The CATESOL Journal

LORENA LLOSA
New York University, New York City

Assessing English Learners' Language Proficiency: A Qualitative Investigation of Teachers' Interpretations of the California ELD Standards

This study investigates teachers' use of the English Language Development (ELD) Classroom Assessment, an assessment of English proficiency used in a large urban school district in California. This classroom assessment, which consists of a checklist of the California ELD standards, is used to make high-stakes decisions about students' progress from one ELD level to the next and serves as one criterion for reclassification. Ten elementary school teachers were interviewed and asked to produce verbal protocols while scoring the ELD Classroom Assessment of two of their students. Through six examples from the data, this paper shows that teachers do not interpret the ELD standards consistently and as a result the scores they assign on the ELD Classroom Assessment to different students have different meanings. The paper concludes by discussing several factors that might affect how teachers interpret standards and the implications of these findings for the use of standards-based classroom assessments within a high-stakes accountability system.

Introduction

The No Child Left Behind Act requires states and school districts to test the English proficiency of all English learners. States, districts, and schools must show annual increases in the number and percentage of students who become proficient in English, as well as in the number and percentage of students who make progress toward that goal. The language tests used to assess students' English proficiency in this accountability system have a direct impact on the educational opportunities of English learners. In many states, these language assessments are aligned to a set of English Language Development (ELD) standards designed to assist teachers in moving English learners to fluency in English and proficiency in the English Language Arts (ELA) Content Standards. In addition to statewide standardized tests, many school districts are using standards-based classroom assessments to monitor the progress of English learners.

In a standards-based system, classroom assessments provide several advantages over standardized tests that are administered once a year. Unlike standardized tests that provide only one score or performance level that represents students' mastery of the standards, classroom assessments have the potential to produce rich information that teachers can use to improve instruction and address the needs of individual students. Also, through classroom assessments a greater number of standards can be assessed based on a broad range of students' performance over a period of time, rather than their performance on a few of the standards on the day of the test (Popham, 2003).

Yet many questions remain as to the reliability and validity of the use of classroom assessments in a high-stakes accountability system (Brindley, 1998, 2001; Rea-Dickins & Gardner, 2000). Brindley (1998, 2001), who researched the use of classroom assessments in Australia and the United Kingdom, found considerable variability in teachers' interpretations of language ability. Similarly, Rea-Dickins and Gardner (2000) also uncovered in their study of teacher assessments in the United Kingdom several factors that threaten the validity of the inferences drawn from

classroom assessments. The present study investigates these issues in the U.S. context by examining teachers' use of the English Language Development (ELD) Classroom Assessment, a standards-based classroom assessment of English proficiency used in a large urban school district in California.

Research on the Use of Standards-Based Classroom Assessments

Brindley (1998) focuses on the problems that arise from the use of standards-based classroom assessments in the context of Australia and the United Kingdom. In particular, he questions the validity and reliability of classroom-based assessments used for accountability. Since individual teachers often devise their own assessment tasks to determine the extent to which their students are achieving the outcomes or standards, he suggests that "it is possible that the scores or ratings derived from a variety of different teacher-generated assessments of unknown validity and reliability are potentially invalid" (p. 64). Rea-Dickins and Gardner (2000) point out the following factors that may affect the validity of the inferences drawn about learners from classroom assessments: variability in the nature of the assessment activity; degree of task preparation; level and detail of rubrics; level and detail of spoken instructions provided by the teacher to individual learners; differences in difficulty levels of assessment activities; amount of assistance learners receive during an assessment; amount of time available to complete an activity; differences in content of the assessment activity; and differences in interlocutors (p. 236). Koretz (1998), who investigated the quality of performance data produced by large-scale portfolio assessment efforts in Kentucky and Vermont, also found that scores on portfolios varied as a function of raters, occasion, the rubrics used for scoring, and the tasks teachers assigned. These tasks, in turn, varied in terms of content, difficulty, and amount and type of assistance students received. Brindley (1998) concluded that "in high-stakes contexts, these inconsistencies could lead to unfair decisions which could adversely affect people's lives" (p. 70).

