# THE IMPORTANCE OF USING MULTIPLE MEASURES TO ASSESS SCIENCE PROFICIEINCY

Patrick Enderle
Learning Systems Institute
The Florida State University

Jonathon Grooms
Learning Systems Institute
The Florida State University

Victor Sampson
School of Teacher Education and FSU-Teach
Florida State University

## Introduction

Scientific proficiency, as described by Duschl, Schweingruber, & Shouse (2007), encompasses a variety of knowledge and skills required by an individual to function effectively in an increasingly complex, information-driven society. The framework of scientific proficiency describes science as "both a body of knowledge and an evidence-based, model-building enterprise that continually extends, refines, and revises knowledge" (p. 2). In this view, individuals that are proficient in science: (a) know, use, and can interpret scientific explanations of the natural world; (b) can generate and evaluate scientific explanations and arguments; (c) understand the nature and development of scientific knowledge; and (d) can participate in the practices and discourse of the various scientific disciplines in a productive manner.

The emerging Common Core State Standards for Science (NRC, 2011), developed and supported by a majority of states, move beyond a primary concern for content knowledge to encompass performance expectations that reflect ideas described in the science proficiency framework. With the advent of this performance aspect and these standards becoming a driving force in future science teachers' professional realities, science teachers and educators must begin the work of creating methods for assessing the development of science proficiency. One of the greatest challenges in working towards such assessments is creating ones that reflect the multi-faceted nature of science proficiency.

Some researchers rely on a well-known standardized instrument, such as the National Assessment of Educational Progress (NAEP) or the Trends in International Mathematics and Science Study (TIMSS), as a way to document student achievement. Unfortunately, the use of these types of standardized assessments, which were not designed to measure the multi-dimensional nature of science proficiency, will result in a limited and often biased view of the overall competency of the students. The style of multiple-choice questions typically employed in these assessments target specific concepts, necessitating students to recall certain definitions or characteristics of the concept. Higher quality assessments will challenge the student further by asking them to use the idea to predict changes in scenarios provided or describe why certain events happen. However, as these assessments are normally multiple-choice selections, an assemblage of answers is already composed from which students choose. Correct selections are used as evidentiary proxies demonstrating students' science proficiency. However, those answer choices are limited in providing evidence of a student's ability to know and use scientific explanations at most. These assessment items provide no insight into a student's ability to generate a scientific argument or productively engage in the practices and discourses of the scientific community.

Another approach often found in the research literature involves focusing solely on one aspect of science proficiency and developing an assessment that targets only a single outcome related to that aspect for the purposes of a study. However, this approach does not represent the complex events that we understand learning, especially in science, to be (Donovon, Bransford, & Pelligrino, 2000; NRC 2011). The aspects of science proficiency, although capable of being studied separately, represent a heavily intertwined collection of knowledge and skills. Therefore, isolating aspects out for study takes away from developing a comprehensive understanding of how overall science proficiency develops. As these practices and skills are so heavily intertwined with each other, assessments aiming to fully measure science proficiency must be designed to engage students in more complex experiences that represent the construct's nature. When only one outcome is privileged at the expense of others, researchers and policy makers run the risk of making decisions about which instructional strategies to implement, curriculum to adopt, or funding priorities based on an incomplete picture.

Single assessments or inappropriate assessments can over estimate or under estimate students' scientific proficiency depending on the nature of the assessment. Thus, we argue that an assessment system that incorporates multiple tasks, rather than relying on a single instrument, will be needed in order to provide a valid assessment of science proficiency. A coherent and well-designed assessment system that uses multiple tasks will be able to target the various aspects of science proficiency in a more comprehensive fashion, including the hard to assess dimensions, such as participation in the practices and discourse of science, and will result in a more comprehensive picture of student knowledge, abilities, and habits of mind. The research presented here describes several assessments used in conjunction to measure students' development of science proficiency over the course of a year in middle school life science and physical science.

## Context of the Study

The assessments described here were tested during year one of a larger, three-year project aimed at refining the Argument Driven Inquiry instructional model (Sampson, Grooms, & Walker, 2011) and assessing students' improvements in science proficiency as a result of experiencing ADI-based instruction (IES Grant #: R205A100909). This work took place in middle school life science and middle school physical science courses offered at a research K12 school. One teacher taught the all the life science classes and one teacher was responsible for the physical science classes. Both of these teachers were involved in the development of the ADI activities in the previous summer, of which 12 were successfully implemented in the life science course and 8 ADI activities implemented in the physical science course.

