Examining Problem Solving Strategies on Multiple-choice Science Items Among English Language Learners Through Cognitive Interviews

Rachel Prosser

University of Colorado at Boulder

&

Guillermo Solano-Flores

University of Colorado at Boulder

Paper presented at the Annual Conference of the American Educational Research Association, New Orleans, Louisiana, April 7-11, 2011.

Abstract

This paper introduces a coding system used to compare the ways English language learners (ELLs) and mainstream students make sense of multiple-choice science items administered in English. Thirty-nine native Spanish-speaking ELLs and thirty-nine monolingual, mainstream students participated in cognitive interviews in which they were asked to report their thinking during and after responding to science items. The coding system was developed based on the analysis of the transcriptions invoking theories of bilingualism, sociolinguistics, and reading comprehension. The coding system allows comprehensive examination of the ways in which each group of student makes sense of items and makes it possible for test developers to investigate the wide range of cognitive resources students use to understand items.

Note: This study is part of a larger study titled, “Design and Use of Illustrations in Test Items as a Form of Accommodation for English Language Learners in Science Assessment,” funded by the National Science Foundation (Award No. DRL 0822362). We are grateful to the funding agency, our colleagues in the project, and the members of our technical advisory board for their support. The opinions expressed are not necessarily those of our colleagues or the funding agency.

English language learners (ELLs) are the fastest growing student population in the U.S. (Flynn & Hill, 2005). Currently, over 11% of students in the U.S. are classified as ELLs (Dalton, Sable, & Hoffman, 2006) and some researchers predict that, by the year 2030, over 40% of students will come from homes where English is not the main language spoken (Thomas & Collier, 2002). This rapid increase in the linguistically diverse student population calls attention to the need for more research in the testing of emerging bilinguals.

No Child Left Behind policy requires all students to participate in standardized testing after a short time of being schooled in English, before they have developed the academic language they need in English to properly benefit from instruction in their second language (Colllier & Thomas, 1989; Hakuta, Butler, & Witt, 2000). While important decisions about students, teachers, and schools are made based on large-scale tests administered in English, these instruments are developed exclusively with native English speaking students (Abedi, Hofstetter, & Lord, 2004). This exclusion of ELLs in the process of test development may be linked to serious issues of validity in ELL assessment (MacSwan, Rolstad, & Glass, 2002; Solano-Flores, 2008).

This paper is an attempt to expand the range of approaches that can be used with the intent to ensure fair, valid assessment for ELLs. Overreliance on existing or commonly accepted practices can lead to the wrong assumption that issues of validity in ELL testing are resolved. For example, in spite of the widespread use of testing accommodations, only certain language-based accommodations used with the intent to eliminate language as a source of construct irrelevant variance (Abedi & Lord, 2001; Mahoney, 2008; Rivera & Stansfield, 2001; Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006) have proven to be effective—although moderately effective—in reducing the score gap between ELL and non-ELL students due to language factors.

In the investigation reported here, we address differences in the ways that monolingual native English speaking students and native Spanish speaking ELLs make sense of standardized test items and the need for appropriate analytical tools for examining these processes. We focus on the conceptual foundation and structure of a system for coding students' verbalizations and responses in cognitive interviews. This coding system should allow test developers to examine the comprehension of multiple-choice items among both monolingual and bilingual students as critical to improving test development practices.

We first discuss the framework that provides the conceptual foundation of the coding system; then we report the process of development of the coding system; and then we describe its content and structure. Reporting results comparing students' understanding of items using this coding system is beyond the scope of this paper. However, in the last section we discuss the kind of data the coding system allows to obtain and the kinds of analyses that can be performed in order to examine, from a cognitive perspective, commonalities and differences in the ways in which ELLs and their non-ELL counterparts reason about and respond to science test items.

Conceptual Framework

While cognitive interviewing has been used in assessment research for at least twenty years (e.g., Hamilton, Nussbaum, Snow, 1997; Norris, 1990; Ruiz-Primo, Baxter, & Shavelson, 2001), with few exceptions (e.g., Prosser & Solano-Flores, 2010; Winter Kopriva, Chen, & Wiley, 2004) this source of information has been used scantly with ELLs. Surprisingly, a strong rationale and the basic principles for using cognitive interviews with ELLs have been available for over a decade (Kopriva, 2000, 2008).

