The Achievement of Student Subgroups on Science Performance Assessments in Inquiry-Based Classrooms

Jerome M. Shaw
University of California, Santa Cruz

Sam O. Nagashima
University of California Los Angeles

Abstract

This study examined student learning in science as measured by performance assessments embedded within inquiry-based units of instruction. These locally developed assessments were implemented in a consortium of districts involved in a multi-year science education reform initiative. The sample consisted of scores from 834 fifth grade students on three performance assessments given in a participating district's 14 elementary schools during the 2004-2005 school year. District-provided data permitted disaggregation of student scores by ethnicity, gender, and socioeconomic status as well as identification as English Learner, Gifted and Talented, and Special Education. Using mean scores as the basis for comparison, results showed the majority of students achieving at the proficient level as defined by initiative-developed rubrics. Statistical analyses indicated significant underperformance on one or more of the assessments by Blacks, Hispanics, low socioeconomic status students, males, non-Gifted, and Special Education students. Depending on the performance assessment and the student subgroup, potential factors related to performance include science discipline and access to economics-related resources (e.g., computers). This study is noteworthy for its comprehensiveness and the nuanced understandings it brings to previously documented achievement gaps.

Correspondence should be addressed to Jerome M. Shaw at jmlshaw@ucsc.edu, or Sam O. Nagashima at samn@ucla.edu

Introduction

Measuring student achievement in science recently has garnered greater attention across the United States. This increase is due in part to the onset of the No Child Left Behind Act's (NCLB) requirement of statewide testing in science starting with the 2007-2008 school year (NCLB, 2002). Federal law now requires local education agencies to annually assess their students' learning in core academic subjects such as science in relation to content that is specified in state-sanctioned standards (U.S. Department of Education, 2002). Said standards routinely include concepts from multiple scientific disciplines such as biology and physics as well as process skills such as designing and conducting experiments.

While NCLB-mandated content standards provide broad guidance as to what should be assessed, state policymakers are left to decide how. Among the many issues to consider when choosing an assessment methodology are feasibility and compatibility. With respect to the former, factors such as time, material requirements and associated costs make selected response testing (e.g., multiple choice) more attractive than approaches such as performance assessment – simply defined as "assessments that allow students to demonstrate their understandings and skills… as they perform a certain activity" (National Research Council [NRC], 2001, p. 31) – which can be difficult as well as costly to develop and implement (Baker, 1997). Science performance assessments in particular may be up to 100 times more expensive than multiple-choice tests (Stecher, 1995) and three times more expensive than open-ended writing assessments (Stecher & Klein, 1997).

Perhaps more important than monetary cost is consideration of an assessment practice's degree of compatibility with the standards it is designed to measure. As the *National Science Education Standards* (NSES) state, "assessments provide an operational definition of standards, in that they define in measurable terms what teachers should teach and students should learn" (NRC, 1996, pp. 5-6). In clarifying this notion of teaching and learning, the NSES and its companion volume on classroom assessment (NRC, 2001), stress that a distinguishing feature of reform science teaching is its focus on engaging students in "active and extended scientific inquiry" (NRC 1996, p. 52). Such inquiry-based instruction leads to the concomitant building of content knowledge and development of process skills.

In theory, more so than selected response methods, performance assessments are especially well suited to measure the complex mix of conceptual understandings and science process skills associated with inquiry-based instruction (NRC, 2001, 2005; Shavelson, Baxter, & Pine, 1991). In practice, most NCLB-compliant tests rely heavily on selected response formats. However, the authors of *Inquiry and the National Science Education Standards* (NRC, 2000) fault multiple-choice tests for being too broad in coverage and focused on recognition and recall of facts, attributes which predispose them to pose "a serious obstacle" to inquiry-based teaching (p. 75). Thus, the appropriateness of selected response assessment for measuring the achievement of students taught inquiry-based science is questionable.

In contrast to the varying approaches to assessing student learning, proponents of current reforms are consistently resolute in asserting that inquiry-based science education is for all students (American Association for the Advancement of Science, 1989; NRC, 1996, 2000). Reform efforts funded by agencies such as the National Science Foundation work to realize this vision in part by aligning multiple elements of the educational enterprise including curriculum, instruction and assessment. One aspect of this alignment is the linking of inquiry-based curriculum and instruction with assessment that is performance-based.

Congruence between inquiry-based science instruction and performance assessment notwithstanding, several studies document student achievement in inquiry-based science classrooms solely using traditional forms of assessment, in part due to the

feasibility issues discussed above (Amaral, Garrison, & Klentschy, 2002; Geier et al., 2008; Johnson, Kahle, & Fargo, 2007; Lee, Deaktor, Hart, Cuevas, & Enders, 2005; Lynch, Kuipers, Pyke, & Szesze, 2005). Other studies of student learning in hands-on science classrooms do utilize performance assessments, either exclusively (Cuevas, Lee, Hart, & Deaktor, 2005; Lee, Buxton, Lewis, & LeRoy, 2006; Shaw, 1997) or in conjunction with traditional assessments (Pine et al., 2006). Still other studies use performance assessment to evaluate student learning in science without specifying the instructional context (Klein et al., 1997; Lawrenz, Huffman, & Welch, 2001).

This sparse literature base provides some indication of the extent to which the inclusive goals of science education reform are realized in terms of the performance of diverse student subgroups. While mixed, the combined findings from these various studies indicate achievement gaps by gender (females outscoring males), ethnicity (Whites outscoring non-whites), socioeconomic status or SES (high SES outscoring low SES), "giftedness" (Gifted students outscoring Non-Gifted), English proficiency (Non-English Learners outscoring English Learners). The studies noting students' native language show no clear pattern. These findings, accompanied with source information, are presented in Table I. Worth noting is that, of the various student subgroups reported on, there is a distinct omission of findings for students designated as Special Education, whether with traditional or performance assessments. Also absent from the literature are findings on Gifted students based on performance assessment. Finally, the most student demographic subgroups reported in any of these recent studies is five by Lee and colleagues in 2005.