The life science teacher was beginning her first year of teaching and was working through an alternative certification program as she had a bachelor's degree in biology. The physical science teacher has been teaching in several contexts for 16 years, working at the research K12 school for 11 years. His degree was originally in Elementary Education, but he had previously obtained his state level certification to teach secondary level science. Life science is the primary science course for 7th grade students at the school. Physical science is the primary science course for 8th grade, which also is the grade when the state mandated science assessment for AYP calculations is administered. The collections of assessments for the two courses were administered at the beginning and end of the school year to at least 78 students for life science and 76 students for physical science. Efforts were made to include all data in the analysis, however due to parent consent and attendance issues, only students with an allowable and complete set of pre and post data for an assessment were included in the analysis. The batteries of assessments given were course specific in nature, dealing with the content and ideas addressed in those courses, however, the structure of the assessments was similar, which is described in general below. Table 1 identifies which assessments were used to measure the development of certain aspects of science proficiency.

**Table 1**: Aspects of Science Proficiency and Associated Assessment

| Science Proficiency | Description | Assessment Instrument |
|---|---|---|
| Aspect 1 | Students know, use, and can interpret scientific explanations of the natural world | Content Knowledge Assessment |
| Aspect 2 | Students can generate and evaluate scientific explanations and arguments | Performance Task - Argument Generation Section |
| Aspect 3 | Students understand the nature and development of scientific knowledge | SUSSI |
| Aspect 4 | Students productively participate in the practices and discourse of the scientific community | Performance Task - Investigation Design Section<br>Scientific Writing Assessment |

### *Assessment Instruments*

*Content Knowledge*: The content knowledge assessment measured students' abilities to know and use scientific explanations of the natural world. The assessment is comprised of eight free response questions, each related to one of several "Big Topics" in Life Science and Physical Science, as determined by the teachers and researchers. National and state level standards documents were used in developing the Big Ideas. Each question includes an opening paragraph that provides a relevant scenario or context, followed by two questions. One question asks the student to *describe* the fundamental science concept (*Know*) and the other asks the student to *apply* that concept to the scenario provided (*Use*). The rubric for this assessment was developed from answers provided for the questions by an expert chemist and expert biologist with experience with K12 science education. A students' score was developed from the rubric based on correct description of several content elements identified in the expert's answer to the question.

*Scientific Writing*: The scientific writing assessment was developed to assess students' abilities to generate and evaluate scientific arguments. This assessment provides a student with a small amount of background information and a related data table followed by a prompt. The prompt presents an argument by a scientist who provides an inaccurate yet plausible explanation for the data. The students are directed to respond to the scientist's claim by generating an argument in support of a countering claim, which includes evidence and a rationale based on the data and information provided in the question, being mindful of writing style and grammar. The rubric, with an overall possible score of 28 points, was divided into three subscales: *Argument Structure* focusing on the inclusion of fundamental argument components including claims, evidence, and rationale (6); *Argument Content* concerning the quality and relevance of the argument components with respect to scientific discourse (10); and *Mechanics* regarding the punctuation, grammar, and technical quality of the writing (12).

*Subject-Specific Performance Task*: The performance task assessment was developed to understand and measure the progress in students' abilities to design an investigation that will allow them to generate an argument in response to a research question. The students must develop an original investigation and make decisions about the appropriate data to collect and evidence to use when generating their argument. This assessment is done in a group of 3-4 students, and the group submits a final product for scoring. The final product includes areas for students to describe the investigation they designed, the data they collected, and the argument they created, along with justification for each of these sections. Initial group composition was maintained as much as possible during separate administrations, and if it was not, the resulting scores were not included in the analysis. The rubric for this assessment followed the structure of the assessment packet and focused on technical and theoretical elements present in each section that related to the nature of scientific inquiry.