Intended to elicit information that is verbally encoded by the participant, the use of verbal protocols is regarded as a valuable resource for accessing mental processes (Ericsson & Simon, 1984) in spite of the fact that some authors (e.g., Smagorinsky, 1998), challenge the notion that cognitive interviewing procedures transfer across cultures. Based on examining the amount of measurement error due to sampling of the questions included in cognitive interviews, Solano-Flores and Li (2009a) found that many cultural factors shape the kind of information obtained in cognitive interviews; students of different cultural backgrounds differ in the length and the level of detail of their responses. However, these authors note that, as long as rigorous coding procedures are used, these differences do not affect the validity of the information obtained for each cultural group.

Kopriva (2000) points out that there is a great need for more in-depth qualitative research with ELLs if we are to learn more about the specific nature of their mental processes. This assertion is consistent with findings from the field of second language acquisition, which show that bilingual students continually use their two languages in their thinking (Bialystok, 2001; Grosjean, 1989). Thus, it is reasonable to expect that even bilinguals who are highly proficient in their second language differ from their monolingual counterparts on some of the cognitive processes that take part during test taking. Thus, differences between monolinguals and bilinguals should therefore be reflected in any coding system used to compare thinking processes.

Unfortunately, studies analyzing the different ways in which language influences ELL performance (e.g., Abedi & Lord, 2001; Mahoney, 2008; Rivera & Stansfield, 2001; Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006) do not address this uniqueness and do not include the kind of in-depth qualitative analysis recommended by Kopriva (2000; 2008). In contrast, we acknowledge that extreme heterogeneity exists among bilinguals in the ways in which they utilize language in testing contexts. As a result of this heterogeneity, the performance of each ELL student on assessments is shaped by the interaction of the individual's language proficiency in English and in their first language, their knowledge of the content being assessed, and the particular set of linguistic challenges posed by each item (Solano-Flores & Li, 2009b).

We focus on what many researchers identify as the first step of problem solving: comprehending the item (Lieghton & Gokiert, 2005; Pretz, Naples & Sternberg, 2003). Item comprehension refers to the ability of the student to make sense of the text of the item. As Polya (1973) points out, "the worst may happen if the student embarks upon computation or constructions without having *understood* the problem." (p. 5). However, understanding the task at hand is not only critical for the participant, but also for the item writers. Exploring in detail how students make sense of test items is a crucial step into investigating whether or not an item functions as item writers intended (Leighton & Gokiert, 2008; Messick, 1989).

Utilizing a sociolinguistic approach to investigate how students make sense of the item, we consider semantic aspects of the item and distinguish between science content and non-science content words. Content words are essential to the item's construct and cannot be removed without changing what the item measures. For example, *photosynthesis* is a science vocabulary word and cannot be replaced with any other word without altering what the question is asking. In contrast, non-content words can be replaced without altering what the item measures. Rather than using *primary function*, an item can state the *main function* or *most important function,* because *primary* is not considered a science content word. It is important to note that for many students, the non-science content words are often as poorly understood as content related terms (Wellington & Osborne, 2001). Also, it is important to mention that the distinction between what is considered to be or not to be academic language is not a clear cut. Rather, it is a matter of usage frequency, involving discursive forms, ways of building arguments, and even ways of socializing (Bayley & Butler, 2003; Scarcella, 2003).

The coding system is also sensitive to differences in the ways that students read aloud. We incorporate aspects of the running record, a type of analysis designed to focus on the systematic actions students carry out when reading a text out loud (Clay, 2002). When running records are combined with questions about students' comprehension, they provide detailed information about students' understanding of the text. Although this type of approach analyzing students' reading abilities was developed for monolingual students, incorporating aspects of second language acquisition makes it appropriate for use with bilingual students.

Development of the Coding System

The coding system was developed based on an exhaustive analysis of the transcriptions of the interviews with both ELL and non-ELL students. Utilizing a constant comparative method (Glaser, 1965), the coding system was developed by incorporating aspects identified as potentially challenging for students in academic texts, including testing register (Abedi, 2006; Abedi, Hofstetter, & Lord, 2004; Solano-Flores, 2006; 2008) and science vocabulary (Osborne & Wellington, 2001). Thus, we devised a coding system that targets challenges students may face interpreting : (1) terms that are specific to the domain of science, (2) terms that are not specific to the domain of science, (3) terms that have more than one meaning, and (4) terms that are common to scientific discourse.