Table I
Synthesized findings of student subgroup performance in inquiry-based science classrooms

	ASSESSMENT TYPE	
SUBGROUP	TRADITIONAL	PERFORMANCE
Gender	• Girls > Boys (3*, *6**) • Girls = Boys (4, 7, 10) • Boys > Girls (*5**)	• Girls > Boys (*5**, *6*, 8, 10*) • Girls = Boys (7, 10)
English Learner	• Non-EL > EL (1*, 7*)	• Exited EL > Non-EL (8)
Ethnicity	• Whites > Non-White (*5**, *6**, 7*, 9) • Whites = Non-Whites (4[1])	• Whites > Non-White (*5**, *6**, 7*, 9)
Gifted	• Gifted > Non-Gifted (7*)	• Not addressed
Native Language	• English > Non-English (7)	• English = Spanish (2) • Spanish > English (8)
Socioeconomic Status	• High SES > Low SES (7*, 9, 10*)	• High SES > Low SES (9, 10*)

Notes:

* = Finding was statistically significant

Italics = unspecified instructional context (all others are inquiry-based)

Code for Studies Cited:

1. Amaral et al. (2002)
2. Cuevas et al. (2005)
3. Geier et al. (2005)
4. Johnson et al. (2007)
5. *Klein et al. (1997)*
6. *Lawrenz et al. (2001)*
7. Lee et al. (2005)
8. Lee et al. (2006)
9. Lynch et al. (2005)
10. Pine et al. (2006)

Research Purpose and Questions

This study was undertaken to broaden the empirical literature base on student achievement in science classrooms as measured by performance assessment. We present an uncommonly comprehensive set of findings by addressing six categories of student demographic subgroups, including Gifted and Special Education, in a single study. Specifically, with respect to fifth grade student scores on three performance assessments

[1] Whites significantly outperformed Non-Whites in non-inquiry-based classrooms

implemented in inquiry-based science classrooms, this study investigated the following questions:

1. What are the patterns of performance for all students and specific subgroups of students? (e.g., comparison of mean scores for students of different ethnicities)
2. Are there statistically significant differences in the performance of student demographic subgroups?

Research Design

Context

The data for this study were drawn from one of several districts that participated in a recently completed multi-year, NSF-funded science education reform initiative known as STEP-uP (Science Teacher Enhancement Program unifying the Pikes Peak region – www.stepupscience.org). STEP-uP's efforts to improve student learning included the development of performance assessments that were embedded within inquiry-based curriculum units taught at grades kindergarten through five. As part of their involvement in STEP-uP, teachers in participating districts engaged in professional development on the science curriculum units and their associated assessments.

The focal district for this study, herein referred to as the Abacus School District, was chosen for its high degree of diverse students. We chose to focus on the fifth grade level in order to reduce potential interference from student lack of familiarity with performance assessment; in other words, to avoid an issue with "opportunity to test" (Shaw, 1997). Assuming prior instruction in the participating districts, fifth grade students were more likely to have previously encountered a STEP-uP science performance assessment than those in lower grades. We chose data from the 2004-2005 school year as it was the first time that all three performance assessments for the fifth grade science units were implemented.

Curriculum

STEP-uP affiliated districts utilize a carefully selected mosaic of research-based science curriculum units from a variety of sources including FOSS, Insights, and Science and Technology for Children. At the time of this study, schools in the Abacus School District taught three science units at the 5th grade level: Ecosystems, Food Chemistry, and Microworlds. All of these units are from the Science and Technology for Children (STC) curriculum developed by the National Science Resources Center (NSRC) at the Smithsonian Institute, and supported by the National Academy of Sciences (NSRC, 1991, 1994, 1996). STC is an inquiry-based curriculum designed to meet the NSES and shown to produce meaningful learning gains with culturally and linguistically diverse students (Amaral et al., 2002).

Assessment

The measures of student learning used in this study were the three STEP-uP-developed performance assessments that correspond to the three science units taught at the fifth grade level: Ecosystems, Food Chemistry, and Microworlds (we use these titles

to refer to the assessments as well). These assessments, which are based on and serve as replacements for lessons already present in the units, engage students in applying understandings and skills from prior lessons to new situations (e.g., for Food Chemistry, determining the nutritional value of previously untested snack foods using techniques and knowledge gained beforehand). While the assessments for Food Chemistry (which focuses on the selection of a nutritious snack) and Microworlds (which involves the observation of live microbes) replace end-of-unit lessons and serve as culminating activities, the assessment for Ecosystems (which includes the study of food webs) begins midway and continues to the end of the unit.

The assessments were developed by Design Teams composed of two to three classroom teachers with prior experience teaching the particular unit and a university scientist knowledgeable in the unit's science content. Design Team members were enrolled in a college level course led by an assessment development expert. They created the assessments and their accompanying manuals as part of the course requirements. Following initial development, the assessments underwent an iterative review and revision process that included pilot and field-testing in project-affiliated schools spread across all five STEP-uP districts. Efforts were made to have test sites reflect the student diversity of the participating districts in terms of ethnicity, socioeconomic status, Special Education, and English Learners. The full development cycle for each assessment spanned a three-year process of initial design (year one), pilot/field-testing (year two), and implementation (year three). Further details regarding the STEP-uP assessment development process can be found in Kuerbis and Mooney (2008).

For each assessment, the development process culminated with the creation of a manual containing information such as administration guidelines and correlations to national and state science content standards. Reading across manuals reveals a common focus on the science process skills of understanding scientific investigation and design as well as appropriate communication of the results from scientific investigations (STEP-uP, 2003; STEP-uP, 2004; STEP-uP, 2005).

