Running head: Science Performance Assessment and English Learners

Science Performance Assessment and English Learners: Results from an Elementary Reform Initiative

Assistant Professor Jerome M. Shaw*, Doctoral Student Sam O. Nagashima[1]**

[1]CRESST/UCLA

Paper presented at the annual meeting of the National Council on Measurement in Education, April 10, 2007; Chicago, IL

Abstract

This post-hoc exploratory analysis examines patterns of student learning as measured by locally developed performance assessments. The assessments are embedded within inquiry-based units of instruction implemented in a multi-year, multi-district, NSF-funded science education reform initiative. The sample consisted of scores from 834 fifth grade students on three performance assessments given in one of the initiative's participating district's 14 elementary schools during the 2004-2005 school year. Mean scores were used as the basis for comparison. The results indicate that, across all three assessments, the majority of students achieved at the proficient level as defined by initiative-developed rubrics. The mean scores of English Learners overall were essentially the same as those of their non-English Learner peers (less than .02 point difference on a 4 point scale). Comparisons between non-English Learners and English Learner subgroups (Exit, Limited English Proficient or LEP, Non-English Proficient or NEP) revealed that only NEP students showed significant difference ($p<.01$) in performance relative to their non-English Learner counterparts on one of the three assessments. Overall, student level demographic variables explained only a small proportion of the variance in the scores for all three assessments. The results indicate the efficacy of the initiative's reform model which includes aligned curriculum, instruction and assessment along with coordinated teacher professional development on each of those components. The results also lend support to the use of performance assessments over selected response assessments as viable measures of inquiry-based science, especially for English Learners.

Science Performance Assessment and English Learners:
Results from an Elementary Reform Initiative

This study was motivated by the confluence of three significant factors on the US K-12 educational landscape: an emphasis on inquiry-based science instruction, the value of performance assessment for measuring student learning in such contexts, and the rise of English Learners as a segment of the school age population. The first two factors speak to a desirable alignment between instruction and assessment while the third acknowledges current and projected demographic realities.

Seminal documents such as the *National Science Education Standards* (National Research Council, 1996) tout the need for science teaching that centrally involves "guiding students in active and extended scientific inquiry" (p. 52). Such inquiry-based science instruction fosters student learning of conceptual understanding and sophisticated process skills. Given their propensity to favor "recognition and recall," the authors of *Inquiry and the National Science Education Standards* (National Research Council, 2000) note the potential of traditional assessments, such as multiple-choice tests, to "pose a serious obstacle to inquiry-based teaching" (p. 75). Conversely, performance assessment's ability to tap into complex thinking and skills leads many to regard it as being more closely aligned with the learning outcomes associated with inquiry-based science (Kim, Park, Kang & Noh, 2000; Lee, 1999; National Research Council, 2000).

Consideration of the students being assessed in inquiry-based science classrooms leads to the issue of linguistic diversity. English Learners (Non-native speakers of English who are developing their proficiency in English) constitute a sizeable and growing portion of the US K-

12 student population. The website for the National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs (NCELA) states the following regarding English Learners in US public elementary and secondary schools: their number has more than doubled in the past fifteen years (school years 1989/1990-2004/2005), their rate of enrollment has increased at nearly seven times the rate of total enrollment; in 2004/2005 (most recent year for which data are listed) their estimated number totaled nearly 5,120,000 – approximately 10.5% of the total public school enrollment – a figure that represents a 56.2% increase over that reported for school year 1994/1995 (NCELA, 2007). With continued immigration and the ongoing growth of the US Hispanic population, these trends are expected to continue into the near future.

The rhetoric of current reform exhorts that high quality science education is for all students, including English Learners (American Association for the Advancement of Science, 1989; National Research Council, 1996). Institutions such as the National Science Foundation have supported this vision by funding large-scale efforts designed to implement inquiry-based science, particularly with diverse or underserved student populations. Reform efforts strive to alignment between instruction and assessment, often by linking inquiry-based curriculum and instruction with performance assessment. Given the above demographic data and the inclusive nature of many science education reform efforts, one can infer that in contexts where inquiry-based science instruction is measured via performance assessment, English Learners are increasingly engaging in such assessments.