Larger research context. The coding system was developed as part of the activities for a larger, three-year study, currently in progress and funded by the National Science Foundation. This larger study investigates the use of illustrations as a form of testing accommodations for

ELLs. Its purpose is to investigate the effect of vignette illustrations—images added to items as a form of visual support without changing their text (Solano-Flores, 2010). As a part of this larger study, we used cognitive interviews to investigate specific technical properties of the illustrations as well as pinpoint the ways in which ELLs and mainstream students used the illustration when responding to the items.

Selection of Items and Development of Vignette-Illustrations. We used ten Grade 8 multiple-choice science items selected from the publicly released item pools of various assessment programs including CSAP (Colorado State Assessment Program), TIMSS (Trends in International Mathematics and Science Study), CST (California Standards Test), and AIMS (Arizona Instrument to Measure Standards). In their original version, these items did not contain any illustrations. An illustrated version of nine of the items was created by adding an image next to the text of the item with the intent to provide non-linguistic support for ELLs based on the procedures detailed by the ITELL project staff. These procedures are described elsewhere (Solano-Flores, 2010a, b). They allow for the systematic design of images, based on the analysis of the linguistic properties of the items and the potential linguistic challenges these properties may pose to ELLs. The tenth item was not illustrated because it was the item all students took in the non-illustrated form.

Participants. The participants in this study were 39 native Spanish-speaking students classified as ELLs and 39 native English-speaking students in grades 6-8 from an English-only middle school in a large district in a Western Mountain state. As with the entire population of ELL students, the ELLs who participated in this study came from a variety of backgrounds, had lived in the United States for varying amounts of time, and were classified at different levels of English proficiency. Of the 39 ELLs, the majority (26 students) were classified as Limited

English Proficient (LEP), 9 ELLs were classified as Fully English Proficient (FEP), and 4 ELLs were classified as Non-English Proficient (NEP). Fourteen students reported having been born in the U.S., while the majority, 21 students, reported having been born in Mexico. For those students not born in the U.S., the length of residence in the country ranged between one and 13 years.

In addition to the students, three kinds of professionals participated in the study, as key actors in the development of the coding system: (1) a professor of science education who was also a licensed science teacher, who examined the science register used in the multiple-choice items, (2) three former bilingual teachers and experts in bilingual literacy, who consulted on the use of running record, and (3) a specialist in science education with extensive knowledge using cognitive interviews who provided guidance on the development of the verbal protocol.

Iterative Process of Development of the Verbal Protocol. Over a year period, several iterations of the verbal protocol were developed. Modifications were made based on information gained from examining pilot students' interviews and the input from teachers participating in the larger study. Once the protocol was developed in English, it was translated to Spanish. Due to the limitations in the number of bilingual students participating in the study, we were unable to pilot the instrument with bilingual students to inform the translation.

Interview Procedures. Following recommendations given by Ericsson and Simon (1993) and Paulsen & Levine (1999), we used a combination of three types of reporting: concurrent reporting, retrospective reporting, and follow-up questions. Concurrent reporting provides insight into the mental processes as they occur in real time, while retrospective reporting provides insight into the participant's understanding of what processes he/she used to complete the task. Finally, follow-up questions allow the researcher to investigate specific aspects of thinking processes in more depth. Although these types of reporting often result in similar information, they can offer slightly different insight into the mental processes.

Students were interviewed individually on the ways in which they interpreted illustrated and non-illustrated items. During the interview, each student was given four items; two items were illustrated and two were non-illustrated. One of the illustrated items and one of the non-illustrated items were the same for all participants. The remaining two items (one illustrated item and one non-illustrated item) were randomly assigned from the remaining pool of eight items.

The interview protocol included five phases: introduction, warm-up, concurrent reporting, retrospective reporting, and follow-up probing questions. Once the introduction and warm-up phases were completed, students participated in concurrent reporting which required them to verbalize their thoughts as they complete the task. If a student fell silent, the interviewer followed recommendations given by Ericsson and Simon (1993) that instruct the interviewer say only one thing: “keep talking” (or “sigue hablando”). Because prompts such as, “tell me what you’re thinking,” can trigger metacognitive processes not naturally occurring in the participant’s mind, no other prompts were given to students during their concurrent reporting.