The administration guidelines specify logistics such as requisite materials, student grouping (e.g., pair or small group), and the number, focus and suggested time length of class sessions. Expected completion times for this group of assessments range from two to seven hours over two to four sessions, each of which may range from 30 minutes to an hour or more. There is also variation with respect to the final product and/or performance on which students will be judged. For example, with Ecosystems students give an oral presentation explaining their poster on a particular ecosystem while Microworlds calls for submission of a written report.

The assessment manuals also contain black line masters of handouts needed for students to engage in the assessment. These include task sheets with instructions and rubrics for scoring student performance as well as previously scored samples of student work. Each assessment is accompanied by at least two rubrics. Returning to the example of Ecosystems, there are separate rubrics for the oral presentation and the poster on which the presentation is based. These and other key features of the assessments are provided in Table II.

STEP-up performance tasks incorporate self-assessment. Teachers are instructed to have students evaluate their own performance using slightly modified versions of the same rubrics used by teachers themselves. Teachers introduce these rubrics to the students at various stages in the assessment process (e.g., prior to creating the Ecosystems poster and while students prepare to give their oral presentation) so that the rubrics may guide student work. As part of completing an assessment, students are expected to rate their own performance using the rubrics' criteria.

Table II
Key features of the three 5th grade science performance assessments

	ECOSYSTEMS	FOOD CHEMISTRY	MICROWORLDS
Task	Research (e.g., using library books and the Internet) an ecosystem and create a poster that presents the relationships within it	Conduct physical and chemical tests on a variety of snack foods to determine the lack or presence of specific nutrients	Examine various water samples with a microscope and determine which is safest to drink
Primary Content Focus	Life science – relationships in ecosystems	Physical science – nutrient composition of foods	Life science – structure and function of microorganisms
Process Focus	Document-based research	Laboratory tests and data collection	Microscope use and observations
Rubrics (Assessment Foci)	• Ecosystem Poster • Oral Presentation	• Testing Chart • Nutrient Testing • Oral Interview	• Lab Work • Lab Report
Student Grouping	Pairs	Small Groups	Pairs
Class Sessions (Total time)	3 (5-7 hours)	4 (3-3.5 hours)	2 (2-3 hours)

Students

The 834 unique fifth grade students in the study sample were from 39 classrooms in the 14 elementary schools of the Abacus School District. District-provided data included information on individual students' membership in multiple demographic categories such as those required for disaggregation by NCLB. Listed in alphabetical order, the six student categories with their associated subgroups used in this study are as follows: English Learner, ethnicity (American Indian/Alaskan Native, Asian, Black, Hispanic, and White), gender (male, female), Gifted and Talented, socio-economic status or SES (based on status as a recipient of free or reduced lunch), and Special Education. While the categories English Learner and Special Education had multiple subgroups (e.g., limited English proficient and non-English proficient for the former; autism and physical disability for the latter), the small number of individuals in these categories warranted

their being collapsed into single variables for this study. Deeper analysis of the English Learner data is reported elsewhere (Shaw, 2009).

Scores

Student scores were derived from the application of STEP-uP developed rubrics to student products or performances such as posters and oral presentations. Teachers scored their own students' responses and submitted them to the school district office. The administration guidelines instruct teachers not to submit scores for students who missed 50% or more of the instruction associated with a particular science unit. Without specifying a particular approach (such as averaging), the guidelines likewise direct teachers to give students one overall score based on their respective scores for an assessment's two or three rubrics. Thus, individual student scores were in the form of single digit numbers on a four-point scale on which 4 = Advanced, 3 = Proficient, 2 = Partially Proficient, and 1 = Unsatisfactory. Although the rubrics contain task-specific criteria, these four performance levels are common to all three assessments.

The Abacus School District provided a combined total of 2,155 individual scores (one score per assessment per student, maximum three scores per student) that represent the full complement of teacher-submitted data from the 2004-2005 school year. Not all students completed all the assessments. For the study sample (n=834), 107 (12.8%) completed only one assessment, 136 (16.3%) completed two assessments, and the remaining 591 (70.9%) completed all three assessments. With respect to the individual assessments, completion rates are 727 (87.2%) for Ecosystems, 694 (83.2%) for Food Chemistry, and 731 (87.6%) for Microworlds. Completion rates for the subgroups in the study are shown in Table III.

Method

This exploratory, post-hoc investigation[2] employed two levels of analysis: descriptive comparisons of overall and student subgroup performance, and significance testing of subgroup differences. Raw score means on each assessment were the basis for the former while the latter were conducted using z-scores. Given their connection to rubric-defined performance levels, raw scores provide a "readily understood reference point" from which to understand the comparisons (Hoover, 1984, p. 13). Conversion of raw scores to z-scores for significance testing is an accepted practice for studies of this nature (see for example, Klein et al., 1997, and Pine et al., 2006) with sound psychometric backing (Binder, 1984; Jaccard & Wan, 1996; Kim, 1975; Labovitz, 1967, 1970; Zumbo & Zimmerman, 1993).

Multiple regression analyses were conducted to determine the existence of significant differences between the performances of the student subgroups in our sample. A unique linear regression analysis was run for each of the three assessments with white females who are non-English Learners, non-Free/Reduced Lunch, non-Special Education,

[2] This study was conducted near the end of funding for the focal project and was not part of that project's original design.

and non-Gifted and Talented serving as the basic comparison group. For each of these analyses, student demographics were the independent variables (e.g., Ethnicity, Gender, SES) with assessment scores (i.e., separate values for Ecosystems, Food Chemistry, and Microworlds) as the dependent variable. Given their relatively large sample sizes, additional analyses were run for the Ethnicity subgroups (e.g., Black, Hispanic). Dummy coding was used where each subgroup was given its own variable with the exception of White, which served as the comparison group. All other student demographic variables (i.e., Gifted and Talented, Male, English Learner, SES, and Special Education,) were dichotomously coded, 0 = non-member, 1 = member.

