The CATESOL Journal

KATIE STOKES-GUINAN
Stanford University

CLAUDE GOLDENBERG
Stanford University

Use With Caution: What CELDT Results Can and Cannot Tell Us

Both California state law and the federal No Child Left Behind Act (NCLB) require that all schools assess the English language proficiency of newly enrolled students who speak a language other than English at home and, annually, all English learners (ELs) already enrolled. California meets this requirement by administering the California English Language Development Test, or CELDT. The CELDT has three primary purposes: to identify students who are ELs, determine their English proficiency level, and assess their progress in acquiring listening, speaking, reading, and writing skills in English through time. We examine data on the validity and reliability of the CELDT to determine if it is an appropriate tool for carrying out these purposes. We conclude that the CELDT is likely a sufficiently valid and reliable tool for making judgments about *groups* of students but not for making crucial educational decisions about *individual* students.

California's public schools educate approximately 6.3 million students in kindergarten through 12th grade. About 1.5 million of these students, or 25%, are classified as English learners (ELs); in other words, they are not native English speakers and do not speak or understand English well enough to benefit adequately from mainstream classroom instruction.

Both California state law and the federal No Child Left Behind Act (NCLB) require that all schools assess the English language proficiency of newly enrolled students who speak a language other than English at home and, annually, all English learners already enrolled (California Department of Education, 2008). California meets this requirement by administering the California English Language Development Test, or CELDT, developed by CTB/McGraw-Hill. The CELDT is a highly consequential assessment. Decisions based on CELDT data affect individual students, schools, districts, and the state as a whole. It is important for educators and policy makers to understand its strengths and limitations. In particular, we argue that the CELDT is likely a sufficiently valid and reliable tool for making judgments about *groups* of students. However, because of issues with construct validity, interrater reliability, and nonstandard administration, the CELDT may not be sufficiently valid and reliable for

making crucial educational decisions about *individual* students. Finally, even though the present paper is focused on the use of the CELDT, issues discussed here could have implications for other language proficiency tests used with ELs nationwide.

Overview of the CELDT and Its Uses

The CELDT is the principal means by which California students are identified as English learners (sometimes referred to as "limited English proficient"). It is aligned to California's English Language Development Standards (California Department of Education, 2002) and provides a measure of students' overall English language proficiency as well as their proficiency in four language domains—listening, speaking, reading, and writing (CTB/McGraw-Hill, 2005).[1]

The CELDT has distinct forms for each of four grade spans: kindergarten through grade 2, grades 3 through 5, grades 6 through 8, and grades 9 through 12. Students receive a performance-level score ranging from 1 to 5 for each of the domains and an overall performance-level score calculated by weighing the domain scores equally. The five performance levels are:

1. Beginning;
2. Early Intermediate;
3. Intermediate;
4. Early Advanced; and
5. Advanced.

CELDT scores are used for initial classification of students as English learners, reclassification of English learners to fluent English proficient, district accountability, and making school-level decisions. We address each of these in turn.

Initial Classification

Upon their children's entry into California schools, all parents must fill out a Home Language Survey that includes questions about the child's first language, the language the child speaks most frequently at home, the language parents speak most frequently to the child, and the language most often spoken by adults in the home (California Department of Education, 2005). If a student's parents indicate on the Home Language Survey that a language other than English is spoken in the home, California law requires schools to administer the CELDT within 30 days of the student's enrollment. In general, a student who receives an overall score of Early Advanced or Advanced and achieves individual domain scores of at least Intermediate on his or her first administration of the CELDT is classified as initially Fluent English Proficient (IFEP). A student whose overall score is *either* (a) Intermediate or below or (b) Early Advanced or Advanced, but with one or more domain scores below Intermediate, is classified as an English learner (EL). However, the California Department of Education's position is that these are guidelines rather than strict rules. Although the education code requires districts to use students' CELDT scores as the primary indicator for initial classification, students who score in the up-

per end of Intermediate on the CELDT *can* be considered as IFEP if other data warrant this classification (California Department of Education, 2009). After initial classification, districts must readminister the CELDT annually to all ELs between July 1 and October 31.