The concurrent reporting section followed the retrospective reporting for the same item where students recalled their steps in responding to the item. Once all concurrent and retrospective reporting was complete for the four items in the interview, students answered follow-up questions about their experiences responding to item. This final section includes questions to investigate what students found difficult and confusing, how the illustration was used, and their interpretation of what the item asked them to do.

Interview procedures remained constant for both ELLs and non-ELLs. The only differences that occurred regarded language use. In the case of ELLs, the interviews began by reminding bilingual students that the interviewer spoke both English and Spanish. Additionally, bilinguals were explicitly told that they could use either language as they preferred at anytime throughout the interview. To create an inviting environment for both languages, the interviewer continually used both English and Spanish throughout the interviews.

On average, the interviews lasted about twenty minutes. All interviews were transcribed verbatim to represent the exact verbalizations, including pronunciations and pauses. Bilingual interviews were transcribed by a bilingual transcriptionist to maintain the student's use of both languages.

Content and Structure of the Coding System

The coding system consists of four parts, shown in Table 1. The first column provides the coding category, followed by a definition of the category, and an example from student verbalizations. We capture aspects at both macro and micro levels of the ways students make sense of the multiple choice items. The first and second categories focus on a more macro level, describing when students face challenges to understanding the item and any strategy they use to arrive at this understanding. The third and fourth categories focus on a micro level, investigating how students deviate from the printed text when reading aloud. In addition, we also investigate sociolinguistic aspects by distinguishing between aspects of student interactions with science and non-science aspects of the items. We also recognize when bilinguals draw from their native language by coding for use or influence of Spanish.

Table 1.
Making sense of the item

Category	Definition	Example
Overall Comprehension		
As is	Student (S) reads item as printed	
Forgets	S unable to recall item content after reading item	So, uhh I think it's uhh, what is it? So I think it's B. Wait, what's the question?
Confused	S is confused by an aspect of the item	*Why does it confuse you?* Um it's cuz at first I read the moon produces no light and then I read and yet it shines at night so it says it doesn't shine but then it shines.
Science term	S reports not knowing the meaning of a science term	A piece of pine wood floating floats on the surface of the lake because the water ex ex extrets, extends, I don't know what that word is.
Non-science term	S reports not knowing the meaning of a non-science term	*Were there any other words you didn't know*? "Excess, photosynthesis"
Syntactical	S reports not understanding part of the item at the phrase or sentence level	*Were there specific words in there that were hard for you?* "No, but I didn't got it that much so," *And so do you know what specific part you weren't getting?* All of it.
Strategies to Comprehend		
Self-monitor		
English	S tracks h. comprehension of the item in English	"The moon produces no light, and yet it shines at night. Why is this?" First I'm gonna read it again, 'cause it didn't really get it.
Spanish	S tracks h. comprehension of the item in Spanish	No example found
Reformulate		
English	S restates the stem into own words in English	"What is the primary function of large leaves found on the seedlings growing in the forest? OK, so why are there big leaves on small trees?
Spanish	S restates the stem into own words in Spanish	C porque decia que sea level would rise y se que es real porque se que el oceanno esta subiendo
Translate	S translates to Spanish or uses cognates	*Y que tal estas, polar ice caps?* Esto lo se, como capas del hielo.
Reread	S reads the item again	I read, first I read the questions and the answers and I tried I read the question again, cuz I didn't get it that well and then I went to look at the answers again and I just knew it was A.
Read Aloud Strategies		
Repeat	S repeats word while reading aloud	
Science term		To provide shade *shade* for the root system
Non-Science term		How should *How should* she arrange Kevin's steps so they are in the correct order?"
Phrase		Which characteristic of this animal shows that it is a mammal?" "It eats other animals." "It feeds its young milk."*It feeds its young milk*. "It makes a nest and lays eggs." "It has webbed feet."
Sound out	S slowly reads the word, syllable by syllable	
Science term		the moon has many cra-cra-ters
Non-Science term		To get rid of ex-exssssess water that is entering through the roots."
Produce Non-Word	S speaks word that is not officially a word	
Science term		downward force equal to the weight of the *dis-ment* water
Non-Science term		A, to proveel shade for the root system,
Omit	S leaves out a word when reading aloud	
Science term		to allow the leaf damage by insects, to gather as much light as possible for.
Non-Science term		C the moon covered with a thin layer of ice.