The variability in teachers' interpretation of language ability presents yet another threat to the validity and reliability of classroom assessments. Brindley (1998) reports that some studies have found that teachers and raters interpret language assessment criteria differently depending on their backgrounds, training, expectations, and preferences (e.g., Brown, 1995; Chalhoub-Deville, 1995; Gipps, 1994; Lumley & McNamara, 1995; North, 1993), even after being trained extensively (Lumley & McNamara, 1995; North 1993).

The present study addresses some of the issues regarding classroom-based assessment identified by Brindley (1998) and others, particularly the variability in teachers' interpretation of language ability, by investigating, through verbal protocol analysis, the way teachers rate students' language ability as they score the ELD Classroom Assessment. The next section briefly reviews the literature on verbal protocol analysis, the methodology used in this study.

Verbal Protocol Analysis

Since the 1980s, the use of verbal reports to study cognitive processes has increased significantly in many areas of psychology, education, and cognitive science. Even though verbal reports were being used to investigate second language learning in the 1980s (Faerch & Kasper, 1987), this methodology was not introduced in studies of language assessment until the field recognized the importance of test-takers' and raters' processes for test validation (Bachman, 1990). Most studies of language assessment that use verbal reports follow the methodology described by Ericsson and Simon (Lumley, 2002).

According to Ericsson and Simon (1993), the two forms of verbal reports that more closely reflect cognitive processes are concurrent verbal reports and retrospective reports. Concurrent verbal reports—the methodology used in the present study—also known as "talk aloud" or "think aloud," take place dur-

ing an individual's performance of a particular task. They are typically used to trace processes, understand problem solving, and shed light on an individual's decision-making processes. Two potential limitations of concurrent verbal reports are the incompleteness of the data gathered—since people think faster than they speak, only thoughts that are verbalized can be analyzed—and the obtrusiveness of the procedure—when conducting concurrent reports, the act of verbalizing thoughts can influence the thoughts themselves. In spite of these limitations, the use of verbal protocols has proved to be invaluable in research that describes raters' behavior in assessments of second language writing and reading (Cohen, 1994; Cumming, 1990; Cumming, Kantor, & Powers, 2001; Lumley, 2000; Lumley 2002; Milanovic, Seville, & Shen, 1996; Vaughan, 1991; Weigle, 1994; Zhang, 1998). In the present study, verbal protocol analysis is used to examine teachers' use of the ELD Classroom Assessment and their decision-making processes as they assess students' language ability. The next section provides a detailed description of the ELD Classroom Assessment.

The ELD Classroom Assessment

The ELD Classroom Assessment, used in a large urban school district in California, documents an English learner's progress toward each of the California ELD standards. The assessment is intended to be used for formative purposes, but it is also used to make high-stakes decisions about English learners' progress from one ELD level to the next (there are five ELD levels), and it serves as one of the criteria for reclassification as Fluent English Proficient. There is a separate set of assessments for different grade levels: kindergarten through grade 2, grades 3 through 5, grades 6 through 8, and grades 9 through 12. Each ELD Classroom Assessment consists of a list of all the ELD standards for a given ELD level divided into three sections—Listening/Speaking, Reading, and Writing—following the same organizational structure as the California ELD Standards. The Reading standards are further divided into the following sections: Word Analysis, Fluency and Vocabulary, Comprehension, and Literary Response. The Writing standards are divided into Strategies, and Applications and Conventions. Teachers are expected to score student progress toward each standard using the following scale:

4 *Advanced Progress:* Exceeds the standards for the identified ELD level
3 *Average Progress*: Meets the standards for the identified ELD level
2 *Partial Progress*: Demonstrates some progress toward mastery of the standards
1 *Limited Progress*: Demonstrates little or no progress toward mastery of the standards

Using this scale, teachers assign a score to each standard by taking into account the overall performance of the student in the class. When the teacher determines that a student has mastered all the ELD standards at a given ELD Level—that is, the student scored a 3 or a 4 on all standards—the student advances to the next ELD level. Teachers must also include in the ELD Classroom Assessment folder selected examples of student work that demonstrate mastery of the ELD standards and that justify scores recorded on the ELD Classroom Assessment. Teachers are expected to update the assessment each reporting period (three times a year) and periodically gather student work samples, such as observation checklists, writing samples, and reading records. The scores on the assessment, however, are not based exclusively on these work samples. Teachers base their scores on a student's overall performance in class and thus, the ELD Classroom Assessment functions somewhat like a report card. Teachers are not systematically trained to score the ELD Classroom Assessment, but they are encouraged to discuss their students' scores with other teachers to develop consistency in scoring across grades and ELD levels.

