assessed, and statistically significant variability indicates the presence of variance that can be explained in subsequent modeling steps.

In the second conditional LGC, besides intercepts and slopes, explanatory covariates can be introduced following SEM principles. Gender, ethnicity, ability, and grade were used as covariates in this study to help explain which groups started higher than others, and which groups increased at a steeper gradient than other groups.

Unconditional LGC modeling for the experimental group. The LGC model of *Concept* and *Content* (see Figure 1) showed an excellent model fit for the data, a nonsignificant $\chi^2(1) = .11$ ($p = .745$), and the low root mean square error approximation (RMSEA = .001), high Comparative Fit Index (CFI = 1.00), and Tucker-Lewis Fit Index (TLI = 1.00) indicated a perfect fit for the data. In addition to the linear growth model above, a quadratic model was also hypothesized to examine whether a nonlinear model fits better than a linear model. To test for a quadratic effect of time, the quadratic parameters were set to 0, 1, and 4, respectively. The χ^2 differences were statistically significant, $\Delta\chi^2(0) = 46.10$, $p > .001$, and CFI differences were greater than .01, indicating that the quadratic effect provided a significantly worse fit. Thus, it can be concluded that the linear LGC model provided a better fit to the data.

The well-fitting linear model enabled us to review the substantive results of the analysis. Table 2 shows means and standard deviations for baseline and first- and second-year follow-up *Concept* and *Content* scores for the experimental group. In

most SEM analyses, regression parameters are the main interest for researchers, but in LGC analyses, the means of the intercept and slope latent variables, covariance between the slope and intercept, and variance of the slope and the intercept are the main interest. A statistically significant mean intercept (M intercept = 12.56, $p < .001$) showed that the students' mean scores at the baseline were greater than 0. This mean intercept estimate is not as important as the mean slope estimate, which is the rate of growth over time. The mean of slope that differs statistically significantly from 0 (M slope = 3.75, $p < .001$) indicated that the mean students' scores were increased by 3.75 points per year for the 3 month intervention period. Furthermore, the variances of the intercept and slope parameters were examined to determine whether individual differences existed in both the initial scores and the rates of growth. The nonsignificant variance parameter of the intercept (S^2 intercept = 1.25, $p = .48$) indicated that students' mean scores in *Concept* and *Content* at the baseline did not statistically significantly differ from each other. The nonsignificant variance parameter of the slope (S^2 slope = 0.89, $p = .47$) indicated that there was no significant difference in the mean rate of growth of the students' scores. The nonsignificant covariance between the intercept and slope ($Cov = .27$, $p = .83$) indicated that students' scores at the baseline did not influence the rates of growth of their scores. This is also supported by the low correlation coefficient ($r = .26$) between the intercept and the slope. Although there were no statistically significant differences in variance of the intercept or slope, further analyses using covariates were conducted to examine any possible between-individual differences of the intercept and slope due to higher power to detect slope variability when covariates are included.

Conditional *LGC Modeling for the Experimental Group*

Gender as a covariate. The standardized regression weight value of −.01 for intercept indicated that although the boys' absolute scores at the baseline were lower than girls', their scores were not significantly different ($p = .94$). The standardized regression weight value of −.26 for slope indicated that although the boys' scores were increased at lower rates than girls', the rates of growth across the 2 years were similar for both boys and girls ($p = .14$).

Ability as a covariate. Because most of the schools in this study did not identify gifted students; *high-ability* students (i.e., those who scored above 115 on the NNAT) were used instead of gifted students to examine whether other students benefit from the ICM as well as high-ability students. The standardized regression weight value of .69 for intercept indicated that although the high-ability students' scores at the baseline were higher than other students', their scores were not statistically significantly different from other students' ($p = .21$). Furthermore, the standardized regression weight value of .37 for slope indicated that although the high-ability students' scores were increased at higher rates than other students', their growth rates across the 2 years were not statistically significantly different from other students' ($p = .45$).