Table III
Completion rates for student groups on the three assessments

	Ecosystems	Food Chemistry	Microworlds
ENGLISH LEARNER			
English Learners	51	61	62
ETHNICITY			
American Indian/Alaskan Native	16	17	17
Asian	28	30	29
Black	199	178	196
Hispanic	247	242	250
White	235	225	237
GENDER			
Female	365	351	368
Male	362	343	363
GIFTED AND TALENTED			
Gifted	34	30	41
SOCIO-ECONOMIC STATUS			
Free/Reduced Lunch	488	467	502
SPECIAL EDUCATION			
Special Educational Needs	74	68	75
Total Sample	727	694	731

Note. The total sample includes 834 unique students.

Findings

We present our findings using the two research questions as frames of reference. First, we provide comparisons of mean scores for all students as well as student demographic subgroups on each of the three assessments (Question 1). These findings appear in order from macro to micro in terms of the level of student groupings: total sample, reference groups (e.g., English Learner and Non-English Learner), and ethnicity subgroups. These descriptive comparisons are followed by a presentation of findings

from the tests of significance for differences in subgroup performance (Question 2). This final section, titled multiple regression analyses, includes effect sizes and the degree to which student level variables explain the variance in the scores of each of the three assessments.

Total Sample Comparisons

With a variance of only .01 of a point on a scale of 1-4, mean scores on all three assessments were essentially identical for students as a whole. Mean scores for the total sample on the three assessments were: 2.80 (*SD = .837*) for Ecosystems, 2.81 (*SD = .858*) for Food Chemistry, and 2.80 (*SD = .804*) for Microworlds. We refer to this result as the "homogeneity of means" pattern, whose "universal mean" is taken to be 2.8 (see Table IV).

In terms of the levels used in the project-developed rubrics, overall student performance on the three assessments was in the "Proficient" range (2.50 – 3.49). There were no mean scores at the "Advanced" (3.50 – 4.0) or "Unsatisfactory" levels (0 – 1.49), and two mean scores at the "Partially Proficient" level (1.50 – 2.49): American Indian/Alaskan Native and Special Education on Food Chemistry, 2.47 and 2.22, (see Table V and Table IV, respectively).

Table IV
Mean scores and standard deviations for student groups on the three assessments

	Ecosystems M (S.D.)	Food Chemistry M (S.D.)	Microworlds M (S.D.)
Total Sample	2.80 (.837)	2.81 (.858)	2.80 (.804)
ENGLISH LEARNER			
Non-English Learner	2.80 (.844)	2.82 (.861)	2.81 (.806)
English Learner	2.82 (.740)	2.80 (.833)	2.69 (.781)
Difference	*0.02*	*0.02*	*0.12*
ETHNICITY			
White	2.87 (.754)	2.95 (.754)	2.89 (.766)
Non-White	2.77 (.873)	2.76 (.898)	2.75 (.818)
Difference	*0.10*	*0.19*	*0.14*
GENDER			
Female	2.92 (.808)	2.99 (.832)	2.92 (.810)
Male	2.69 (.851)	2.64 (.849)	2.67 (.779)
Difference	*0.23*	*0.35*	*0.25*
GIFTED AND TALENTED			
Non-Gifted	2.78 (.840)	2.79 (.856)	2.76 (.789)
Gifted	3.29 (.579)	3.47 (.629)	3.46 (.602)
Difference	*0.51*	*0.68*	*0.70*
SOCIO-ECONOMIC STATUS			
Non-Free/Reduced Lunch	2.92 (.747)	2.91 (.826)	2.92 (.810)
Free/Reduced Lunch	2.75 (.872)	2.78 (.870)	2.75 (.796)
Difference	*0.17*	*0.13*	*0.17*
SPECIAL EDUCATION			
Non-Special Educational Needs	2.85 (.813)	2.89 (.815)	2.87 (.762)
Special Educational Needs	2.39 (.934)	2.22 (1.01)	2.21 (.920)
Difference	*0.46*	*0.67*	*0.66*

Note. Maximum score = 4.

Other patterns worth noting are that the mean scores on all three assessments for females (2.92, 2.99 and 2.92, respectively for Ecosystems, Food Chemistry and Microworlds) and for Gifted and Talented students (3.29, 3.47 and 3.46, respectively for Ecosystems, Food Chemistry and Microworlds) were consistently above those for the total sample. The reverse was true for males (2.69, 2.64 and 2.67, respectively for Ecosystems, Food Chemistry and Microworlds) and students classified as Special Education (2.39, 2.22 and 2.21, respectively for Ecosystems, Food Chemistry and Microworlds). Mean scores for Non-white students as a group (2.77, 2.76 and 2.75,

respectively for Ecosystems, Food Chemistry and Microworlds) reflect the homogeneity of means pattern.

Reference Group Comparison

The subgroups and corresponding reference groups presented here are as follows (SUBGROUP / Reference Group): ENGLISH LEARNER / Non-English Learner, ETHNICITY / White, GENDER / Female, GIFTED AND TALENTED / Non-Gifted, SOCIOECONOMIC-STATUS / Non-Free/Reduced Lunch, SPECIAL EDUCATION / Non-Special Education (see Table IV). With the exception of the Gifted and Talented subgroup, each reference group outperformed its counterpart in nearly all cases. For example, mean scores on all three assessments for females were consistently above those for males (2.92:2.69, 2.99:2.64 and 2.92:2.67, respectively for Ecosystems, Food Chemistry and Microworlds). Although slight, the lone departure from this pattern was the mean score for English Learners which was marginally higher than that for Non-English Learners on the Ecosystems assessment only (2.80:2.82, respectively).