Research Questions

The large-scale use of the ELD Classroom Assessment as part of a high-stakes accountability system allows for the investigation of several issues related to the assessment of English proficiency via standard-based classroom assessments. This paper focuses on the following research questions: When scoring the ELD Classroom Assessment, how do teachers make decisions about students' English language proficiency as defined by the California ELD standards? More specifically, to what extent do teachers interpret the ELD standards consistently as they determine student mastery of the standards?

Method

The sample for this study consists of 10 fourth-grade teachers in five schools with large populations of English learners. Eight of the teachers were female and 2 were male. Six of the teachers were white, 2 were Asian-American, 1 was Hispanic, and 1 was African American. The number of years working in their school district ranged from 1 to 26 years.

This sample is part of a larger study designed to investigate the constructs assessed by the California English Language Development Test (CELDT) and the ELD Classroom Assessment. For the larger study, 20 schools were randomly sampled among schools in the district that have a majority population of English learners (50% or more). Eight principals of the 20 schools invited to participate in the study agreed to participate. (Five principals declined to participate and the remaining principals did not respond). In each of the eight schools, all fourth-grade teachers were invited to participate in the qualitative study. The first 10 teachers who accepted the invitation were included in the study and paid a stipend of $50 for their participation.

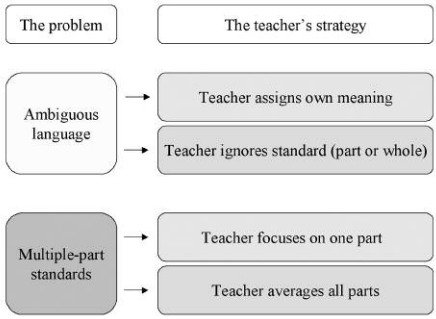
Meetings with each teacher were scheduled at the teachers' convenience, typically after school, and lasted between 1 and 1.5 hours. Each teacher was asked to sign a consent form and for permission to record the session. During the session, the teacher was first trained on how to produce a "think-aloud" verbal report. The teacher was then asked to engage in a "think aloud" as he scored two of his students' ELD Classroom Assessments. Because the majority of English learners in fourth grade are ELD levels 3 and 4, teachers were asked to score the assessments of students at one of those two ELD levels. (The instructions provided to teachers for producing the verbal protocols are included in the Appendix.)

These concurrent verbal protocols were audio-recorded and later transcribed. The data were analyzed qualitatively. All transcripts were read carefully for themes and patterns in the data. The verbal protocols were then grouped by ELD standard. For each ELD standard, teachers' interpretations and strategies for scoring students on that standard were examined and then compared across teachers. Finally, themes and patterns in teacher interpretations and strategies for scoring were identified across all the ELD standards in the assessment.

Findings

The analysis of the data revealed that, when scoring the ELD Classroom Assessment, teachers did not always interpret the standards consistently. Of the 29 standards in the ELD level 3 Classroom Assessment, 16 (55%) were interpreted in different ways by the teachers. Different interpretations occurred most often when a standard contained ambiguous language or when a standard contained multiple parts. When confronted with ambiguous language teachers would (a) assign their own meaning to the standard, (b) ignore the part that they perceived as ambiguous, or (c) ignore the entire standard. When faced with a standard that contained multiple parts, teachers would either (a) focus on one part only, or (b) find a way to "average" all parts. This process is illustrated in Figure 1:

Figure 1
How Teachers Interpret the Standards



This section presents six examples from the data that illustrate how teachers interpret the California ELD standards and make decisions about students' language ability.