Category	Definition	Example
Mispronounce	S pronounces word other than the standard form	
Science term		To gather as much light as possible for foto-SIGN-thesis
Non-Science term		to a-low for leaf damage by insects
Insert	S adds a word to those written in the item	
Science term		What is the primary function of large *forest* leaves found on the seedlings growing in the forest?"
Non-Science term		Hold *up* the watch in your hand
Substitutions	S substitutes a word for those written on the page	
Science		
English: meaning chaned		B, to get rid of *existing* water that is entering through the roots
English: meaning kept		What is the primary *fraction* of the large leaves found on the seedling growing in the forest?
Spanish: meaning changed		No example found
Spanish: meaning kept		No example found
Non-Science		
English: meaning changed		to divide shade for the root system
English: meaning kept		The moon is covered with tiny layer of ice
Spanish: meaning changed		No example found
Spanish: meaning kept		Step one: hold the stopwatch in *the* hand.
Corrections		
Non-word	S properly says a word initially said as non-word	
Science term		OK, when ma, when *Ma..Magnilium* - MG metal is burned, in the presence of Oxi-gent - O2 *magnesium* oxide MGO is produced.
Non-Science term		No example found
Omission	S inserts a word initially left out	
Science term		No example found
Non-Science term		How should she arrange Kevin's steps so they are in the order, in the *correct* order?
Mispronunciation	S properly pronounces after reading it twice	
Science term		"If the temperature of *earse-earth* rose over time, which of the following would occur?"
Non-Science term		Step one: hold the stop-watch in *on* in *one* hand.
Substitution	S reads printed word initially substituted by another	
Science term		A piece of pine wood floats on the surface of a lake because the water exerts: an upward force *equals*, *equal* to the weight of the wood,
Non-Science term		Alexander Flemming noticed that bacteria growing on a plate agar did not grow next to the mold that *it*, that *was* growing on the same plate.
Insertion	S removes an added word	
Science term		No example found
Non-Science term		To provide shade for the root system, to get rid of ex-cess water to that, that is entering through roots

The first part of the coding system, Overall Comprehension, tracks students understanding of the item. For example, students may forget the content of the item after reading it in entirety. Or, students may report that an aspect of the item is confusing. In addition, this category identifies when students do not understand science terms, non-science terms, or when they understand individual words but cannot make sense of them at a syntactical level.

The second part of the coding system, Strategies to Comprehend, identifies specific activities students carry out when making sense of the item. Here, it is possible bilingual students may draw from their native language to carry out a specific strategy, such as self-monitor their comprehension. Or, a student may restate the problem in their own words, using English, Spanish, or a combination of both. In addition, students may draw from the use of cognates or need to reread the question to make sense of it.

Next, Read Aloud, utilizes a bilingual view of the running record analysis to investigate students understanding of the item at a micro level. Focusing on the strategic behavior students carry out when reading aloud, this part of the coding system identifies when students verbalizations deviate from the printed text. Note, while we are coding deviations, these have been shown to be strategic behavior of good readers (Clay, 2002). We do not consider these deviations to be incorrect, but rather different from the printed text.

For example, students may substitute words when reading aloud, informing these substitutions based on their knowledge of their language system(s). When students substitute, they inform their choice with visual, syntactical, or semantic information from the other words on the page (Clay, 2002). Substitutions may or may not change the meaning of the item. Take the following excerpt, “What is the primary *fraction* of the large leaves found on seedlings growing in a forest?” Here, the student has changed the meaning of the science term *primary function* by substituting it with *primary fraction*. Most likely this student is relying on visual and semantic information for this substitution, seeing the word *primary* followed by another word that begins with the letter f which triggers him or her to say a common mathematical term, *primary fraction*. Students may also substitute words that preserve the printed text's meaning. For example, "the moon is covered with a *tiny* layer of ice" is an example of a student substituting a non-science word (*tiny* in place of *thin*) as both words start with the letters "ti" and refer to a small amount.