Using performance assessments to measure student learning in inquiry-based science classrooms holds promise and peril. As discussed above, many consider performance assessments to be more congruent with the teaching practices called for in current conceptions of educational reform. Noting the bias in traditional assessments, proponents of performance assessment argue its potential to narrow achievement gaps between ethnic, socioeconomic, and gender groups (Lee, 1999). Nevertheless, challenges associated with the use of performance assessments include difficulty and costliness in development and implementation (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). Moreover, studies have found that while girls tend to have higher overall mean scores than boys, patterns of achievement gaps among ethnic and socioeconomic groups are generally the same with both performance assessment and traditional tests (Klein et al., 1997; Lee, 1999; Lynch, 2000; Stecher & Klein, 1997). As Lee (1999) concluded, "The limited available research does not support the contention that performance assessment is more equitable with diverse students than traditional multiple-choice tests" (p. 103).

A similar conclusion can be drawn when focusing on assessment of English Learners in content areas such as science. In fact, there is evidence indicating inequitable aspects of performance assessments when used with this student population. Studies have documented significant links between students' level of proficiency in English and their performance on content-based assessments (Abedi, 2002, Abedi & Lord, 2001; Abedi, Lord, Hofstetter, & Baker, 2000). Understandably, low levels of proficiency in English have been shown to correspond with low levels of achievement. However, confounding variables include the linguistic demands of the assessment which may be mitigated by the use of accommodations such as customized dictionaries (Abedi, 2001).

When looking exclusively at science performance assessment with English Learners, the research base is quite sparse. Nevertheless, there are indications of similar linkages between English proficiency and student scores (see for example Shaw, 1997). Once again, the limited findings are insufficient to determine a consensus on the relative merit of using performance assessments to measure the achievement of English Learners in inquiry-based science classrooms.

The present study was conducted to address these gaps in the literature. While not comparative (i.e., traditional versus performance) in nature, the purpose of the study was to investigate a context in which performance assessments were indeed being used to measure the achievement of English Learners who in fact were taught science via inquiry-based instruction. Accordingly, this post-hoc analysis examined the learning of fifth grade English Learners in inquiry-based science classrooms as measured by a set of classroom-based performance assessments. Specifically, the study investigated the following questions:

Regarding 5th grade students' scores on three classroom-based science performance assessments:

1. What are the patterns of performance for all students, for English Learners in general, and for English Learners at different levels of proficiency in English?
2. What are key similarities and differences in the patterns of performance for the above groups?
3. To what extent does status as an English Learner impact student performance?

Research Design

Context. This study is based on the work of students taught science by teachers who were participants in a recently completed multi-year, multi-district, NSF-funded science education reform initiative known as STEP-uP (Science Teacher Enhancement Program unifying the Pikes Peak region). STEP-uP efforts to improve student learning included the development of performance assessments integrated with inquiry-based curriculum units taught at each grade level, K-5. A prominent feature of the STEP-uP model is the engagement of participating teachers in professional development on the curriculum units as well as the assessments. This study focused on the scores of 5th grade students in a single STEP-uP district, referred to as the Abacus School District (ASD), during the 2004-2005 school year.

Assessments. STEP-uP-developed performance assessments were used to measure student learning for the three science units taught at the 5th grade level: Ecosystems, Food Chemistry, and Microworlds. These end-of-unit assessments engage students in applying previously learned knowledge and skills to novel situations (e.g., researching and reporting on ecological aspects of a not yet studied endangered species). The assessments were designed to incorporate the content and procedures presented in previous lessons and were integrated into the units as part of the standard course of instruction. In fact, the assessments are based on existing lessons in the kits. Oftentimes this adaptation centers on adding scoring guides or rubrics to the elements in the lesson.