Example 1

One of the "Reading—Literary Response and Analysis" standards for ELD level 3 states: "Apply knowledge of language to derive meaning/comprehension from literary texts." While scoring Mayra's assessment, one of the teachers said:

> Again, we went over this again today—going over our vocabulary in class. And she was very proficient at knowing that if a "–tion" is at the end of a word, that means that it's a noun. So if I know it's a noun I can start to understand what that word means. And so we actually went over that today and she did a great job at that. (Wrote down a score of 4).

This teacher interpreted "knowledge of language" to mean knowledge of vocabulary, specifically knowledge of word parts (affixes and suffixes) to derive the meaning of words. Another teacher decided to ignore the part of the standard that she perceived as ambiguous when scoring Raul's assessment:

> I'm never sure what they mean by "apply knowledge of language," but he can definitely find meaning and comprehension in the text so for these kinds of things I just use my best judgment. So I'd give him a 4.

For Raul's teacher, a score of 4 on this standard means that Raul can find meaning and comprehension in the text, whereas for Mayra's teacher a score of 4 means that Mayra has knowledge of word parts and can use that knowledge to understand the meaning of words. Clearly, the scores of 4 for these two students mean different things. When faced with the ambiguous phrase, "apply knowledge of language," in the standard, their teachers chose different strategies: Mayra's teacher assigned her own meaning to the phrase, whereas Raul's teacher ignored it. As a result, the students' identical score on the standard represents mastery of different abilities.

Example 2

The following is another example of teachers' strategies for dealing with a standard that contains ambiguous language. The "Reading—Word Analysis" standard for ELD level 3 states: "Use common English morphemes in oral and silent reading."

This is what Manuel's teacher said while scoring his assessment:

> These are so hard because how do I know he's applying knowledge of morphemes especially in such a big class? Unless he says: "Oh, I'm noticing that this word has an ending and that this ending is changing the meaning of the word," it's really hard to tell. So what I do, if he's doing well in reading, I say a 3. He is deriving meaning from what he's reading.

Manuel's teacher seemed confident about the meaning of the standard but was not sure how to gather evidence of the student's mastery of this standard. Thus, she decided to base her score on Manuel's general reading ability and his ability to derive meaning from text, which is a skill not mentioned in the standard. Another teacher, on the other hand,

chose to completely ignore the standard when scoring Ernesto's assessment:

> I really don't know for sure; they say in oral and silent reading. Oh, that's a strange one. I would say a 3 because I can't be sure. I don't think anyone can be sure.

Ernesto's teacher was unsure about the meaning of the standard and to be "safe" and not penalize him, she assigned a score of 3. Manuel's score of 3 on this particular standard means that he can read and derive meaning from text. Ernesto's score of 3 is not very meaningful.

Example 3

The following example involves a standard that at first glance does not appear to be ambiguous but that lends itself to many interpretations upon closer inspection. One of the "Reading—Fluency and Systematic Vocabulary Development" standards for ELD level 3 reads: "Create a simple dictionary of frequently used words." When scoring Jessica's assessment, her teacher said the following:

> She doesn't do that yet. I mean, she doesn't create one for herself. She doesn't take the initiative to say, "Oh, that's a new word. I should put that down." She doesn't do that. . . . Some do. I have quite a few students that do that, but in her case she doesn't.

This teacher interpreted this standard as an activity that students are responsible for doing proactively: She expected Jessica to create a dictionary on her own. It is unclear whether the teacher has taught this activity as a strategy, but it does not seem that she requires students to create a dictionary. Another teacher, however, had a very different interpretation of the standard. While scoring Gustavo's assessment, she said:

> I should have them do that. Because I don't and so it's a funny thing to be grading them on. It's something that I should do.

This teacher understood the standard to mean that she is responsible for assigning the activity of creating a dictionary to her students, while the former teacher expected students to take the initiative to create a dictionary on their own. Neither Cecilia nor Gustavo created a dictionary, but while Cecilia was given a score of 2 for not taking the initiative to do it, Gustavo was not scored at all because the teacher did not expect him to do it.