*SUSSI*: The Student Understanding of Science and Scientific Inquiry (SUSSI) (Liang, Chen, Chen, Kaya, Adams, Macklin, & Ebenezer, 2006) instrument was adapted to measure students' understanding of the development and nature of scientific information. The assessment was comprised of 44 statements about science with Likert-scale agreement responses offered. Analysis of these answers assigned raw points to each response in relation to the nature of the item. Statements representing accurate ideas about science and scientific inquiry were scored a minimum of zero points (strongly disagree) to a maximum of four points (strongly agree). Statements representing inaccurate ideas about science were scored in a reverse manner. The authors of this instrument originally separated the assessment into several subscales representing major NOS concepts; however, the researchers grouped these subscales in appropriate groups relating to Aspect 2 of the science proficiency framework.

*Standardized Comparison:* As a point of comparison for these assessments, a more "traditional" assessment was constructed using released multiple-choice questions from several prominent standardized tests used as benchmark measures of student learning, including the NAEP, PISA, and TIMMS. These questions resemble the typical measures for gains in content knowledge that are ubiquitous in K12 education.

## *Analysis*

The research team developed a scoring rubric for each of these assessments. Sets of two team members then used these rubrics to score the assessments for each course. Each rater scored subsets of each assessment, representing 25% of the whole sample, and those scores were used for reliability analyses. Due to multiple raters for these assessments, the intra-class correlation coefficient (ICC), a measure of agreement similar to Cohen's Kappa, was used to calculate inter-rater reliability. All analysis teams achieved an intra-class correlation above 0.70, which is considered substantial agreement between scorers (Landis & Koch, 1977). Once the groups had achieved this level of agreement, the remaining data sets were equally distributed among individual raters. Table 2 provides the ICC's for each of the assessments, not including the SUSSI or Standardized Comparison test, which did not require multiple raters.

**Table 2**: ICC for Science Proficiency Assessments Using Multiple Raters

| Assessment | Life Science | Physical Science |
|---|---|---|
| Content Knowledge | 0.99 | 0.97 |
| Scientific Writing | 0.76 | 0.74 |
| Performance Task | 0.79 | 0.79 |

## Results and Discussion

Using multiple assessments provides a level of detail regarding students' performance and progress toward science proficiency that is not obtainable from a single instrument. In order to illustrate this point two students were selected from each of the two middle school courses that completed the battery of assessments described above. The students in each course were first divided into two groups, those students that exhibited no learning gains on the instrument used for a standardized comparison and those students who achieved at least a .50 normalized learning gain. One student was then randomly selected from each of these two groups for the life science and physical science courses.

These two subgroups, less than .50 normalized gain and greater than or equal to .50 normalized gain, were identified prior to the random selection of students such that the selection process would ensure diverse performances were represented. Because of the potential advantages afforded to high scoring, or "proficient," students and the additional barriers created for low-scoring students these groups were purposefully identified to serve as an initial selection criterion.

The following figures display the normalized learning gains for four students; those students identified as "student A" were randomly selected from the students in their course whom demonstrated a 0.50 normalized learning gain, or higher, on the multiple choice instrument used as the standardized measure for comparison purposes. Recall, this assessment was constructed of released items from popular state, national, and international assessments. The students identified as "student B" were randomly selected from the students in their course whom demonstrated zero normalized gain on the same instrument. The normalized learning gains displayed in each table below have been aligned to the four aspects of science proficiency discussed above. Each aspect of science proficiency corresponds to a specific assessment administered to the students or specific subscales within a given assessment.

### *Middle School Physical Science*

Physical Science Student A scored at a level of 63% on the initial multiple choice comparison assessment and then achieved a 0.67 normalized gain on the post-year assessment. This type of gain is considerable and might indicate that the student has made significant progress in their learning and understanding of physical science. However, when this result in evaluated in light of the other aspects of science proficiency this student has not made such and impressive gain. Student A only demonstrates a 0.13 normalized learning gain in his ability to describe scientific explanations in a short-answer format and furthermore, demonstrates no learning gain in his ability to use scientific explanations to solve a problem or make sense of a scientific situation. Student A also regresses in his ability to generate scientific explanation and arguments (-0.13) and in his understanding of the development of scientific knowledge (-0.14). Physical Science Student A does make positive learning gains in aspect four of science proficiency in his ability to design and conduct and investigation (0.24) and his ability to communicate in writing (0.15), as well as a small gain in his understanding of the nature of scientific knowledge (0.14). Once again the science proficiency profile for Physical Science Student A shows that a narrow focus on only the results generated by the multiple choice comparison instrument would misrepresent the actual abilities of this student, which further strengthens the need for multiple instruments when assessing science proficiency.