Reclassification

ELs are eligible for reclassification to Fluent English Proficient (RFEP) once they score at least Early Advanced overall with no domain scored below Intermediate. Note that this score makes ELs *eligible* for reclassification; it does not automatically mean that they are reclassified. Before a student can be reclassified, state law requires teacher evaluation of the pupil's mastery of the curriculum, parental opinion and consultation, and at least one measure of academic achievement (often, but not necessarily, a standardized test score) indicating that the student is "sufficiently proficient in English to participate effectively in a curriculum designed for pupils of the same age whose native language is English" (California Department of Education, 2008, p. III-2).

District Accountability

Schools and districts use CELDT data in other ways besides informing the decision to classify or reclassify students as English learner (EL) or fluent English proficient (FEP). CELDT scores are used for federally mandated district accountability. Districtwide CELDT data determine whether districts have met the two Annual Measurable Achievement Objectives (AMAOs) for ELs' English language acquisition (Linquanti & George, 2007). AMAOs are performance targets that districts receiving Title III federal funds must meet each year. AMAO 1 addresses the percentage of ELs making adequate progress on the CELDT; AMAO 2 measures the percentage of ELs who have attained English proficiency. Districts failing to meet AMAOs over multiple years face increasingly serious sanctions under the federal No Child Left Behind Act (NCLB, 2002).

School Level Uses

Finally, school administrators use CELDT data in various ways to monitor, improve, and report on programs. Williams, Hakuta, Haertel, et al. (2007) report that 95% of principals say they use CELDT data to:

> evaluate the progress of students and communicate with parents, and nearly as many used it to identify struggling students (87%) and develop strategies for moving them toward English-language proficiency (78%). A substantial majority (71%) of principals said they used CELDT data to examine school-wide instructional practices. (p. 13)

Clearly, the CELDT is used to provide important information and to make consequential decisions. Students may be placed into special programs based on their CELDT performance, districts may face sanctions under No Child Left Behind for failing to increase CELDT scores, and parents and others—such as com-

munity members, policy makers, and the public at large—draw conclusions based on students' CELDT scores. These consequences clearly require that the tool be both a valid and reliable measure of students' English language proficiency.

Is the CELDT an Adequate Tool?

The CELDT has three purposes: (a) identify students who are ELs; (b) determine the English proficiency level of students who are ELs; and (c) assess the progress of ELs in acquiring listening, speaking, reading, and writing skills in English through time (CTB/McGraw-Hill, 2007; Educational Data Systems, 2009).[2] Does the CELDT adequately serve these purposes? NCLB mandates—and sound educational practice requires—the use of valid and reliable English proficiency assessments for English learners (NCLB, 2002). So before we attempt to answer this question, we will begin by discussing the key concepts of validity and reliability as they pertain to the CELDT.

Construct Validity

Construct validity refers to the degree to which a test accurately reflects or assesses the specific knowledge, skills, or abilities it purports to measure. As defined by Messick (1995), construct validity "comprises the evidence and rationales supporting the trustworthiness of score interpretation in terms of explanatory concepts that account for both test performance and score relationships with other variables" (p. 743). In the case of the CELDT, the test is supposed to measure English language proficiency. One immediate problem is that there is no generally accepted definition of "language proficiency" or of the various levels that describe degrees of proficiency, such as Beginning, Early Intermediate, and so forth (Abedi, 2008; Bialystock, 2001). Although this can hardly be blamed on the CELDT, it does create an enormous challenge for ensuring the CELDT's construct validity, since different definitions of English proficiency can lead to different conclusions regarding students' presumed proficiency levels. At a minimum, educators and policy makers must realize how tenuous our definition of "language proficiency" is and the uncertainty this creates for language testing and the interpretation and use of test scores.