The performance assessments were developed as part of an "embedded assessment package" that includes constructed response assessments and rubrics for scoring both performance and constructed response assessments. These "assessment manuals" include samples of student responses collected during the development process. All of the assessments and supporting documents are provided in English only.

The assessments were developed by design teams that included two to three classroom teachers with prior experience teaching the particular kit and a college/university scientist knowledgeable in the kit's science content. As part of the development process, design teams enrolled in a college level course led by an assessment development expert. Design Teams created the assessments and their accompanying manuals as part of the course. Following initial development, the assessments underwent and iterative review and revision process that included pilot and field-testing in project-affiliated schools (all five of the participating 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 latter group was distinctly under-represented. In total, the entire development process spanned two years.

Students. The study sample consists of 834 fifth grade students from 39 classrooms in the 14 elementary schools of Abacus School District (ASD). Within the subgroup English Learners, students are classified into the following mutually exclusive sub-designations (listed in order of increasing proficiency in English): Non-English Proficient (NEP), Limited English Proficient (LEP), reclassified less than one year as Fluent English Proficient (EXIT), one year as Fluent English Proficient (EXIT1), two years as Fluent English Proficient (EXIT2), and three years as Fluent English Proficient (EXIT3). Note that our sample did not include any EXIT2 students. Of the 834 total students in the sample, 68 (8.2%) are classified as English Learners. The distribution of English Learners by sub-designation includes 10 (1.2%) NEP, 47 (5.6%) LEP, 2 (0.2%) EXIT, 8 (1.0%) EXIT 1, and 1 (0.1%) EXIT 3.

Scores. Teachers used STEP-uP developed rubrics to score their own students' responses. These assessment-specific rubrics use a common 4-point scale with 4 = Advanced, 3 = Proficient, 2 = Partially Proficient, 1 = Unsatisfactory. Data provided by ACD included individual student scores on each of the three performance assessments (Ecosystems, Food Chemistry, and Microworlds) as well as student level variables including gender, ethnicity (including sub-designations American Indian/Alaskan Native, Asian, Black, Hispanic, and White), English Learner status (including the sub-designations NEP, LEP, etc.), Free/Reduced Lunch status, Special Educational Needs status (including Autism, Multiple Disabilities, etc.), and Gifted and Talented status (see Table 1 for completion rates of all student groups on the three assessments).

Not all students completed all the assessments. For the total sample (n=834), 107 (12.8%) completed only one assessment, 136 (16.3%) completed at least two assessments, and the remaining 592 (70.9%) completed all three assessments. For English Learners (n=68), 6 (8.8%) completed only one assessment, 18 (26.5%) completed at least two assessments, and the remaining 44 (64.7%) 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. English Learners' assessment-specific completion rates are 54 (79.4%) for Ecosystems, 61 (89.7%) for Food Chemistry, and 63 (92.6%) for Microworlds.

Method. In order to explore whether there are significant differences between the performances of different groups of students in our sample, a multiple regression analysis was conducted using student demographics as the independent variables (e.g., Gender, Socioeconomic status (using free/reduced lunch as an indicator) and Special Education status) and student assessments scores (i.e., Ecology, Food Chemistry, and Microworlds) as the dependent variable. 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, and non-Gifted and Talented serving as the basic comparison group.

For the subgroups English Learner and Ethnicity, analyses were run using the sub-designation categories (e.g., LEP for English Learner and Hispanic for Ethnicity). Given the limited numbers of those students in our sample, analyses involving students identified by the English Learner sub-designation EXIT, EXIT1, and EXIT3 were collapsed into a single variable "EXIT." Thus, for analytical purposes, the English Learner subgroup is composed of the following three sub-designations: Exit, LEP and NEP. Each sub-designation under Ethnicity was given its own variable with the exception of White, which served as the comparison group. All other variables (i.e., Male, SES, Special Education, Gifted and Talented, and the English Learner sub-designations) were dichotomously coded, 0 = non-member, 1 = member.