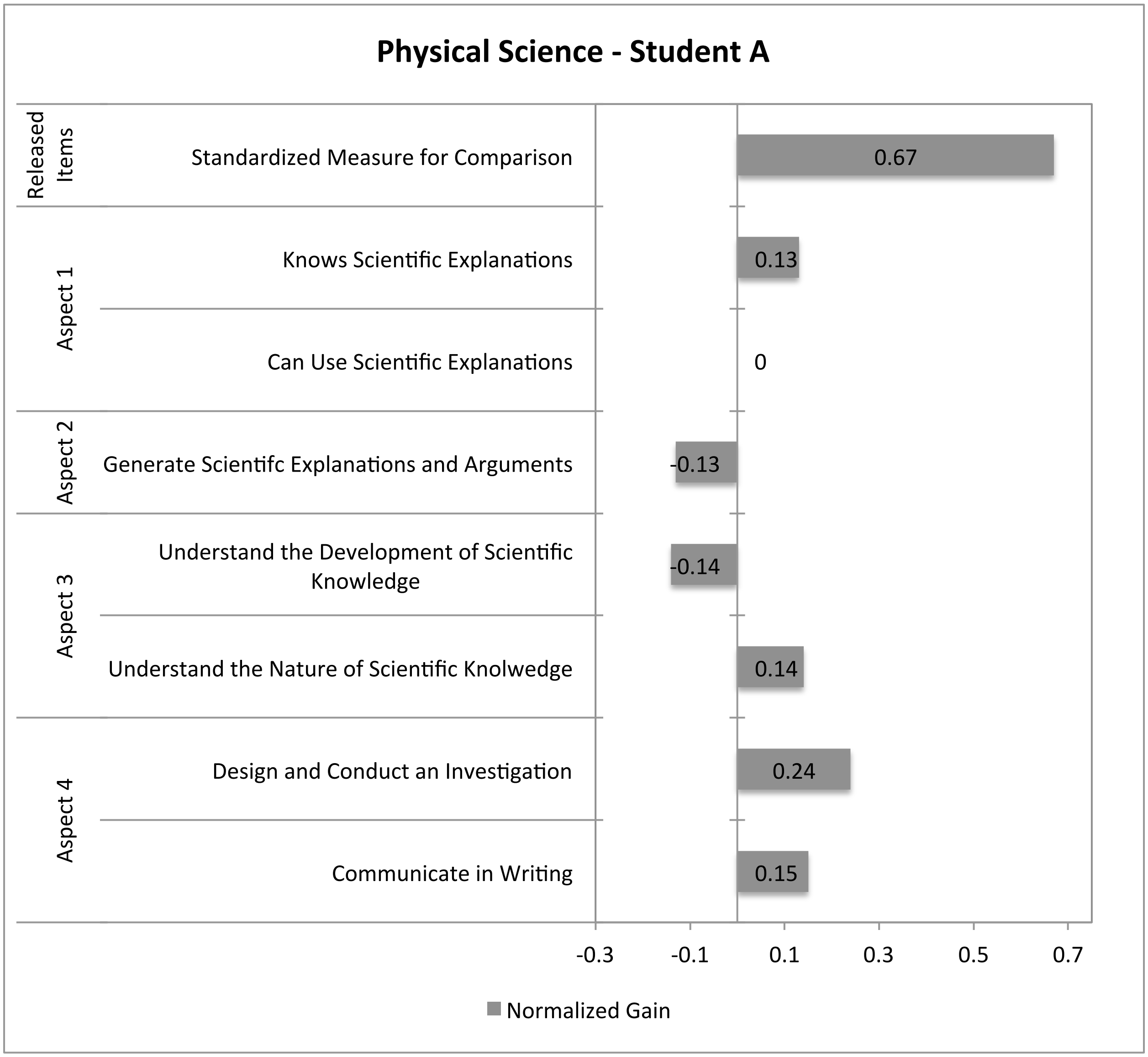

**Figure 1**: Normalized learning gains for Physical Science Student A

Physical Science Student B demonstrated no normalized gain on the assessment used as the comparison measure. As with other low performing students, Student B had substantial room for improvement after scoring only 38% on the pre-assessment at the beginning of the school year. In the case of Student B his performance in many of the aspects of science proficiency seem to mirror the lack of improvement on the comparison assessment. Physical Science Student B demonstrated no learning gains in his ability to use scientific explanations (0.0) or design and conduct and investigation (0.0). Additionally, this student regressed in aspect three, his understanding of the development of scientific knowledge (-0.17) and his understanding of the nature of scientific knowledge (-0.03). In contrast, he was able to achieve a 0.09 normalized gain with regard to his understanding of scientific explanations when asked in a short-answer format. He also experienced positive gains in his ability to communicate in writing (0.18) and quite a substantial gain in his ability to generate scientific explanations and arguments (0.56). Once again these data support the ineffectiveness of one assessment in capturing a full picture of a students' abilities within the framework of science proficiency.