Yet a third teacher, unsure of what the standard means, assigned an entirely different meaning to the standard. While scoring Cecilia's assessment, she said:

> She'd be able to do that with some help. She might not know, and I'm not sure exactly what they're asking when they ask this, but I'm assuming that they mean like alphabetical skills and those kinds of things and the meanings of words, and she could possibly do that. But I'd probably give her a 2 because she would not be able to define as many words as, say, Alex would or somebody at the same ELD level.

This teacher focused on Cecilia's alphabetical skills, knowledge of the meaning of words, and the ability to define words. She did not interpret this standard as requiring students to literally create a dictionary in the way the other two teachers did. Cecilia's score of 2 has a very different meaning from Jessica's score of 2.

The examples above paint a picture of the problems present when teachers are confronted with language in a standard that is—or is perceived as being—ambiguous. The next three examples show the strategies that teachers use when faced with standards that have multiple parts.

Example 4

The following standard from the "Writing—Strategies and Applications" section of the ELD level 3 Classroom Assessment

states: "Independently create cohesive paragraphs that develop a central idea with consistent use of standard English grammatical form. (Some rules may not be in evidence)." When scoring Edwin's assessment, his teacher focused on one part of the standard—the part that requires independent action:

> A 2. He's not really independent. If I sit and I work with him, he does a lot. He can write a paragraph, but I have to really keep him on task and keep him focused. So independently he can't do it. It would be a 2.

But Cinthia's teacher focused on the second part of the standard when scoring her assessment—"creat[ing] cohesive paragraphs that develop a central idea":

> Right now I'd say a 2 only because she's one that has problems with that in her writing. The cohesive paragraphs, trying to put that together, having a topic sentence and supporting one—it's a complex process and we work on main ideas. She's doing it, better and better with each writing we do, but it's still…a work-in-progress situation, so…a 2.

Once again, Edwin's and Cinthia's identical scores mean different things. A score of 2 means that Edwin can write a paragraph, but only with assistance. In Cinthia's case, a 2 means that she has difficulties writing a paragraph regardless of whether she is doing it independently or not. When scoring José on this same standard, another teacher chose to focus on the third part of the standard: "with consistent use of standard English grammatical form. (Some rules may not be in evidence)":

> [W]hat I read from this is that they're mostly concerned that he can write a paragraph using standard English because writing a paragraph, a cohesive paragraph that develops the central idea, is difficult for any fourth-grader. So I would give him a 3 because I think they're mostly focusing on the fact that he can use standard English.

José was assigned a score of 3, which means that unlike Cinthia and Edwin who received a 2, he has mastered the standard. Yet according to the teacher's statement, José cannot write a cohesive paragraph that develops a central idea; he can only write a paragraph using standard English.

In short, the data suggest that when a standard contains multiple parts, the extent to which a student is determined to have mastered that standard often reflects mastery of the part of the standard on which the teacher chose to focus.

Example 5

The following is another example of a standard with multiple parts and the ways in which a teacher struggles to consider all parts before assigning a score. The "Reading Comprehension" standard for ELD level 3 states**: "**Read and orally identify examples of fact/opinion, and cause/effect in literature and content area texts." While scoring Ruby's assessment, a teacher expressed the following:

> I would say she would get a 3 for fact—she could probably even get a 4 for fact and opinion. But the cause and effect is a little harder for her to grasp, and so we'll just go with a 3. The cause and effect, they're still getting them confused, which is the cause and which is the effect. . . . But the fact and opinion, all the kids know that.

In this example, the teacher admitted that Ruby exceeds the standard for understanding fact versus opinion, but she is not yet able to understand cause and effect. Because both skills are part of the same standard, the teacher felt compelled to "average out" the two skills and assigned Ruby a score of 3, even though she has not mastered cause and effect.

Example 6

The last example also shows how a teacher struggles to address all parts of the standard to come up with one score. One of the "Listening and Speaking" standards for ELD level 3 reads: "Actively participate in social conversations with peers and adults on familiar topics by asking and answering questions and soliciting information." Alex's teacher carefully tried to attend to each part of the standard to come up with a single score:

> I would give him a 3. He's usually a 2. Well, he would be a 4 with his peers. He has no problem with his peers. With adults, it would probably be a 2. But when it's a familiar topic he does great. Because…[H]e had attended space camp, and he wants to be an astronaut. So when we had the story about that, he was sharing every day. I would've given him a 4 that week. But overall, I would say a 3.