This challenge is illustrated in a study conducted by CTB/McGraw-Hill (2005) examining how accurately the CELDT categorizes students' English proficiency levels. Since there are no absolute criteria that determine "language proficiency," the best we can do is compare the results of one approach to the results of another. CTB/McGraw-Hill had English language development experts conduct independent assessments of the English proficiency of 1,384 students and then compared the experts' ratings to the proficiency levels assigned by the CELDT.[3]

The results showed that the experts and the CELDT classified students into the same proficiency level just 40% of the time; 50% of the time the experts and the CELDT were one proficiency level off—for example, the CELDT indicated Early Intermediate and the expert rater indicated Intermediate. In 90% of cases, therefore, the CELDT and the experts either agreed exactly or were within

a proficiency level of each other. This means, then, that *60% of the time* the CELDT and the experts disagreed on students' exact English proficiency levels, but only 10% of the time was the difference greater than one proficiency level.[4] Readers should keep in mind that these data are from the 2003-2004 CELDT. It might be that with more recent versions, and a new test publisher in 2009-2010 (Educational Data Systems), concordance with "expert raters" might improve. We know of no data to gauge this possibility, however.

A test that might incorrectly estimate students' proficiency level up to 60% of the time is problematic. But it is possible that the CELDT, while not very good at pinpointing a student's exact level of proficiency, might still effectively distinguish ELs from non-ELs. Making this distinction is arguably the most important use of the CELDT. On this criterion, CELDT scores held up better. Experts and CELDT scores agreed whether students should be classified as EL or English proficient 70% of the time. Although substantially better than the 40% agreement on proficiency levels, this still indicates that up to 30% of students might be incorrectly classified by the CELDT. CTB/McGraw-Hill's (2005) study raises questions about the validity of CELDT scores when used to make decisions about individual students' language proficiency classification. On the other hand, it might raise questions about the validity of the expert raters' judgments, since we cannot say with certainty which classification is correct. But perhaps most important, the substantial lack of agreement raises questions about any attempt to measure precisely, using a single instrument, a construct such as language proficiency that at present lacks a consistent and sufficient definition. This lack of a precise definition is a serious problem because important decisions are made based on which side of the EL/non-EL dividing line students fall.

Messick (1995) brings up these types of issues in his discussion of *consequential validity*, which he defines as an aspect of construct validity that deals with the consequences of score interpretation:

> Social consequences of testing may be either positive, such as improved educational policies based on international comparisons of student performance, or negative, especially when associated with bias in scoring and interpretation or with unfairness in test use. (p. 746)

The notion of consequential validity is relevant to the CELDT because the magnitude of the problem of inaccurate classification varies depending on whether the CELDT is being administered to a student for the first time for initial classification or is being administered to an already identified EL to assess whether he or she should be reclassified to fluent English proficient. When a student is administered the CELDT for the first time (i.e., initial classification), the CELDT score is typically the only criterion used to make the decision. This fact means that an inappropriately high CELDT score will prevent a student who may need language supports from being identified as an EL, and an inappropriately low CELDT score will result in a student's receiving an unwarranted EL designation.

When an EL student is administered the annual CELDT for assessment purposes, however, he or she is not reclassified based on CELDT scores alone. State guidelines for reclassifying students from EL to RFEP require that schools consider multiple criteria in addition to the CELDT, such as performance on the California Standards Test in English Language Arts, local district criteria, and parent opinion and consultation (California Department of Education, 2008). For reclassification decisions, this means that an inappropriately high CELDT score is unlikely to result in an EL's being reclassified because the student will most likely not meet the other performance criteria for reclassification. Thus, an inappropriately high score on the annual CELDT is not likely to have negative repercussions for students.

An inappropriately low CELDT score, on the other hand, will prevent a student from being considered for reclassification—a potentially harmful consequence. This will be particularly true if other indicators, such as achievement measures and teacher judgments, are not even consulted until students first reach a determined performance threshold on the CELDT. Because we cannot assume that language proficiency is being measured precisely by the CELDT, as students are approaching the reclassification criterion (i.e., Early Advanced overall with no domain scored below Intermediate), districts should look to other indicators, such as achievement measures and teacher evaluation, and consider taking these into account as they consider reclassification.[5] Looking to other indicators before students have reached the required performance level on the CELDT obviously runs the risk of prematurely reclassifying them as fully proficient in English. Therefore, schools and districts must of course use judgment and caution when making these decisions. But there are no simple solutions, since as we have indicated, making absolute judgments about language proficiency levels is fraught with challenges.