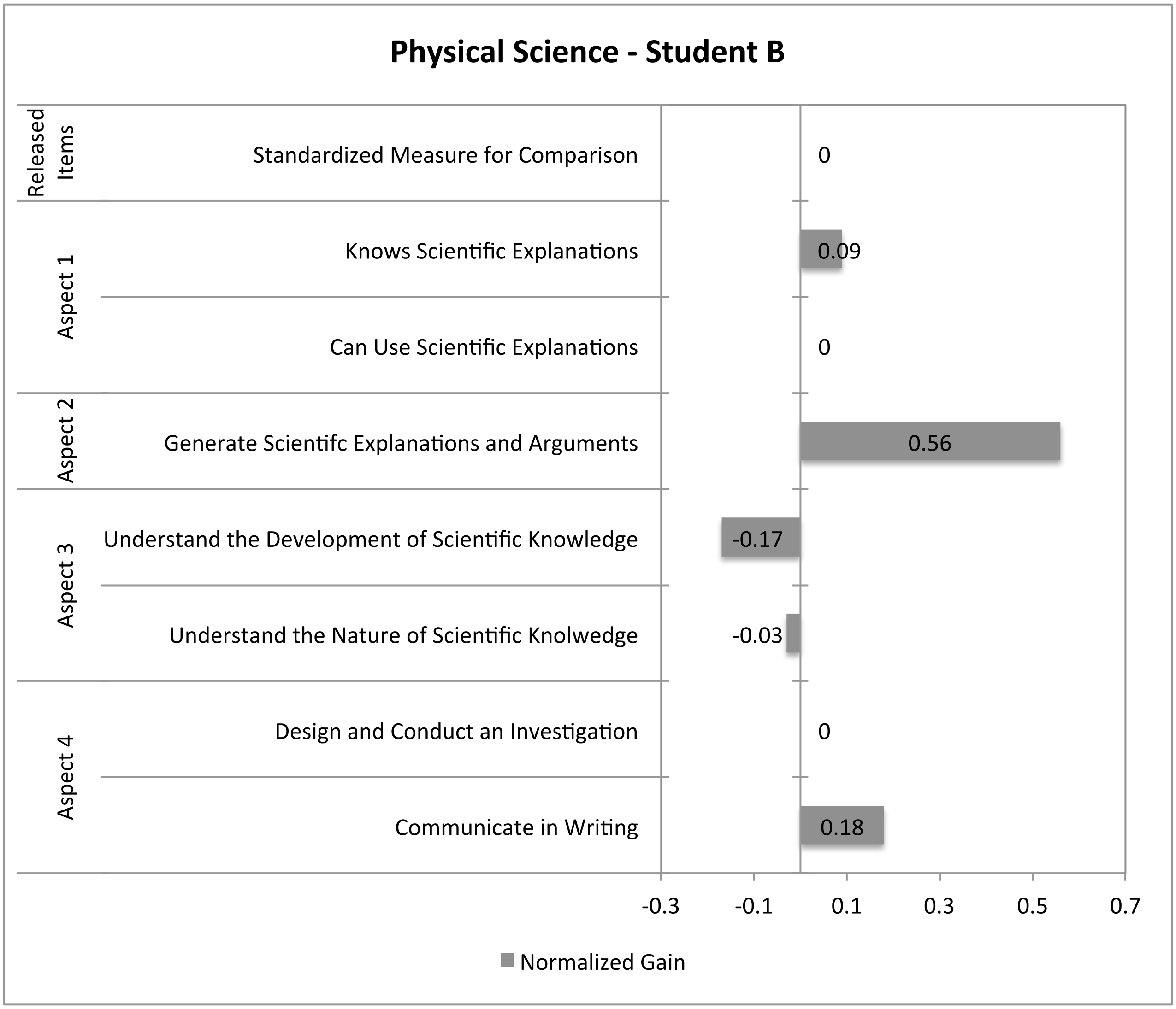

**Figure 2**: Normalized learning gains for Physical Science Student B

### *Middle School Life Science*

Life Science student A demonstrated a .64 normalized learning gain on the standardized assessment, moving from 39% correct on the pretest to 78% correct on the post test. This student's performance on the multiple assessments used to create this profile resonates with the standardized score, showing sizable increases across all aspects. The student increased in both their knowledge of scientific explanations (0.19) and the ability to apply them to novel situations (0.29). The student's greatest gain of 40% is seen in Aspect 2, the ability to generate scientific explanations and arguments. The next largest gain for this student (0.31) relates to their ability to design and carry out an investigation. The 0.17 normalized learning gain regarding the student's ability to communicate in writing provides evidence that even students who are high performers can continue to improve. The area where this student showed the least amount of learning is in Aspect 3 dealing with the nature (-.05) and development (0.10) of scientific knowledge. This result is not terribly confounding in light of the strong body of research that indicates these more abstract ideas about science are some of the most difficult to learn and change, especially in middle school students.