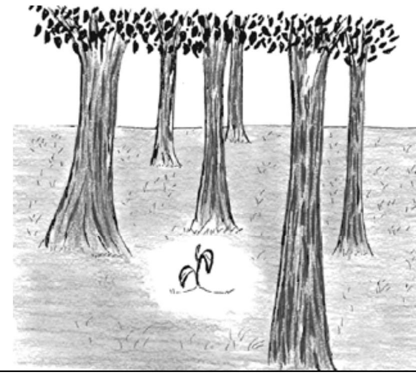
Finally, the last section of the coding system, corrections, identifies when students correct an initial deviation from the printed text. These corrections describe when a student changes their original verbalization to match the printed text of the item, identifying in the previous section of the coding system. Corrections may occur immediately following the verbalization of a deviation or significantly after the initial verbalization.

Once the features of this coding system are identified in a student's think aloud of a give item we can add the total number of features. This number provides an estimate of the cognitive resources students devote to understanding a given item. While this number is not an exact representation of the cognitive resources students use to make sense of an item, it does illustrate aspects of the interaction between student and item. In addition, this coding system allows us to focus on problematic aspects of test items. For example, analyzing the number of students who repeat, sound out, produce a non-word, or substitute a particular term in an item could suggest that item writers consider an alternative ways of asking the question.

Below is an excerpt from a student's interview, coded using this system as he answers the following problem:

What is the primary function of the large leaves found on seedlings growing in a forest?

A. To provide shade for the root systems
B. To get rid of excess water that is entering through the roots
C. To allow for leaf damage by insects
D. To gather as much light as possible for photosynthesis



Interview	Code
What is a prim primar primary[1] function of the large leaves found on seed[2] growing in a forest. To pro-vide[3], to provide[4] shade from the root system, to get ride[5] rid[6] of excess water that is entering through the roots. *Rereads under breath.*[7] To allow from leaf damage by insects, to gather as much light as possible for photo- sine- this this[8]. I think um, um the primary is a um something that helps a leaf or a tree grow.[9] I'm not sure though. And I think my answer is, to um, to provide shade for the roots systems.	[1]Sound out science word [2]Substitute science word: meaning changes [3]Sound out non-science word [4]Repeat non-science word [5]Substitute non-science word [6]Correct substitution [7]Rereads question [8]Sound out science word [9]Does not understand science word
Ok, great, can you tell what you did first, then what you did?	
First I read the paragraph then I um thinked what would be the answer then I just um grabbed each one and see which was better and describe it better. And yah	
Was there any part of this question that helped you solve it?	
Yah, um which is the primary function.	
MmHm, and what does that mean - primary functio?	
Like what does it do.	
What does what do?	
The primary, yah	
What does the primary do?	
Yah	
And what was the hardest part?	
Um, the question	
The question? Was there any specific part of it that was the hardest part? Was it hard because you didn't understand it or was I hard because you didn't know the answer?	
Cuz I didn't understand it[10].	[10]Does not understand phrase
What part?	
Where it said large leaves found on seedlings growing in a forest.	
MmHm,okay. And was there anything confusing? Other than you said this question was confusing up here	
Seed-lines[11]	[11]Does not understand science term
Seedling. You didn't know what that word was? Ok. Anything else?	
No	

The example shown above identifies eleven codes that describe how the student understands the item, focusing on challenging aspects. When reading aloud, the student sounds out and substitutes both science and non-science words. He also rereads the question to better understand it. In addition, he states that he does not understand two science terms (*primary* and *seedling*) nor does he understand part of the question at a phrasal level. This coding scheme permits detailed analysis of the student's experiences making sense of the item by identifying precisely how he uses his cognitive resources. By using this coding system to describe how ELLs and non-ELLs make sense of items, we are better able to understand differences in the ways these two groups of students interact with test items.

Perspectives on Data Analysis

As mentioned above, discussing results of analyzing empirical data with this coding system is beyond the scope of this paper. However, we report main issues in data collection and the forms of analyses being conducted to examine differences and commonalities in the ways in which ELL and non-ELL students make sense of science test items. Because our coding system incorporates aspects of second language acquisition and sociolinguistics we are able to more accurately understand ELLs' understanding of multiple-choice science items.

Upon completing qualitative analysis utilizing this coding system, we will add the number of features identified to calculate the number that describes the cognitive resources students devote to making sense of the item. Using this number, we will compare the resources ELLs and non-ELLs devote across items, both with and without illustrations.