As in the previous example, this one shows the problems teachers face when they are forced to assign one score to assess a student on a standard that has multiple parts. The resulting score is not a true reflection of the student's mastery of the standard but rather a compromise or "average" of the multiple abilities assessed. These examples demonstrate that when a standard has multiple parts and only one score can be used to rate mastery of the standard, the score is not particularly meaningful. Students' abilities might be better described with individual scores for each part of the standard.

In summary, the data show that teachers do not always interpret the ELD standards consistently when scoring the ELD Classroom Assessments. When faced with ambiguous language or standards with multiple parts, teachers resort to one or more strategies: They assign their own meaning to the standard, they ignore it in part or in whole, they focus on one part only, or they average all parts. As a result, the same scores on the assessment can represent different levels of mastery or types of proficiency for different students.

Discussion

Bachman and Palmer (1996) define validity as "the meaningfulness and appropriateness of the interpretations that we make on the basis of test scores" (p. 21). They explain that, to justify a particular score interpretation, evidence must be gathered to confirm that "the test score reflects the area(s) of language ability we want to measure, and very little else" (p. 21). The present study shows that the scores on the ELD Classroom Assessment reflect more than just students' language ability; scores also reflect teachers' interpretations of the standards. In addition, other factors not explored in this paper are reflected in the scores as well, including teachers' interpretation of the scoring criteria, students' personalities and behavior, external pressure to advance students to the next level, and teachers' beliefs about assessment and grading. Overall, the evidence gathered from the present study calls into question the validity of score interpretations.

So, why are teachers *not* interpreting and using the ELD standards consistently when assessing their students? There are at least three possible explanations.

One reason might have to do with the standards themselves. Content standards have often been criticized for being too broad, too ambitious, and too numerous for teachers to address effectively (Marzano & Kendall, 1997; Popham, 2003). Teachers in the present study often complained about the "vagueness" of the ELD standards, particularly when the language of the standards included "caveats" or "qualifications" of the ability—statements such as "may include some inconsistent use of" or "some rules may not be in evidence." Teachers also identified as ambiguous standards that included modifiers such as "more expanded vocabulary," "detailed," "more complex." Given the central role that standards play in instruction and assessment, studies should be conducted to examine the

quality, clarity, and importance of content standards (Lane, 2004).

Another reason why teachers do not interpret the ELD standards consistently might be lack of professional development. Of the 10 teachers in the study, 3 had never received training on the ELD standards or the use of the ELD Classroom Assessment, while the remaining 7 vaguely remembered attending a professional development session some years before. This finding is consistent with those of a recent survey of teachers of English learners in California that found that, in the last 5 years, many teachers of English learners had received little or no professional development on issues related to English learners (Gándara, Maxwell-Jolly, & Driscoll, 2005). At the beginning of the school year, teachers receive a memo about the ELD Classroom Assessment that encourages them to discuss their scores with other grade-level teachers. But teachers often do not have the time to engage in these conversations. In fact, in his 1998 study, Stecher found that classroom assessments place additional burdens on teachers, who are already strapped for time.

A third possible explanation might have to do with alignment. Even though there is a perfect alignment between the California ELD standards and the ELD Classroom Assessment—after all, the standards listed in the assessment correspond directly to the California ELD standards—there is no direct alignment between the standards in the ELD Classroom Assessment and the instructional program. The Reading/Language Arts curriculum that teachers use is aligned to the English Language Art (ELA) standards, not the ELD standards. Even though there might be some overlap, there is not a clear alignment between the ELD standards and what teachers are focusing on in the classroom. As evident in some of the examples discussed in this paper, the lack of alignment between what is covered in class and the standards measured in the assessment makes it difficult for teachers to determine whether a standard has been met, especially when they have not covered the content in class or engaged students in activities that would allow them to demonstrate mastery.