Ethnicity Subgroup Comparisons

Listed in alphabetical order, subgroups within the Ethnicity category are American Indian/Alaskan Native, Asian, Black, Hispanic, and White (see Table V). Ethnicity subgroup mean scores range from a high of 3.23 (Asians on Food Chemistry) to a low of 2.47 (American Indian/Alaskan Native on Food Chemistry). From high to low, the pattern on Ecosystems and Microworlds was: Asian, White, Hispanic, Black, American Indian/Alaskan Native. Blacks and Hispanics reversed their rankings on Food Chemistry, resulting in the following pattern: Asian, White, Black, Hispanic, American Indian/Alaskan Native.

Table V
Mean scores and standard deviations for Ethnicity subgroups on the three assessments

	Ecosystems M (S.D.)	Food Chemistry M (S.D.)	Microworlds M (S.D.)
ETHNICITY			
American Indian/Alaskan Native	2.50 (1.030)	2.47 (0.874)	2.53 (0.800)
Asian	3.14 (0.705)	3.23 (0.568)	3.17 (0.602)
Black	2.68 (0.972)	2.74 (0.864)	2.65 (0.936)
Hispanic	2.82 (0.778)	2.73 (0.938)	2.80 (0.719)
White	2.87 (0.754)	2.95 (0.754)	2.89 (0.766)
Total Sample	2.80 (0.837)	2.81 (0.858)	2.80 (0.804)

The reference group for all Ethnicity subgroups is White. Mean scores for Whites closely resemble the homogeneity of means pattern: 2.87/2.80, 2.95/2.82, 2.89/2.80, respectively, for Ecosystems, Food Chemistry and Microworlds. Whites were outperformed by Asians on all three assessments (2.87:3.14 / Ecosystems, 2.95:3.23 / Food Chemistry, 2.89:3.12 / Microworlds). Conversely, Whites outperformed all other Ethnicity subgroups on all three assessments.

Multiple Regression Analyses

Individual multiple regression analyses were conducted using z-scores as outcome variables and student background demographic information as the predictor variable to identify statistically significant differences in group performance. Results of theses analyses are presented in Tables VIa-c. On each of the three assessments, four to five subgroups showed statistically significant results. Common to all three assessments was the pattern of underperformance by males and Special Education students and overperformance by Gifted students. Assessment-unique instances of underperformance were low SES students on Ecosystems, Blacks and Hispanics on Food Chemistry, and Blacks on Microworlds.

Table VIa
Summary of regression coefficients for Ecosystems assessment

	ECOSYSTEMS		
Variables	B	SE B	β
Constant	.296	.085	
English Learner	-.006	.147	-.002
American Ind./Alaskan Ntv.	-.323	.252	-.047
Asian	.285	.194	.055
Black	-.146	.094	-.065
Hispanic	-.003	.095	-.001
Male	-.224	.072	-.112**
Gifted	.512	.177	.104**
SES	-.187	.078	-.088*
Special Education	-.468	.120	-.141***

*p<.05, **p < .01, ***p< .001

Table VIb
Summary of regression coefficients for Food Chemistry assessment

	FOOD CHEMISTRY		
Variables	B	SE B	β
Constant	.414	.086	
English Learner	.071	.137	.020
American Ind./Alaskan Ntv.	-.372	.239	-.058
Asian	.256	.184	.052
Black	-.199	.095	-.087*
Hispanic	-.274	.094	-.131**
Male	-.346	.072	-.173***
Gifted	.641	.177	.130***
SES	-.097	.078	-.045
Special Education	-.688	.122	-.205***

*p<.05, **p < .01, ***p< .001

With respect to gender, females outperformed their male counterparts by an average of .271 of a point on the three assessments: .224 on Ecosystems, .346 on Food Chemistry, and .223 on Microworlds. These differences were statistically significant at the p<.01 level for Ecosystems and at the p<.001 level for Food Chemistry and Microworlds.

Regarding ethnicity, two groups underperformed relative to their white counterparts: Blacks underperformed by -.199 on Food Chemistry and -.241 on Microworlds, and Hispanics by -.274 on Food Chemistry. These differences were significant at the p<.05, p<.001, and p<.01 levels, respectively. The Ecosystems assessment showed no significant difference in student performance based solely on ethnicity.

Table VIc
Summary of regression coefficients for Microworlds assessment

	MICROWORLDS		
Variables	B	SE B	β
Constant	.315	.083	
English Learner	-.156	.134	-.043
American Ind./Alaskan Ntv.	-.348	.237	-.052
Asian	.245	.180	.049
Black	-.241	.091	-.107***
Hispanic	-.072	.091	-.035
Male	-.243	.069	-.122***
Gifted	.718	.153	.164***
SES	-.141	.077	-.065
Special Education	-.701	.114	-.215***

*p<.05, **p < .01, ***p< .001

Low SES students underperformed relative to their high SES peers by -.187 on Ecosystems. This difference was significant at the p<.05 level. They also underperformed on Food Chemistry and Microworlds, however these differences were not significant.

Students classified as Special Education underperformed relative to their non-Special Education counterparts with a difference of -.468 on Ecosystems, -.688 on Food Chemistry, and -.701 on Microworlds, for an average of -.619. Each of these differences was significant at the p<.001 level.

Conversely, students classified as Gifted and Talented outperformed non-Gifted and Talented students an average of .624 on all three assessments: .512 on Ecosystems, .641 on Food Chemistry, and .718, on Microworlds. These differences were significant at the p<.01, p<.001, and p<.001 levels, respectively.

Overall, student level variables explained only a small proportion of variance in the scores for all three assessments: Ecosystems, $R^2 = .062$; Food Chemistry, $R^2 = .120$; and Microworlds, $R^2 = .127$. In general, less than 12% of the total variability in student scores is accounted for by student level variables. Estimates of effect size (f^2) suggest marginal effect due to English Learner status alone and a very small effect due to Ethnicity alone (see Table VII).