Findings

In this section we first present the mean scores of the total sample and selected subgroups using the total sample as the comparison group. Next we provide findings for selected subgroups with their associated reference group (e.g., English Learners and Non-English Learners). These results are followed by findings for English Learner sub-designations (i.e., Exit, LEP, and NEP). The section closes with findings from our regression analyses including general observations on the source of variance within our sample. Mean scores on each of the three performance assessments served as the basis of comparison throughout.

Total Sample Comparisons. With a variance of only .01 points on a scale of 1-4, mean scores on all three assessments were essentially identical for all students on all three assessments. For the total sample, mean scores 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. Similarly, mean scores for non-English Learners as a whole closely matched those of the total sample: 2.80 (*SD = .844*) for Ecosystems, 2.83 (*SD = .861*) for Food Chemistry, and 2.81 (*SD = .806*) for Microworlds. Also close in value were the mean scores for the aggregate group of non-English Learners: 2.82 (*SD = .837*) for Ecosystems, 2.80 (*SD = .858*) for Food Chemistry, and

2.69 (*SD = .804*) for Microworlds. This result is hereon referred to as the "homogeneity of means" pattern. For ease of comparison, the "universal mean" for this pattern can be taken as 2.80.

Table 2 presents the above findings along with mean scores for additional subgroups, such as gender and ethnicity, provided for informational purposes (analyses with these groups is the topic of another study). Some patterns worth noting here 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 (3.29, 3.47, and 3.46 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 Comparisons. The subgroups and corresponding reference groups presented here are as follows: GENDER/Female, ENGLISH LEARNER/Non-English Learner, ETHNICITY/White, SOCIOECONOMIC-STATUS/Non-Free/Reduced Lunch, SPECIAL EDUCATION NEEDS/Non-Special Education Needs, GIFTED AND TALENTED/Non-Gifted (see Table 2). 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 higher than that for Non-English Learners on the Ecosystems assessment only (2.80:2.82, respectively).

English Learner Sub-designation Comparisons. Listed in order of proficiency in English from high to low, sub-designations within the English Learner subgroup are Exit, Limited English Proficient (LEP) and Non-English Proficient (NEP). Among these sub-designations, mean scores on all three assessments correspond to level of proficiency in English. Using Ecosystems as an example, mean scores were 3.11, 2.82 and 2.50, respectively for Exit, LEP and NEP English Learners (see Table 3 for mean scores for all English Learner subgroups on all three assessments).

The reference group for all English Learner sub-designations is Non-English Learner. Mean scores for LEP English Learners resemble the homogeneity of means pattern (2.82:2.80, 2.74:2.82 and 2.77:2.81, respectively for Ecosystems, Food Chemistry and Microworlds). Deviations from the homogeneity of means pattern emerge when looking at the other two sub-designations within this category. Exit English Learners outscored Non-English Learners on all three assessments (3.11:2.80, 3.18:2.82 and 3.00:2.81, respectively for Ecosystems, Food Chemistry and Microworlds). Mean scores for NEP English Learners were consistently below those for Non-English Learners on all three assessments (2.50:2.80, 2.63:2.82 and 1.88:2.81, respectively for Ecosystems, Food Chemistry and Microworlds). Mean scores and standard deviations for all these comparisons are provided in Table 3.

Multiple Regression Analyses. Results of the regression analyses are presented in Table 4. When considering English Learner sub-designations, NEP was the only classification to show statistically significant difference in performance on one of the three assessments. Using the standard $p<.01$, students designated as NEP exhibited significantly lower performance on the

Microworlds assessment relative to the population (including Exit and LEP students) with a mean difference of -.817.

With respect to gender, females outperformed their male counterparts by an average of .230 point on the three assessments: .191 on Ecosystems, .298 on Food Chemistry, and .200 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.