Next, we will perform a series of two-way factorial ANOVA to examine the main and interaction of two factors: language status (ELL vs. non-ELL) and illustration (present vs. absent) on the frequency of the different coding categories shown in Table 1. Next, we will perform a series of Chi-square analyses to examine the strength of association of item comprehension, the inclusion or absence of illustrations; and whether or not students arrived at the correct answer. This quantitative analysis will allow examination of statistically significant differences of the ways ELLs and non-ELLs makes sense of items with and without vignette illustrations.

Final Remarks

Based on knowledge from the fields of cognitive science, problem solving, reading comprehension, and bilingualism, we have created a coding system that allows in-depth examination of how students make sense of multiple-choice science items. This coding system allows the researcher to compare how ELLs and mainstream students interact with test items, and the challenges they face. This coding system utilizes a holistic approach that considers bilinguals' entire linguistic repertoire when comparing these students to mainstream monolingual students. By adding the number of features present when a student responds to an item, we are able to identify the cognitive resources students use to make sense of the item. While these numbers should be used in caution, they may provide insight into the differences in challenges faced by ELLs and mainstream students.

This paper has focused on the actions *students* carry out to make sense of the item. To comprehensively investigate the interaction between the student and the item, consideration must be given to the linguistic features of the item. Future investigations should draw from previous investigations (e.g., Abedi, Hofstetter, & Lord, 2004; Rosebery, O'Connor, Noble, Suarez, Hudicourt-Barnes, & Warren, 2010) to identify item features that are potentially problematic for second language learners. Next, using the coding system described here, one could investigate how the actions students carry out to make sense of an item differ based on its linguistic features.

References

Abedi, J. (2006). Psychometric issues in the ELL assessment and special educational eligibility. *Teachers College Record, 108*(11), 2282-2303.

Abedi, J., & Hefri, F. (2004). Accommodations for students with limited English proficiency in the national assessment of educational progress. *Applied Measurement in Education, 17*(4), 371-392.

Abedi, J., Hofstetter, C., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, (74)*1, 1-28.

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

Bailey, A. L. & Butler, F. A. (2003). An evidentiary framework for operationalizing academic language for broad application to K-12 education: A design document. CSE Report 611. Los Angeles, CA: Center for the Study of Evaluation, Standards, and Student Testing. http://PAREonline.net/getvn.asp?v=4&n=6

Baxter, G.P., Elder, A.D., & Glaser, R. (1996). Knowledge-based cognition and performance assessment in the science classroom. *Educational Psychologist*, 31(2), 133-140.

Baxter, G.P. & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice*, 17(3), 37-45.

Bialystok, E. (2002). Cognitive processes of the L2 user. In V. J. Cook (Ed.), Portraits of the L2 user (pp. 145-165). Clevedon, England: Multilingual Matters.

Collier, V. P. & Thomas, W. P. (1989). How quickly can immigrants become proficient in school English? *Journal of Educational Issues of Language Minority Students, 5*, 26-38.

Dalton, B., Sable, J., and Hoffman, L. (2006). *Characteristics of the 100 Largest Public Elementary and Secondary School Districts in the United States: 2003–04* (NCES 2006-329). U.S.

Ericsson, K.A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data.* MITE Press.

Ericsson, K.A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity, 5*(3), 178-186.

Flynn, K., & Hill, J. (2005). English-language learners: A growing population. *Mid-continent research for education and learning (MCREL policy brief).* Retrieved July, 28, 2009 from http://www.mcrel.org/topics/products/257/.

Glaser, B. (1965). The constant comparative method of qualitative analysis. *Social Problems, 12,* 436-445.

Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language*, 36, 3-15.

Hakuta, K., Butler, Y. G., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* Santa Barbara: University of California Linguistic Minority Research Institute. Policy Report 2000-1.

Hamilton, L.S., Nussbaum, E.M., & Snow, R.E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education*, 10, 181-200.

Kopriva, R. J. (2000) *Ensuring accuracy in testing for English language learners: A practical guide for assessment development*. Washington, DC: Council of Chief State School Officers.

Kopriva, R. J. (Ed.), (2008). *Improving testing for English language learners* (pp. 169-200). New York: Routledge.

Leighton, J. & Gokiert, R. (2005). *The cognitive effects of test item features: Informing item generation by identifying construct irrelevant variance.* Paper presented at the annual meeting of the National Council on Measurement in Education, April, 2005.