Table VII
Effect sizes of models

	ECOSYSTEMS	
	Adj. R^2	Effect size (f^2)
Complete Model[a]	.062	.066
Ethnicity Only[b]	.012	.012
English Learner Only[c]	.000[d]	-
	FOOD CHEMISTRY	
	Adj. R^2	Effect size (f^2)
Complete Model[a]	.120	.136
Ethnicity Only[b]	.026	.027
English Learner Only[c]	.000d	-
	MICROWORLDS	
	Adj. R^2	Effect size (f^2)
Complete Model[a]	.127	.138
Ethnicity Only[b]	.021	.022
English Learner Only[c]	.016	.016

Note. Effect sizes (f^2) of 0.02, 0.15, and 0.35 are considered small, medium, and large, respectively (Cohen, 1988).

[a]Includes all student level demographic variables (English Learner, Ethnicity, Gender, Gifted and Talented, SES, and Special Education)

[b]Includes only Ethnicity variables (American Indian/Alaskan Native, Asian, Black, Hispanic, and White)

[c]Includes only English Learner variables (EXIT, LEP, NEP)

[d]Models non-significant at $p < .01$

Discussion and Implications

The purpose of this exploratory study was to shed light on student learning in inquiry-based science classrooms. Given this context, our post-hoc analysis used results from three locally developed, curriculum-embedded, fifth grade performance assessments to inform an understanding of the achievement of multiple NCLB-accountability related student subgroups. Within these parameters, our research questions essentially ask: To what degree are students learning science?, and, Are there appreciable differences in the level of science learning based on group affiliation? As seen in the previous section and discussed below, the answers are mixed. Interpretations of student scores analyzed in this study need to be made with at least two considerations in mind. Although reported for individual students, the scores are (a) based on collaborative work (i.e., students worked in pairs or small groups on the assessments), and (b) represent teachers' appraisal of student overall proficiency on the various constructs measured (through unspecified processes, teachers' global scores were based on the two or three rubric-based scores

students received per assessment). Therefore, the scores are relevant for discussing groups of students, not individuals.

This study is limited in scope; the data come from only one of the initiative's participating districts for one school year at one grade level. Future studies can address this shortcoming by expanding the data set along each of those dimensions (i.e., multiple districts, school years, and grade levels). An additional track along which to amplify the current study's scope would be to compare findings from the performance assessments with those from other measures of science achievement. A logical candidate for such a comparison is student scores on statewide science tests. While such a test was not in operation at the time of the study, one has been implemented since. Studies including scores from the two sources would need to consider factors such as the comparability or lack thereof in the constructs measured by each assessment.

An additional concern is the lack of comprehensive validity evidence to support the interpretation of scores from the assessments. In his discussion of validity issues pertinent to performance assessments, Messick (1996) notes the importance of "transparency," meaning that "the performance standards are understood and facilitate learning" (p. 13). This issue is partially addressed by the assessments' administration guidelines that call for teachers to share and discuss the scoring guidelines or rubrics with the students before, during and after the assessment. However, the degree to which students understand the rubrics and such teacher-student interactions facilitate learning is beyond the purview of this study. Moreover, evidence stands to be gathered along the six aspects of Messick's "unified" concept of construct validity: content, substantive, structural, generalizability, external, consequential (1996, p. 7). While an ongoing and evolving process, obtaining such evidence is important if the assessments are to be used past the existence of STEP-uP and outside the project's geographic boundaries.

Bearing those factors in mind and turning to our results, we found that, in broad terms, the universal mean of 2.8 – or "Proficient" when rounded to 3.0 – indicates that students as a whole exhibited desirable levels of science knowledge and skills. It is encouraging to note that this level of achievement held constant on all three assessments, indicating comparable performance across the different content, process, and assessment foci (see Table II). For example, students as a whole were proficient at demonstrating knowledge of ecosystems, nutrition, and microorganisms through graphic, oral, and written means, respectively.

Parsing the total sample, we uncovered patterns of underperformance for specific student subgroups on each assessment (see Table VIII). We applied two criteria to put these findings into perspective. First, we identified those differences that were statistically significant. Second, we invoked a criterion of "practical" significance meaning that, following standard conventions for rounding to the nearest whole number, the mean scores of comparison groups translated to different levels of performance on the rubric. It should be noted that there are no student groups that meet the practical significance criterion that do not also meet the criterion for statistical significance. Thus, applying them in the stated order does not eliminate any group from being identified as underperforming.

Application of the above two criteria yields a more focused appraisal of student subgroup underperformance. As shown in Table VIII, statistically significant differences were observed with respect to gender, Gifted, and Special Education across all three assessments. In addition, assessment-specific underperformance (discussed below) was as follows: low SES students on Ecosystems, Black and Hispanic students on Food Chemistry, and Black students on Microworlds. However, rounding student group mean scores to the nearest whole number and applying the practical significance criterion narrowed the field of underperformers to non-Gifted and Special Education students. These were the only two groups meeting both significance criteria and they did so on all three assessments. In relative terms, the degree of practical significance was not great – only one rubric level (i.e., Proficient versus Partially Proficient, the next lowest level). Separation by non-adjacent levels (e.g., Proficient vs. Unsatisfactory) would present cause for greater concern.

Table VIII
Significant underperformance by student groups on the three assessments

ECOSYSTEMS	FOOD CHEMISTRY	MICROWORLDS
---	Black	Black
---	Hispanic	---
Male	Male	Male
Non-Gifted	*Non-Gifted*	*Non-Gifted*
Low SES	---	---
Special Education	*Special Education*	*Special Education*

Note. All student groups listed showed statistically significant differences from regression analyses (see Table VI). Groups in italics also showed practical significance (defined as having a rubric-based performance level lower than the corresponding reference group).