Ethnicity as a covariate. About 69% of the students were Caucasian, with the remaining students represented in one of several other ethnicity categories (e.g., African American, Asian, Hispanic, Native American, etc.). Given the relatively small percent of students represented in the various ethnic groups, Caucasian or non-Caucasian was used as another covariate. The standardized regression weight value of .46 for intercept indicated that although the Caucasian students' absolute scores at the baseline were higher than others', their scores were not significantly different from other students' ($p = .32$). The standardized regression weight value of −.42 for slope indicated that although the Caucasian students' scores increased at lower rates than other students' scores, their growth rates across the 2 years were not statistically significantly different from other students' ($p = .31$).

Grade as a covariate. Because 28% of participants were first graders, we wanted to examine the extent to which these younger students benefited from the ICM. Being first graders or nonfirst graders, therefore, was used as yet another covariate. The standardized regression weight value of −.93 for intercept indicated that the first graders' scores at the baseline were significantly lower than other students' ($p < .05$). However, the standardized regression weight value of 2.85 for slope indicated that the first graders' scores were significantly increased at higher rates than other students' scores across the 2 years ($p < .001$).

Discussion

Because the Project Clarion science units were intricately designed to support deep understanding of science concepts through a focus on multiple and mutually supportive skills and concepts, multiple measures including a standardized achievement test and PBAs were necessary to obtain accurate results about what students know and are able to do as a result of exposure to the ICM. After the 1-year intervention, as well as after the 2-year intervention, the experimental group performed better on the MAT and the TCT, indicating that both science achievement and critical thinking were improved among students with low SES exposed to the ICM.

Students' Growth in Concept and Content

The results also indicate that students' learning in *Concept* and *Content* through the ICM gradually increases over time. The mean scores in *Concept* and *Content* attainment among all participants were similar at the beginning of the ICM. Furthermore, the rates of growth of their scores were not influenced by their scores at the beginning of the ICM, indicating that the ICM benefits all of the students, regardless of their initial achievement levels. This is also confirmed by the subsequent analyses with gender, ability, ethnicity, and grade as covariates. The ICM benefited both boys and girls, both high-ability and other students, and both students who are Caucasian as well as students from other ethnic groups. The first graders' mean scores were lower than the second or third graders' mean scores at the beginning of the ICM, as expected.

However, their scores increased at a considerably faster rate than other grade-level scores. Across both years of the ICM, the youngest students in the study grew the most in both content attainment and concept mastery.

Conclusion

This study suggests that the ICM, designed according to exemplary science strategies with differentiation approaches for high-ability learners and instructional scaffolding for other learners, is effective when implemented students from all walks of life, regardless of grade, gender, ability level, and so on. Moreover, the study results suggest that the use of such curricula helps enhance students' reasoning skills and science achievement, as measured by standardized tests; advanced, complex science concepts and content, as measured by PBAs, were similarly enhanced for all learners.

Implications

The study represents an important step forward for the field of gifted education in crafting science interventions that work with a broad range of students, from the youngest ages on. Such curriculum also serves well the gifted learner. For example, students' performance on the macro-concept showed that primary-age students demonstrated significantly high learning gains and were able to transfer learning about a macro-concept, make generalizations, and understand rigorous scientific content when provided with satisfactory direct instruction. Schools need to recognize then that young children, including first graders, can benefit from more direct instructional time targeted on all aspects of learning science and in enhancing reasoning skills. The results from this study showed that greater growth occurred with the students who began using the ICM at earlier grades in their schooling.

The study also suggests that, when multiple types of understandings, learning, or skills are taught, multiple measures need to be implemented to examine all of the dimensions of learning. By using both traditional (i.e., standardized tests) and nontraditional assessments (i.e., PBAs), the ICM demonstrated enhanced student learning at young ages. For example, this study demonstrated the viability of using concept mapping (i.e., a component of the PBAs) as a valid and reliable method of assessing science understanding. It also showed changes in conceptual understanding and content attainment when administered as a pre- and post-assessment. The study also suggests that such PBAs allow room for growth across all levels of ability, as the open-ended nature of the PBAs prevented a ceiling effect.