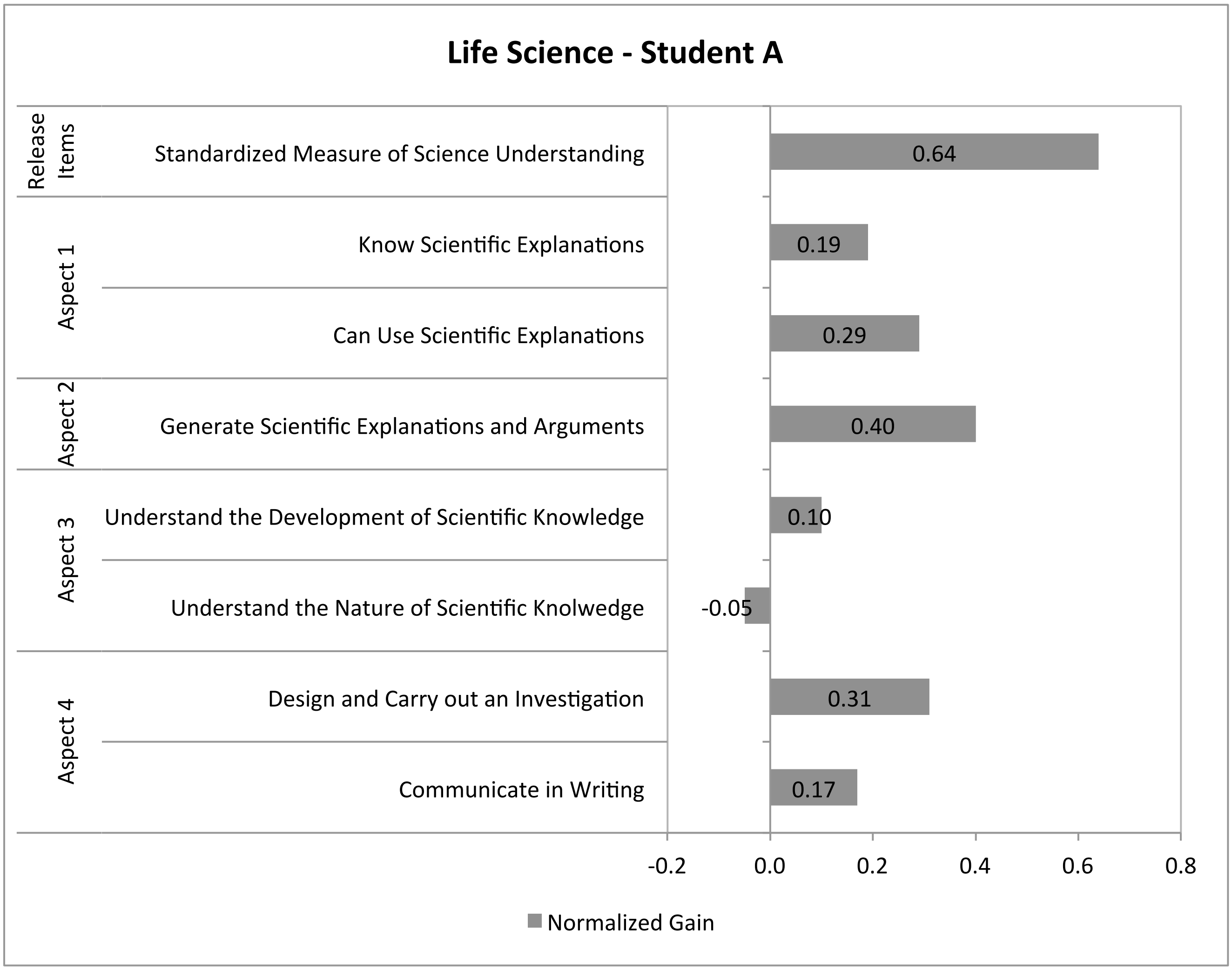

**Figure 3**: Normalized learning gains for Life Science Student A

Life Science student B achieved a 28% correct score on the pretest of standardized items and a 33% on the posttest. This difference in scores resulted in an 8% normalized learning gain on this assessment. This type of score would indicate that this student did not achieve a noticeable amount of science learning. The results from the multiple assessments argued for in this paper reflect a similar trend, indicating minimal learning gains or decreases. The student exhibited a 14% learning gain in the knowledge of scientific explanations, yet this student also demonstrated a 5% learning decrease in their ability to use those explanations. The student also indicated a learning decrease in their understanding of the nature of (-8%) and development of (-6%) scientific knowledge. The student did not show any learning gain in their ability to communicate in writing. However, this student still offers another case where the score on the standardized items masks to spectrum of learning that did occur over the course of the year. The student's two most significant learning gains occurred in the Aspect 2, the ability to generate scientific explanations and arguments (100%) and part of Aspect 4, the ability to design and carry out a scientific investigation (24%). These results support the impact of an argument based instructional strategy; insights that may not be captured by more traditional standardized assessments.