Our documentation of the underperformance of non-Gifted students and Special Education students as measured by performance assessments in science classrooms, while new to the literature, is not surprising. It is likely attributable to the achievement disparities inherent in the definition of those categories. Nevertheless, studies could be undertaken to discern whether or not performance assessments can enlighten our understanding of the nature of these gaps and how they might be addressed in inquiry-based science classrooms.

Significance ratings aside, important nuances are present in the achievement patterns observed in this study. Many of our findings corroborate those of prior studies. The over-performance of females, Whites, and high SES students in relation to their respective counterparts are reflective of findings by Geier and colleagues (2005), Johnson and colleagues (2007), and Lee and colleagues (2006). Conversely, the lack of difference in performance between English Learners and Non-English Learners runs counter to results reported by Amaral and colleagues (2002) as well as Lee and colleagues (2005). The small size of the English Learner sample in our study limits the power of this

finding. Future studies with larger populations are needed to determine the veracity of this claim. While not available in our data set, analyses incorporating English Learner's native language (see Cuevas et al., 2005 and Lee et al., 2006) would further clarify the nature of the complex relationship between language proficiency and student achievement.

Likewise contrary to published results (Lee et al., 2005, Lynch et al., 2005, and Pine et al., 2006) is the lack of significant performance differences between high and low SES students on two of the three performance assessments in our study. With counts in the hundreds (low SES close to 500 and high SES near 200), small sample size does not appear to be an issue here. The fact that a statistically significant difference arose only on the Ecosystems assessment may be related to the nature of that task. Given the long-term, document-based research focus of the task (conceivably including out of school time), student performance arguably might be influenced by economic factors such as the quantity and quality of relevant library reference materials as well as access to computers and the Internet. In contrast, the other two assessments are dependent on resources previously used during instruction (e.g., nutrient testing equipment for Food Chemistry and microscopes for Microworlds). Further study noting the particular location of different SES students (e.g., high SES versus low SES schools) may help clarify this ambiguous finding.

Our findings in relation to ethnicity raise similar assessment-specific concerns. Except for the performance of American Indians/Alaskan Natives on Food Chemistry (mean -2.47 equivalent to "Partially Proficient"), all Ethnicity subgroups performed at the "Proficient" level (i.e., means within the range of 2.5 – 3.4). In comparison to Whites, statistically but not practically significant differences were found for Blacks and Hispanics on Food Chemistry and for Blacks alone on Microworlds (see Table VIII). It is worth investigating whether or not the design of or scoring systems for the Food Chemistry and Microworlds assessments somehow disadvantage these subgroups. Given the wide variability of individuals within ethnic groups it is unproductive to speculate on potential explanations for these observed differences. Research employing techniques such as focus groups or think aloud protocols have the potential to provide insights on these important yet perplexing results (see for example, Martiniello, 2008).

More sense can be made of the findings in relation to gender. Gender gaps have been shown to be sensitive to assessment type and content orientation. While the general pattern is one of girls outperforming boys, Klein and colleagues (1997) found that boys outperformed girls on a multiple-choice test. However, that same study found that girls outperformed boys on performance assessments, a pattern our findings uphold. In a study using scores from four different performance assessments, Pine and colleagues (2006) found comparable gender performance on physical science tasks while girls outperformed boys on the one life science task. However, in our study girls outperformed boys on both life (Ecosystems and Microworlds) and physical science (Food Chemistry) tasks. This apparent contradiction might be explained by the strong connection of the particular physical science content to the life sciences on the Food Chemistry assessment (the chemical determination of nutrients in food was tied directly to nutritional value for human consumption). Future research should explore the persistence of this discipline-

specific bias with performance assessments. Bearing in mind that different assessment formats measure different competencies (Lawrenz et al., 2001; Shavelson et al., 1991), careful attention to the particular constructs measured by the assessments yielding these results would be a necessary step to make further sense of these outcomes.

In sum, our findings bring greater comprehensiveness and additional insights to the understanding of student performance in inquiry-based science classrooms. This study is noteworthy for its inclusion of six student demographic subgroup variables (more than any recent study on the topic) and the nuanced understandings it brings to previously documented achievement gaps, gender in particular. While not intending to provide definitive answers or explanations, we point to potentially fruitful areas of further research on this issue. With their close alignment to the precepts of reform-based instruction, continued studies employing performance assessments have important contributions to make to the ultimate goal of attaining high levels of achievement for all students.

Acknowledgments

This study was conducted with support from National Science Foundation grant ESI-0196127. The opinions expressed herein are those of the authors and not necessarily those of the funding agency. The authors wish to acknowledge STEP-uP Principal Investigators Linda Mooney and Paul Kuerbis as well as STEP-uP Assessment Coordinator Cynthia Pechacek for sharing their time, expertise, and materials in support of this study.

References

Amaral, O.M., Garrison, L., & Klentschy, M. (2002). Helping English learners increase achievement through inquiry-based science instruction. *Bilingual Research Journal 26*(2), 213-239.

American Association for the Advancement of Science. (1989). *Science for all Americans.* New York: Oxford University Press.

Baker, E. (1997). Model-based performance assessment. *Theory Into Practice, 36*(4), 247-254.

Binder, A. (1984). Restrictions on statistics imposed by method of measurement: Some reality, some myth. *Journal of Criminal Justice*, *12,* 467-481.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed.). Hillsdale, NJ: Erlbaum.

Cuevas, P., Lee, O., Hart, J., & Deaktor, R. (2005). Improving science inquiry with elementary students of diverse backgrounds. *Journal of Research in Science Teaching, 42*(3), 337-357.