As for ethnicity, two groups underperformed relative to their white counterparts: Blacks by -.171 on Food Chemistry and -.190 on Microworlds, and Hispanics by -.237 on Food Chemistry. These differences were significant at the $p<.05$, $p<.01$, and $p<.01$ levels, respectively.

Low socioeconomic status (SES) students (i.e., those classified as Free/Reduced Lunch), underperformed relative to their high SES peers by -.153 on Ecosystems. This difference was significant at the $p<.05$ level.

Students classified as Special Education underperformed relative to their non-Special Education counterparts with a difference of -.381 on Ecosystems, -.587 on Food Chemistry, and -.558 on Microworlds, for an average of –509. Each of these differences was significant at the $p<.001$ level. Correspondingly, students classified as Gifted and Talented outperformed non-Gifted and Talented students an average of .520 on all three assessments: .438 on Ecosystems, .553 on Food Chemistry, and .568, on Microworlds.

Overall, student level demographic 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.

Discussion

This study explored 5th grade students' scores on three science performance assessments embedded within inquiry-based units of instruction. Students were taught the units, which included taking the assessments, by teachers in schools of one of five districts that participated in a multi-year, National Science Foundation supported science education reform project. As researchers external to the project, we set out to answer the following questions:

1. What are the patterns of performance for all students, for English Learners in general, and for English Learners at different levels of proficiency in English?
2. What are key similarities and differences in the patterns of performance for the above groups?
3. To what extent does status as an English Learner impact student performance?

In a nutshell, the answer to these questions can be stated as follows: Overall, status as an English Learner was not a predictor of student performance on the assessments as a whole. As a general rule, compared with all students taken together, all English Learners and all non-English Learners performed equally well on each of the three assessments.

Of interest was the emergence of a "universal mean" of 2.8 points (on a scale of one to four) across all of the above groups. Rounding to the nearest whole number (i.e., 3), this translates to the "proficient" level according to the rubrics by which the assessments were scored. An unsurprising nuance to this pattern is that, on average, English Learners with higher degrees of proficiency in English scored higher than their less English proficient peers (i.e., Exit English Learners scored higher than Limited English proficient (LEP) English Learners, who in turn scored higher than Non-English Proficient (NEP) English Learners). However, on all except the Microworlds assessment – and here, only for NEP English Learners – these differences were so small (i.e., within half a point of each other) as to be negligible. Rounding again to the nearest whole number, all English Learner subgroups achieved at the proficient level. The exception again is NEP English Learners who scored at the next highest level, namely “partially proficient,” on the Microworlds assessment. This difference was significant and signals an area for further investigation. For example, might the language demands placed on students by the Microworlds assessment be more challenging than those of either Ecosystems or Food Chemistry?

Moving beyond proficiency in English, our results likewise indicate that, with few exceptions, socioeconomic status (SES) and ethnicity played insignificant roles as predictors of student performance. Instances of underperformance were observed with low SES students on Ecosystems, with Hispanics on Food Chemistry and with Blacks on both Food Chemistry and Microworlds. Conversely, predictable, but problematic nonetheless, patterns of underperformance were observed for males, Special Education students and non-Gifted and Talented students on all three assessments. The disparity was greatest for the latter two groups, particularly on Food Chemistry and Microworlds where differences of at least half a point were observed. Might there be aspects of gender and cultural bias on the assessments?

Summary. The results of this study indicate that the performance assessments and their concomitant inquiry-based science instruction leveled the playing field as a whole for students with respect to proficiency in English, ethnicity, and socioeconomic status. Statistically significant, yet practically insignificant differences were observed for Blacks on the Food Chemistry assessment and for low SES students on the Ecosystems assessment. Ethnicity had a noteworthy negative impact only for Blacks on Microworlds and for Hispanics on the Food Chemistry and Microworlds assessments. Gender, Special Education status, and Gifted and

Talented status appear to be an important predictors of student performance on all three assessments, with females favored over males and non-Special Education and Gifted students favored over their counterparts.