Finally, this study suggests that LGC modeling can be used as an additional analysis to show the extent of growth from an intervention. LGC provides researchers with a tool for modeling longitudinal data, and allows researchers to examine patterns across multiple measurement points in time (Bollen & Curran, 2006). Furthermore, LGC allows researchers to explain students' development across a number of behavioral domains with increased complexity, which broadens the understanding of changes within the context of increasing larger and more complex

frameworks that mirror more closely the actual educational environment (e.g., Duncan & Duncan, 2004).

Areas for Further Investigation

The results of this study indicate several areas for potential investigation in education with respect to curriculum implementation and its impact on learning. It is necessary to know the nature and quality of professional development needed to ensure high-level implementation of innovative curricula. This study provided teachers with 2 days of initial professional development, follow-up, planning time, and additional professional development in subsequent years of the study. Professional development did not occur as a single isolated incident, but was sustained over time, building on teachers' knowledge base. By varying the amount of professional development provided administrators can determine how much professional development is necessary to ensure implementation fidelity at the most efficient cost for schools. It is also necessary to know what school climates and support structures are needed to foster and sustain innovation. While this study shows that student learning increases with increased teacher professional development, it did not investigate deeply the support structures of teacher and school characteristics that also influence student learning. Future studies should be conducted to investigate these possible support structures.

This study demonstrated significant and educationally important gains after a short interval of the ICM implementation. An important investigation would be to determine the possible longitudinal learning gains in all science, if all science curricula, instruction, and assessment were revamped to use the ICM as applied to all science.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author disclosed the receipt of the following financial support for the research, authorship, and/or publication of this article: Project Clarion was supported by the Jacob K. Javits Grant under grant number S206A020059 (for research support by U.S. Department of Education).

References

Adams, C. M., & Callahan, C. M. (1995). The reliability and validity of the performance task for evaluating science process skills. *Gifted Child Quarterly*, *39*, 14-20.

Arbuckle, J. L. (2008). *AMOS software*. Chicago, IL: SPSS.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: John Wiley.

Borman, G. D., & Hewes, G. M. (2002). The long-term effects and cost-effectiveness of success for all. *Success for All*, *24*, 243-266.

Boyer, P., Bedoin, N., & Honore, S. (2001). Relative contributions of kind- and domain-level concepts to expectations concerning unfamiliar exemplars: Developmental change and domain differences. *Cognitive Development*, *15*, 457-479.

Bracken, B. A., Bai, W., Fithian, E., Lamprecht, S., Little, C., & Quek, C. (2004). *Test of Critical Thinking*. Williamsburg, VA: The Center for Gifted Education. Retrieved from http://cfge.wm.edu/publications.htm

Byrne, B. B. (2010). *Structural equation modeling using AMOS: Basic concepts, applications, and programming* (2nd ed.). New York, NY: Routledge.

Chi, M. T. H., Hutchinson, J. E., & Robins, A. F. (1989). How inferences about novel domain-related concepts can be constrained by structural knowledge. *Merrill-Palmer Quarterly*, *35*, 27-62.

Cooper, C., Baum, S., & Neu, T. (2004). Developing scientific talent in students with special needs: An alternative model for identification, curriculum, and assessment. *Journal of Secondary Gifted Education*, *15*, 162-169.

Duncan, T. E., & Duncan, S. C. (2004). An introduction to latent-growth curve modeling. *Behavior Therapy*, *35*, 333-363.

Educate America Act, H.R. 1804, 103rd Cong. (1994). Retrieved from http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=103_cong_bills&docid=f:h1804pp.txt.pdf

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction* (Executive summary: "The Delphi Report"). Millbrae: California Academic Press. Retrieved from http://assessment.aas.duke.edu/documents/Delphi_Report.pdf.

Feng, A., VanTassel-Baska, J., Quek, C., O'Neill, B., & Bai, W. (2005). A longitudinal assessment of gifted students' learning using the integrated curriculum model: Impacts and perceptions of the William and Mary language arts and science curriculum. *Roeper Review*, *27*, 78-83.