Geier, R., Blumenfeld, P. C., Marx, R. W., Krajcik, J. S., Fishman, B., Soloway, E., & Clay-Chambers, J. (2008). Standardized test outcomes for students engaged in inquiry-based science curricula in the context of urban reform. *Journal of Research in Science Teaching, 45*(8), 922-939.

Hoover, H. D. (1984). The most appropriate scores for measuring educational development in elementary schools: GE's. *Educational Measurement: Issue and Practice, 3*(4), 8-14.

Jaccard, J., & Wan, C. K. (1996). *LISREL approaches to interaction effects in multiple regression.* Thousand Oaks, CA: Sage Publications.

Johnson, C. C., Kahle, J. B., & Fargo, J. D. (2007). A study of the effect of sustained, whole-school professional development on student achievement in science. *Journal of Research in Science Teaching, 44*(6), 775-786.

Kim, J. O. (1975). Multivariate analysis of ordinal variables. *American Journal of Sociology*, *81,* 261-298.

Klein, S. P., Jovanovic, J., Stecher, B. M., McCaffrey, D., Shavelson, R. J., Haertel, E., Solano-Flores, G., & Comfort, K. (1997). Gender and racial/ethnic differences on performance assessments in science. *Educational Evaluation and Policy Analysis, 19*(2), 83-87.

Kuerbis, P. J., & Mooney, L. B. (2008). Using assessment design as a model of professional development. In J. Coffey, R. Douglas, & C. Stearns (Eds.),

Assessing science learning: Perspectives from research and practice, (pp. 409-426). Arlington, VA: National Science Teachers Association Press.

Labovitz, S. (1967). Some observations on measurement and statistics. *Social Forces*, *46,* 151-160.

Labovitz, S. (1970). The assignment of numbers to rank order categories. *American Sociological Review*, *35*(3), 515-524.

Lawrenz, F., Huffman, D., & Welch, W. (2001). The science achievement of various subgroups on alternative assessment formats. *Science Education, 85*(3), 279-290.

Lee, O., Buxton, C., Lewis, S., & LeRoy, K. (2006). Science inquiry and student diversity: Enhanced abilities and continuing difficulties after an instructional intervention. *Journal of Research in Science Teaching, 43*(7), 607-636.

Lee, O., Deaktor, R. A., Hart, J. E., Cuevas, P., & Enders, C. (2005). An instructional intervention's impact on the science and literacy achievement of culturally and linguistically diverse elementary students. *Journal of Research in Science Teaching, 42*(8), 857-887.

Lynch, S., Kuipers, J., Pyke, C., & Szesze, M. (2005). Examining the effects of a highly rated science curriculum unit on diverse students: Results from a planning grant. *Journal of Research in Science Teaching, 42*(8), 912-946.

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review, 78*(2), 333-368.

Messick, S. (1996). Validity of performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1-18). Washington, DC: US Government Printing Office, Report No. NCES-96-802. Available from ERIC Document Reproduction Service, no. ED 399 300.

National Research Council. (1996). *National science education standards.* National Committee on Science Education Standards and Assessment. Washington, DC: National Academy Press.

National Research Council. (2000). *Inquiry and the national science education standards: A guide for teaching and learning.* Washington, DC: National Academy Press.

National Research Council. (2001). *Classroom assessment and the National Science Education Standards.* Committee on Classroom Assessment and the *National Science Education Standards*. J. Myron Atkin, Paul Black, and Janet Coffey (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council. (2005). *How students learn: History, mathematics, and science in the classroom.* Committee on *How People Learn*, A Targeted Report for Teachers, M.S. Donovan and J.D. Bransford, Editors. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Science Resources Center. (1991). *Microworlds teacher's guide.* Burlington, NC: Carolina Biological Supply.

National Science Resources Center. (1994). *Food chemistry teacher's guide.* Burlington, NC: Carolina Biological Supply.

National Science Resources Center. (1996). *Ecosystem's teacher's guide.* Burlington, NC: Carolina Biological Supply.

No Child Left Behind Act of 2001, Public Law No. 107-110, 115 Stat. 1425 (2002).

Pine, J., Aschbacher, P., Roth, E., Jones, M., McPhee, C., Martin, C., Phelps, S., Kyle, T., & Foley, B. (2006). Fifth graders' science inquiry abilities: A comparative study of students in hands-on and textbook curricula. *Journal of Research in Science Teaching, 43*(5), 476-484.

Shavelson, R. J., Baxter, G.P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education, 4*(4), 347-362.

Shaw, J. M. (1997). Threats to the validity of science performance assessments for English language learners. *Journal of Research in Science Teaching, 34*(7), 721-743.

Shaw, J. M. (2009). Science performance assessment and English learners: An exploratory study. *Electronic Journal of Literacy Through Science, 8*(3).

Stecher, B. M. (1995, April). *The cost of performance assessment in science: The RAND perspective.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Stecher, B. M., & Klein, S. P. (1997). The cost of science performance assessments in large-scale testing programs. *Educational Evaluation and Policy Analysis, 19*(1), 1-14.

STEP-uP (Science Teacher Enhancement Program unifying the Pikes Peak Region). (2003). *Embedded assessment package for ecosystems.* Colorado Springs, CO: Author.

STEP-uP (Science Teacher Enhancement Program unifying the Pikes Peak Region). (2004). *Embedded assessment package for food chemistry.* Colorado Springs, CO: Author.

STEP-uP (Science Teacher Enhancement Program unifying the Pikes Peak Region). (2005). *Embedded assessment package for microworlds.* Colorado Springs, CO: Author.

U.S. Department of Education, Office of Elementary and Secondary Education, *No Child Left Behind: A Desktop Reference*, Washington, D. C., 2002.

Zumbo, B. D., & Zimmerman, D. W. (1993). Is the selection of statistical methods governed by level of measurement? *Canadian Psychology*, *34*, 390-399