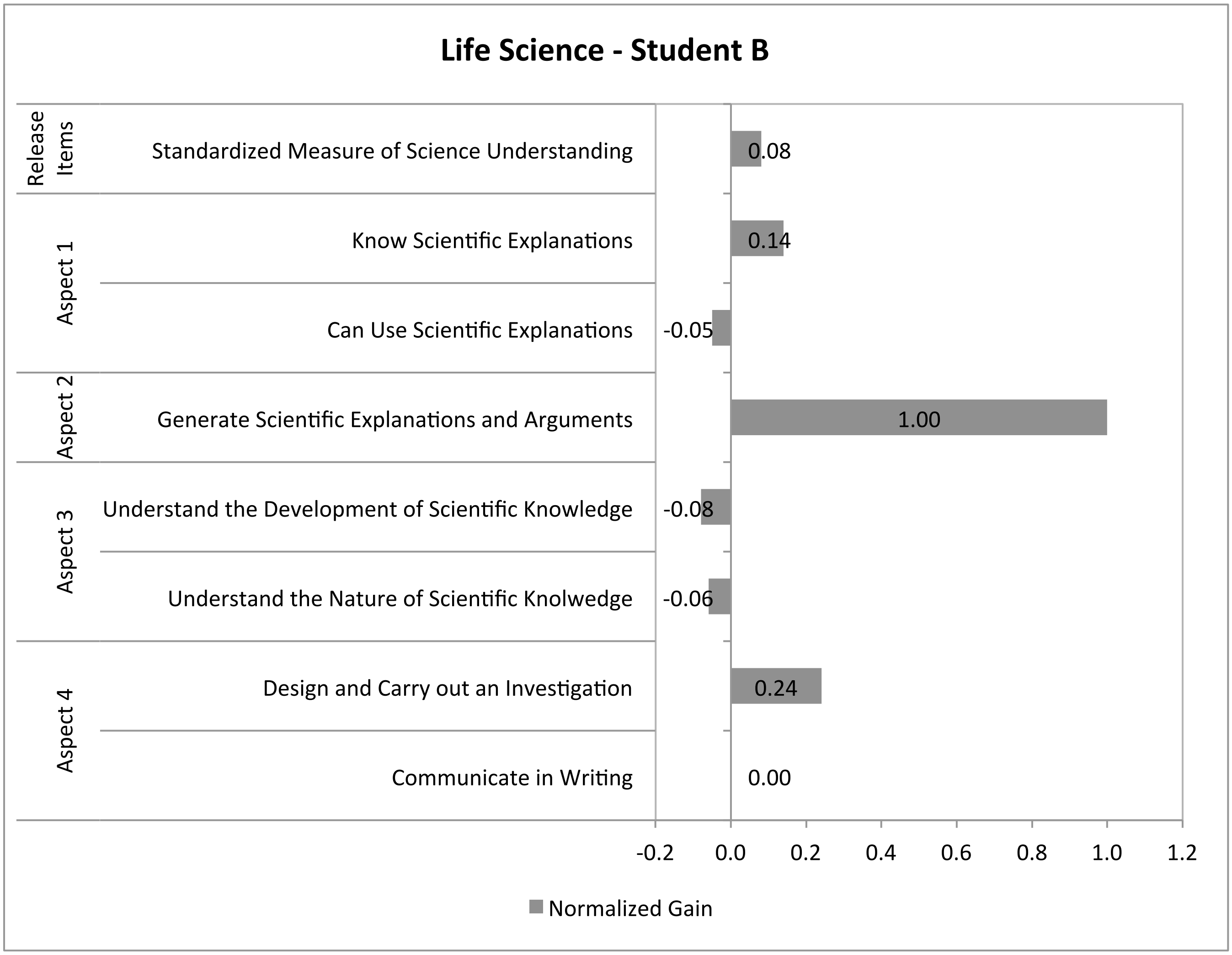

**Figure 4**: Normalized learning gains for Life Science Student B

## Conclusions

Typical high-stakes standardized assessments are limited in that they do not offer a complete picture of a student's science proficiency. The case studies examined above demonstrate how students may perform poorly or exemplary on one assessment (i.e. the standardized comparison instrument), however that assessment may not target all aspects of science proficiency. In order to fully capture a student's abilities within a science proficiency framework, multiple assessments that reflect the variety of skills and abilities associated with science proficiency are needed.

In many cases high-stakes multiple-choice assessments provide the basis for decisions concerning the student's academic future. Students who demonstrate no normalized learning gains within a school year or from year to year may be required to complete remediation tasks, may not be given access to upper-level courses, or may be required to repeat a course or grade level. Additionally, students demonstrating a .50 normalized learning gain, or higher, may be advanced to the next grade level or more advanced course work, they may be given the opportunity to opt-out of courses or specific course work, or be rewarded for achieving adequate yearly progress. Basing such academic decisions on one test is not in the best interest of students in either category.