Regardless of the reasons, the fact that the extent to which a student is determined to master a standard is in part based on a particular teacher's interpretation of that standard not only has implications for the validity of the ELD Classroom Assessment itself, but it also calls into question the effectiveness of standards-based reform, at least as it concerns English learners. The goal of standards-based reform is to improve the quality of education by aligning the entire system—curriculum, assessment, professional development, and funding—to standards set at the state level. These standards are at the core of such a system so, unless there is a clear and common understanding of the standards, the system cannot improve the quality of education. In areas such as English Language Arts and mathematics, where the pressures of accountability are greatest, the school district in this study adopted and supported the implementation of a curriculum aligned to California's set of standards. But with so much attention focused on English Language Arts and mathematics, the ELD standards are not a priority for teachers. This "second-class" status of ELD standards and ELD instruction might ultimately explain why teachers do not receive adequate professional development on the standards and as a result do not interpret standards consistently when assessing their students.

In spite of some of the issues raised in this paper, classroom assessments such as the ELD Classroom Assessment have the potential to be useful assessment instruments to promote student learning. For one, these assessments provide teachers with immediate and practical information about their students that can inform their instruction. In fact, 7 of the 10 teachers reported that they use the assessment in formative ways. A few use the assessment to identify students who are not doing well and provide those students with additional assistance, sometimes through an instructional aide. Others reported that scoring the assessment gives them

new teaching ideas and reminds them of content and activities that they should cover in class. Among the many benefits of using classroom assessments, Stecher (1998) also found greater enthusiasm for teaching, higher expectations for students, and desired changes in educational goals, content, and instructional procedures.

Many of the issues raised in this paper might be resolved if teachers were trained and retrained on a regular basis on the ELD standards and the use of the ELD classroom assessment. Professional development should focus on the meaning of the standards, the scoring criteria, and examples of student work. Also, the alignment between the ELD Classroom Asessment and the instructional program needs to be explicitly articulated for teachers to take advantage of the connections that already exist between the instructional program and the assessment. The Map of Standards for English Learners created by WestEd (Carr & Lagunoff, 2003) is a helpful resource that maps the ELD standards to the ELA standards, but it provides minimal guidance for designing instructional activities. There is a need for specific instructional activities to help teachers address those ELD standards that are not matched directly to the ELA standards and as a result are not addressed by the English Language Arts curriculum.

In conclusion, several issues need to be resolved to ensure that standards-based classroom assessments are used effectively to promote student learning and for accountability. But their potential should not be wasted. As Lane (2004) and many other researchers have pointed out, we cannot rely on any one measure alone. The effective use of high-quality classroom assessments that reflect the standards should be an integral part of a "cohesive, balanced assessment system" aimed at promoting student learning (p. 13).

This research was supported by grants from the University of California Linguistic Minority Research Institute (UC LMRI) and the Spencer Foundation. Opinions reflect those of the author and do not necessarily reflect those of the grant agencies.

Author

Lorena Llosa completed her Ph.D. in Applied Linguistics at the University of California, Los Angeles in June 2005. She is now an assistant professor at New York University's Steinhardt School of Education. Her research interests include second and foreign language learning and teaching, language testing, program evaluation, and research methods.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.

Brindley, G. (1998). Outcomes-based assessment and reporting in language learning programmes: A review of the issues. *Language Testing, 15*(1), 45-85.

Brindley, G. (2001). Outcomes-based assessment in practice: Some examples and emerging insights. *Language Testing, 18*(4), 393-407.

Brown, A. (1995). The effect of rater variables on the development of an occupation-specific language performance test. *Language Testing 12*(1), 1-15.

Carr, J., and Lagunoff, R. (2003). *The map of standards for English learners: Integrating instruction and assessment of English language development and English language arts standards in California* (4th ed). San Francisco: WestEd.

Chalhoub-Deville, M. (1995) Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*(1), 16-33.

Cohen, A .D. (1994). English for academic purposes in Brazil: The use of summary tasks. In C. Hill & K. Parry (Eds.), *From testing to assessment: English as an international language*. London: Longman.

Cumming, A. (1990). Expertise in evaluating second language composition. *Language Testing, 7*(1), 31-51.