Fowler, M. (1990). The diet cola test. *Science Scope*, *13*, 32-34.

Frayer, D. A., Frederick, W. C., & Klausmeier, H. J. (1969). *A schema for testing the level of concept mastery* (Tech. Rep. No. 16). Madison: The University of Wisconsin Press.

Gavin, M. K., Casa, T. M., Adelson, J. L., Carroll, S. R., Sheffield, L. J., & Spinelli, A. M. (2007). Project M3: Mentoring mathematical minds: Challenging curriculum for talented elementary students. *Journal of Advanced Academics*, *18*, 566-585.

Gonzales, P., Williams, T., Jocelyn, L., Roey, S., Kastberg, D., & Brenwald, S. (2008). *Highlights from TIMSS 2007: Mathematics and science achievement of U.S. fourth- and eighth-grade students in an international context* (NCES 2009–001 Revised). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Grigg, W., Lauko, M., & Brockway, D. (2006). *The nation's report card: Science 2005* (NCES 2006-466). Washington, DC: U.S. Government Printing Office.

Harcourt Educational Measurement. (2001). *The Metropolitan Achievement Tests–Eighth Edition (MAT-8)*. San Antonio, TX: Pearson.

Herberg-Davis, H., & Callahan, C. M. (2008). Gifted students' perceptions of advanced placement and international baccalaureate programs. *Gifted Child Quarterly*, *52*, 199-216.

Kim, K. H., VanTassel-Baska, J., Bracken, B. A., Feng, A., Stambaugh, T., & Bland, L. (2012). Project Clarion: Three years of science instruction in title I schools among K-Third grade students. *Research in Science Education*, *42*, 813-829. doi:10.1007/s11165-011-9218-5

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York, NY: Guilford.

Linn, M. C., & Songer, N. B. (1991). Cognitive and conceptual change in adolescence. *American Journal of Education*, *99*, 379-417.

Little, C. A., Feng, A. X., & VanTassel-Baska, J. (2007). A study of curriculum effectiveness in social studies. *Gifted Child Quarterly*, *51*, 272-284.

Liu, O., Lee, L., Hofstetter, C., & Linn, M. (2008). Assessing knowledge integration in science: Construct, measures, and evidence. *Educational Assessment*, *13*, 33-55. doi:10.1080/10627190801968224

Lubinski, D., & Benbow, C. P. (2006). Study of mathematically precocious youth after 35 years: Uncovering antecedents for the development of math-science expertise. *Perspectives on Psychological Science*, *1*, 316-345.

Mioduser, D., & Betzer, N. (2008). The contribution of project-based-learning to high-achievers' acquisition of technological knowledge and skills. *International Journal of Technology and Design Education*, *18*, 59-77.

Moon, T. R., Brighton, C. M., Callahan, C. M., & Robinson, A. (2005). Development of authentic assessments for the middle school classroom. *Journal of Secondary Gifted Education*, *16*, 119-133.

Nafiz, K. (2008). A student-centered approach: Assessing the changes in prospective science teachers' conceptual understanding by concept mapping in a general chemistry classroom laboratory research. *Science Education*, *38*, 91-110.

Naglieri, J. A. (1997). *Naglieri Nonverbal Ability Test*. San Antonio, TX: The Psychological Corporation.

National Center for Education Statistics. (2010). *Program for international student assessment (PISA): Overview*. Retrieved from http://nces.ed.gov/surveys/pisa/index.asp

National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.

National Research Council. (2002). *Learning and understanding: Improving advanced study of mathematics and science in U.S. high schools*. Washington, DC: National Academy Press.

Novak, J. D. (1998). *Learning, creating, and using knowledge: Concept maps as facilitative tools in schools and corporations*. Mahwah, NJ: Lawrence Erlbaum.

Paul, R. (1992). *Critical thinking: What every person needs to survive in a rapidly changing world*. Rohnert Park, CA: Center for Critical Thinking and Moral Critique, Sonoma State University.