As resources are directed to developing new science standards and frameworks that privilege scientific knowledge and using that knowledge to generate and evaluate scientific arguments, resources must also be directed toward developing assessments to assess such skills. Additionally, frameworks that emphasize students' abilities to complete performance based tasks, i.e. designing and conducting scientific investigations, quality assessments must be developed to target these skills and abilities in a manner other than typical multiple choice format. Similarly, attempts to measure students' abilities to communicate through scientific writing requires that assessments engage students in those practices rather than solely critiquing the communication of others. Assessments tend to serve as a driving force in education policy and curricular decisions, thus increased focus and a concerted effort to develop more authentic and educative assessments are needed.

## References

Donovan, M. S., Bransford, J. D. & Pellegrino, J. W. (2000). *How people learn: Brain, mind, experience, and school: Expanded edition*. Washington, DC, National Academy Press

Duschl, R., Schweingruber, H., & Shouse, A. (Eds.). (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: National Academies Press.

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

Liang, L., Chen, X., Chen, S., Kaya, O., Adams, A., Macklin, M. & Ebenezer, J. (2006). Student Understanding of Science and Scientific Inquiry (SUSSI): Revision and Further Validation of an Assessment Instrument. *Paper presented at the 2006 Annual Conference of the National Association for Research in Science Teaching (NARST), San Francisco, CA, April 3-6*

Sampson, V., Grooms, J., & Walker, J. (2011). Argument-Driven Inquiry as a way to help students learn how to participate in scientific argumentation and craft written arguments: An exploratory study. *Science Education, 95*(2), 217-257.