Cumming, A., Kantor, R., & Powers, D. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making, and development of a preliminary analytic framework.* (TOEFL Monograph Series.) Princeton, NJ: Educational Testing Service.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data (Rev. ed.).* Cambridge, MA: The MIT Press.

Faerch, C., & Kasper, G. (1987). *Introspection in second language research.* Clevedon, Great Britain: Multilingual Matters.

Gándara, P., Maxwell-Jolly, J., & Driscoll, A. (2005). *Listening to teachers of English language learners: A survey of California teachers' challenges, experiences, and professional development needs*. Santa Cruz, CA: Center for the Future of Teaching and Learning.

Gipps, C. (1994). *Beyond testing: Towards a theory of educational assessment*. London: The Falmer Press.

Koretz, D. (1998). Large-scale assessment assessments in the US: Evidence pertaining to the quality of measurement. *Assessment in Education, 5*(3), 309-334.

Lane, S. (2004). Validity of high-stakes assessment: Are students engaged in complex thinking? *Educational Measurement: Issues and Practice, 23*(3), 6-14.

Lumley, T. (2000). *The process of the assessment of writing performance: The rater's perspective.* Unpublished doctoral thesis, University of Melbourne, Melbourne, Australia.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246-276.

Lumley, T. J. N., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing 12*(1), 54-71.

Marzano, R. J., & Kendall, J. S. (1997). *The fall and rise of standards-based education: A National Association of School Boards of Education (NASBE) issues in brief*. Aurora, CO: Mid-Continent Research for Education and Learning.

Milanovic, M., Saville, N., and Shen, S. (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th language testing research colloquium, Cambridge and Arnhem* (pp. 92-114). Cambridge, England: Cambridge University Press and University of Cambridge Local Examinations Syndicate.

North, B. (1993). *The development of descriptors on scales of language proficiency*. Washington, DC: The National Foreign Language Center.

Popham, J. W. (2003). The trouble with testing: Why standards-based assessment doesn't measure up. *American School Board Journal, 190*(2), 14-17.

Rea-Dickins, P., and Gardner, S. (2000). Snares and silver bullets: Disentangling the construct of formative assessment. *Language Testing, 17*(2), 215-243.

Stecher, B. M. (1998). The local benefits and burdens of large-scale portfolio assessment. *Assessment in Education, 5*(3), 335-351.

U.S. Department of Education (n.d.) *No child left behind act of 2001.* Retrieved September 12, 2005, from http://www.ed.gov/nclb/landing.jhtml

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts,*(pp. 111-125). Norwood, NJ: Ablex.

Weigle, S. C. (1994). *Effects of training on raters of English as a second language compositions: Quantitative and qualitative approaches*. Unpublished doctoral dissertation, University of California, Los Angeles.

Zhang, W. (1998). *The rhetorical patterns found in Chinese EFL student writers' examination essays in English and the influence of these patterns on rater*

response. Unpublished doctoral thesis, Hong Kong Polytechnic University, People's Republic of China.

Appendix
Instructions Given to Teachers for Think-Aloud Task
(Adapted from Lumley, 2002)

"I am now going to ask you to score two of your students using the ELD Classroom Assessment. I would like you to score them as far as possible in the usual way, that is, just as you have done it in the previous reporting period. However, there will be one important difference this time: As I have previously mentioned, I am conducting a study of the processes used by teachers when they score the ELD Classroom Assessment, and I would now like you to talk and think aloud as you score these assessments, while this tape recorder records what you say.

First, you should state the student's current ELD level. Then you should read each standard out loud as you start to score. Then, as you score each standard, you should vocalize your thoughts, and explain why you give the scores you give.

It is important that you keep talking all the time, registering your thoughts all the time. If you spend time reading from the ELD Classroom Assessment, you should do that aloud also, so that I can understand what you are doing at that time. In order to make sure there are no lengthy silent pauses in your scoring, I propose to sit here and prompt you to keep talking if necessary. I will sit here while you rate and talk. I will say nothing more than give you periodic feedback such as 'mhm,' although I will prompt you to keep talking if you fall silent for more than 10 seconds."