These findings need to considered in the light of the following limitations: (a) the assessments have not undergone any formal validity study, (b) student responses were scored by their respective teachers with no information on inter-rater reliability, (c) demographic data do not include information on students' native languages (e.g., English Learners may perform differently based on the degree of difference between the alphabets of their native language and that of English), (d) demographic data lack finer distinctions of students' ethnic group affiliations (e.g., Hispanics grouped as a whole versus identification as Mexican American, Cuban, Puerto Rican, etc.), and (e) there is no indication of the degree to which students actually engaged in the learning activities on which the assessments are based.

That said, the curriculum-embedded nature of the assessments may be a contributing factor to the positive findings noted above. As Lee (1999) states: "Achievement gaps among ethnic, socioeconomic, and gender groups tend to be larger on items that call on outside-of-school knowledge and experiences" (p. 100). Our results reflect the converse of this pattern; that is, assessments based on inside-of-classroom knowledge and experiences tend to narrow achievement gaps among ethnic, socioeconomic, and gender groups. Moreover, our findings support the veracity of this pattern for English Learners in particular.

Conclusion

As one of the few studies documenting English Learners' performance on science performance assessments in the context of inquiry-based instruction, this study provides valuable insights on purportedly equitable practices in science education. Contrary to common patterns of underperformance, the findings indicate that when their inquiry-based science learning is measured with performance assessments, English Learners can and do exhibit levels of achievement comparable to their native English-speaking peers. Contextual factors such as a coherent curriculum, assessments aligned with and embedded within this curriculum, and coordinated teacher professional development on the curriculum, instruction and assessment are likely important contributing factors to this positive outcome. Further study is required to disentangle these interconnected elements so as to provide finer grained guidance to the educators of America's increasingly diverse student population.

References

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometric issues. *Educational Assessment, 8*(3), 231-257.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219-234.

Abdei, J., Lord, C., Hofstetter, C. & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3), 16-26.

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.

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

Kim, E., Park, H., Kang, H., & Noh, S. (2000, April). *Developing a framework for science performance assessment.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, New Orleans.

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.

Lee, O. (1999). Equity implications based on the conceptions of science achievement in major reform documents. *Review of Educational Research, 69*(1), 83-115.

Lynch, S. (2000). *Equity and science education reform.* Mahwah, NJ: Lawrence Earlbaum Associates.

National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs (no date). Retrieved March 9, 2007 from http://www.ncela.gwu.edu/.

National Research Council. (1996). *National science education standards.* 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.

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.

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.

TABLE 1. Completion rates for student groups on the three assessments.[1] (N)

	Ecosystems	Food Chemistry	Microworlds
GENDER			
Female	365	351	368
Male	362	343	363
ENGLISH LEARNER			
Exit[2]	9	11	11
LEP	34	42	43
NEP	8	8	8
ETHNICITY[3]			
American Indian/Alaskan Native	16	17	17
Asian	28	30	29
Black	199	178	196
Hispanic	247	242	250
White	235	225	237
SOCIO-ECONOMIC STATUS			
Free/Reduced Lunch	488	467	502
SPECIAL EDUCATIONAL NEEDS			
Special Educational Needs[4]	74	68	75
GIFTED AND TALENTED			
Gifted	34	30	41
Total Sample	727	694	731

[1] Total sample includes 834 unique students.

[2] Includes Exit, Exit1, Exit3

[3] Two students declined to report ethnicity

[4] Includes Autism, Multiple Disabilities (MD), Perceptual or Communicative Disability (PCD), Physical Disability (PD), Significant Identifiable Emotional Disability (SIED), Significant Limited Intellectual Capacity (SLIC), Speech-language Disability (S/L).