MacSwan, J., Rolstad, K., & Glass, G. V. (2002). Do some school-age children have no language? Some problems of construct calidity in the Pre-LAS Espanol. *Bilingual Research Journal*, *26*, 395-420.

Mahoney, K. (2008). Linguistic Influences on Differential Item Function for Second Language Learners on the National Assessment of Educational Progress. *International Jornal of Testing, (8)*1, 14-33.

Norris, S. P. (1990). Effects of eliciting verbal reports of thinking on critical thinking test performance. *Journal of Educational Measurement,* 27, 41-58.

Paulsen & Levine, (1999). The applicability of cognitive laboratory method to the development of achievement test items. Paper presented at the annual meeting of American Educational Research Association, Canada, April.

Payan, R. M., & Nettles, M. T. (2006, August). Current State of English-Language Learners in the U.S. K-12 Student Population. www.ets.org/Media/jConferences_and_Events/pdf/.

Prosser R. & Solano-Flores, G. (2010)

Rivera, C. & Stansfield, C. (2001). The Effects of Linguistic Simplification of Science Test Items on Performance of Limited English Proficient and Monolingual English-Speaking Students. Paper presented at the annual meeting of the American Educational Research Institute, Seattle, WA.

Rosebery, A., O'Connor, M., Noble, T., Suarez, C., Hudicourt-Barnes, J., & Warren, B. (2010). Understanding children's Performance on multiple-choice items in achievement tests in science. Paper presented at the annual meeting of the American Educational Research Institute, Denver, CO.

Ruiz-Primo, M.A., Baxter, G. P., & Shavelson, R. J., (2001). On the validity of cognitive interpretations of scores from alternative concept-mapping technqiues. *Educational Assessment, 7*(2), 99-141.

Scarcella, R. 2003. *Accelerating Academic English: A Focus on the English Learner.* Oakland, California: Regents of the University of California.

Shaftel, J., Belton-Kocher, E., Glasnapp, D. & Poggio, J. (2006). The Impact of Language Characteristics in Mathematics Test Items on the Performance of English Language Learners and Students with Disabilities. *Educational Assessment, 11*(2), 105-126.

Smagorinsky, P. (1998). Thinking and speech and protocol analysis. *Mind, Culture, and Activity,* 5(3), 157-177.

Solano-Flores, G. (2006). Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of English-language learners. *Teachers College Record*. *108*(11), 2354–2379.

Solano-Flores, G. (2008). Who is given tests in what language by whom, when and where? The need for probabilistic views of language in the testing of English Language Learners. *Educational Researcher, 37*(4), 189-199.

Solano-Flores, G., & Li, M. (2009a). Generalizability of cognitive interview-based measures across cultural groups. *Educational Measurement: Issues and Practice, 28* (2), 9-18.

Solano-Flores, G., & Li, M. (2009b). Language variation and score variation in the testing of English language learners, native Spanish Speakers. *Educational Assessment, 14, 1-15.*

Solano-Flores, G. (2010a). *Vignette illustrations as a form of testing accommodation for English language learners: A design methodology for use in large-scale science assessment*. Paper presented at the Annual Conference of the National Council of Measurement in Education, Denver, Colorado, April 29-May 3.

Solano-Flores, G. (2010b). The use of pictorial supports as an accommodation for increasing access to test items for students with limited proficiency in the language of testing. Paper presented at the 7th Conference of the International Test Commission. Hong Kong, July 19-21, 2010.

U. S. Department of Education. (2007). *Fact sheet: NCLB provisions ensure flexibility and accountability for Limited English Proficient students.* Retrieved June 10, 2010, from http://www.ed.gov/nclb/accountability/schools/factsheet-english.html.

Thomas, W.P., & Collier, V. (2002). *A national study of school effectiveness for language minority students' long-term academic achievement*. Santa Cruz, CA and Washington, DC: Center for Research on Education, Diversity & Excellence. Available: http://www.crede.ucsc.edu/research/llaa/ 1.1_final.html.

Winter, P., Kopriva, R.J., Chen, C.S., and Wiley, D. (2004b) *Exploring student and item factors that affect assessment validity: Results from large-scale cognitive lab*. (C-SAVE Report No. 232). Madison, WI: University of Wisconsin.