Pine, K. J., & Messer, D. J. (2000). The effect of explaining another's actions on children's implicit theories of balance. *Cognition and Instruction*, *18*, 35-51.

Quinn, P. C., & Eimas, P. D. (1997). A reexamination of the perceptual-to-conceptual shift in mental representations. *Review of General Psychology*, *1*, 271-287.

Ramey, C. T., & Ramey, S. L. (1998). Early intervention and early experience. *American Psychologist*, *53*, 109-120.

Rayneri, L., Gerber, B., & Wiley, L. (2006). The relationship between classroom environment and the learning style preferences of gifted middle school students and the impact on levels of performance. *Gifted Child Quarterly*, *50*, 104-118.

Rutherford, F. J., & Ahlgren, A. (1991). *Science for all Americans*. New York, NY: Oxford University Press.

Salvalei, V. (2008). Is the ML chi-square ever robust to nonnormality? A cautionary note with missing data. *Structural Equation Modeling*, *15*, 1-22.

Schappe, J. F. (2005). Early childhood assessment: A correlational study of the relationships among student performance, student feelings, and teacher perceptions. *Early Childhood Education Journal*, *33*, 187-193.

Spektor-Levy, O., Eylon, B., & Scherz, Z. (2009). Teaching scientific communication skills in science studies: Does it make a difference? *International Journal of Mathematics and Science Education*, *7*, 875-903.

Swanson, J. D. (2006). Breaking through assumptions about low-income, minority gifted students. *Gifted Child Quarterly*, *50*, 11-25.

Taba, H. (1962). *Curriculum development, theory and practice*. New York, NY: Harcourt Brace.

Taba, H. (1966). *Teaching strategies and cognitive functioning in elementary school children*. San Francisco, CA: San Francisco State College. Retrieved from ERIC database. (ED025448)

Tali Tal, R., & Miedijensky, S. (2005). A model of alternative embedded assessment in a pull-out enrichment program for the gifted. *Gifted Education International*, *20*, 166-186.

VanTassel-Baska, J. (1986). Effective curriculum and instructional models for the gifted. *Gifted Child Quarterly*, *30*, 164-169.

VanTassel-Baska, J., Avery, L. D., Hughes, C. E., & Little, C. A. (2000). An evaluation of the implementation of curriculum innovation: The impact of William and Mary units on schools. *Journal for the Education of the Gifted*, *23*, 244-272.

VanTassel-Baska, J., Bass, G., Ries, R., Poland, D., & Avery, L. D. (1998). A national study of science curriculum effectiveness with high ability students. *Gifted Child Quarterly*, *42*, 200-211.

VanTassel-Baska, J., & Bracken, B. (2004). *Project Clarion: An integrative study of science teaching to young children*. Washington, DC: United States Department of Education.

VanTassel-Baska, J., Feng, A. X., & Brown, E. (2008). A study of differentiated instructional change over three years. *Gifted Child Quarterly*, *52*, 297-312.

VanTassel-Baska, J., Feng, A. X., & de Brux, E. (2007). A study of identification and achievement profiles of performance task-identified gifted students over six years. *Journal for the Education of the Gifted*, *31*, 7-34.

VanTassel-Baska, J., Feng, A. X., & Evans, B. L. (2007). Patterns of identification and performance among gifted students identified through performance tasks: A three-year analysis. *Gifted Child Quarterly*, *51*, 218-231.

VanTassel-Baska, J., Johnson, D., & Avery, L. D. (2002). Using performance tasks in the identification of economically disadvantaged and minority gifted learners: Findings from Project STAR. *Gifted Child Quarterly*, *46*, 110-123.

VanTassel-Baska, J., & Little, C. (Eds.). (2003). *Content-based curriculum for gifted learners*. Waco, TX: Prufrock Press.

VanTassel-Baska, J., & Little, C. (Eds.). (2011). *Content-based curriculum for gifted learner* (2nd ed.). Waco, TX: Prufrock Press.

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

Zeigler, E. F. (1995). Competency in critical thinking: A requirement for the "allied professional." *Quest*, *47*, 196-211.