TABLE 2. Difference in mean scores (absolute) for student groups on the three assessments. (Range = 1-4)

	Ecosystems		Food Chemistry		Microworlds	
	M	s.d.	M	s.d.	M	s.d.
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)*	
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)*	
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 EDUCATIONAL NEEDS						
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)*	
GIFTED AND TALENTED						
Non-Gifted	2.78	.840	2.79	.856	2.76	.798
Gifted	3.29	.579	3.47	.629	3.46	.602
Difference	*(0.51)*		*(0.68)*		*(0.70)*	
Total Sample	2.80	.837	2.81	.858	2.80	.804

TABLE 3. Mean scores and standard deviations for student groups on the three assessments.

	Ecosystems		Food Chemistry		Microworlds	
	M	s.d.	M	s.d.	M	s.d.
GENDER						
Female	2.92	.808	2.99	.832	2.92	.810
Male	2.69	.851	2.64	.849	2.67	.779
ENGLISH LEARNER						
Exit	3.11	.333	3.18	.603	3.00	.632
LEP	2.82	.673	2.74	.857	2.77	.611
NEP	2.50	1.20	2.63	.916	1.88	1.25
ETHNICITY						
American Indian/Alaskan Native	2.50	1.03	2.47	.874	2.53	.800
Asian	3.14	.705	3.23	.568	3.17	.602
Black	2.68	.972	2.74	.864	2.65	.936
Hispanic	2.82	.778	2.73	.938	2.80	.719
White	2.87	.754	2.95	.754	2.89	.766
SOCIO-ECONOMIC STATUS						
Free/Reduced Lunch	2.75	.872	2.78	.870	2.75	.796
SPECIAL EDUCATIONAL NEEDS						
All Special Education	2.39	.934	2.22	1.00	2.87	.762
GIFTED AND TALENTED						
Gifted	3.29	.579	3.47	.629	3.46	.596
Total Sample	2.80	.837	2.81	.858	2.80	.804

TABLE 4. Summary of regression coefficients for all three assessments.

	Ecosystems			Food Chemistry			Microworlds		
Variables	B	SE B	β	B	SE B	β	B	SE B	β
Constant	3.05	.072		3.18	.073		3.08	.068	
Male	-.191	.061	-.114**	-.298	.062	-.174***	-.200	.056	-.124***
Exit	.263	.276	.035	.453	.250	.066	.156	.233	.024
LEP	-.007	.149	-.002	-.003	.136	-.009	-.092	.126	-.027
NEP	-.304	.291	-.038	.002	.290	.002	-.817	.271	-.106**
Indian/Alaskan	-.270	.211	-.047	-.319	.205	-.058	-.274	.190	-.051
Asian	.238	.163	.055	.226	.157	.054	.199	.149	.048
Black	-.124	.079	-.066	-.171	.082	-.087*	-.190	.073	-.105**
Hispanic	-.006	.080	-.003	-.237	.081	-.132**	-.044	.074	-.026
SES	-.153	.066	-.086*	-.084	.067	-.046	-.107	.062	-.062
Special Education	-.381	.101	-.138***	-.587	.105	-.204***	-.558	.093	-.211***
Gifted	.438	.144	.111**	.553	.152	.131***	.568	.127	.157***

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

TABLE 5. Effect sizes of models.

	Ecosystems		Food Chemistry		Microworlds	
	Adj. R^2	Effect size (f^2)	Adj. R^2	Effect size (f^2)	Adj. R^2	Effect size (f^2)
Complete Model[1]	.062	.066	.120	.136	.127	.138
Ethnicity Only[2]	.012	.012	.026	.027	.021	.022
EL Only[3]	.000^	-	.000^	-	.016	.016

[1] Includes all variables (all sub-designations for Gender, English Leaner, Ethnicity, SES, Special Education, and Gifted and Talented)

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

[3] Includes only English Learner variables (Exit, LEP, NEP)

^Models non-significant at p<.01

Note: An effect size (f^2) of 0.02, 0.15, and 0.35 are considered small, medium, and large, respectively (Cohen, 1988).