Yet we must make our classification decisions as accurate as possible because a student's classification may make him or her eligible for some types of programs and services but ineligible for others. Students identified as ELs, for example, are eligible to receive special language supports and may be placed into English Language Development programs. At the same time, many EL students are also denied access to elective courses or Advanced Placement classes (Callahan, 2005). Non-EL students may have opportunities to take advanced classes but may not have access to language supports (which presumably are not needed). Clearly, the different opportunities available to EL and non-EL students underscore the importance of accurate classification.

The problems identified above are substantial when we refer to making judgments about individual students. They might be less serious, however, for making judgments for relatively large *groups* of students. Assuming that discrepancies in proficiency designations or classification decisions are random, the CELDT is equally likely to produce incorrectly high and low scores. Thus, average scores for groups of students are typically less problematic than any one individual student's score. This fact is generally true for all tests (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999; Haertel, 2006).

Reliability

Reliability refers to the degree to which a measure is consistent and dependably produces the same result. Reliability is distinct from validity in the sense that a test can measure something reliably or consistently even if it is not a true or valid measure of whatever is being measured, such as oral language proficiency. However, a test that is not reliable can never be valid, since an unreliable score cannot be considered a dependable gauge of the construct being measured.

There are several different types of reliability. One type that is often reported is internal consistency reliability. Internal consistency means that all of the items in a measure are indeed measuring the same construct. For example, an internally consistent reading-comprehension measure will have items that to one degree or another correlate with each other. This internal consistency then gives test users confidence that there is adequate coherence and substance to the measure and that it does not comprise a group of random items each measuring something different. CTB/McGraw-Hill (2007) reports internal consistency reliability coefficients for the 2006-2007 version of the CELDT (Form F) ranging from .80 to .93 across all grades and domains. It notes that "these are typical coefficients for assessments of these lengths" (p. 49). We do not disagree with this appraisal; nevertheless we note there is no hard and fast universal standard for judging what is "adequate" internal consistency.

A second type of reliability is test-retest reliability. Test-retest reliability indicates whether a student who takes the same test twice within a short span of time receives the same, or similar, score each time. There do not appear to be any data on the CELDT's test-retest reliability.

Another reliability estimate, for which we do have CELDT data, is interrater reliability. This is an indication of the degree to which a measure will yield similar results when two or more individuals administer and/or score the assessment independently. If two judges consistently give the same score to a writing sample, that indicates high interrater reliability.

Written responses on the CELDT are scored at a central location, and since 2005-2006, 100% of student responses to the writing constructed response items are scored by two people to verify the scores. CTB/McGraw-Hill (2007) reported interrater agreement percentages for the sentence-writing and composition-writing sections of the CELDT Form F. For the sentence-writing questions, two raters provided the same rating for 77.2% to 85.8% of the questions. Score discrepancies greater than 1 point on a 4-point rubric scale occurred in only 0.8% to 5.6% of the cases, indicating reasonably high interrater reliability within 1 score-point. Agreement on the writing-composition section of the test was lower. Depending on the grade span, two raters scoring a written response provided the same score just 66.0% to 71.2% of the time. Discrepancies greater than 1 point occurred 1.6% to 5.3% of the time. CTB/McGraw-Hill (2007) states that these percentages are considered to be within acceptable industry standards. Again, we have no quarrel with this statement but would note that there is no universal agreement about what constitutes adequate agreement.

Nonstandard Administration

The CELDT, as with any standardized measure, particularly one that is individually administered, is also susceptible to problems of nonstandard administration. These problems can affect both the validity and reliability of the scores.

Both the original CELDT contractor, CTB/McGraw-Hill, and the new contractor, Educational Data Systems, offer training for school and district staff who administer the CELDT to promote uniform administration and scoring of the test. In practice, however, schools often send to the training one or two people who then train others at their site, making the chances of nonstandard administration substantial.

In fact, we have at least anecdotal evidence that administration varies. A CELDT administrator at a Northern California school interviewed for this article reported directing her staff to ask questions out of order and split some questions into two parts to make them more comprehensible to students. She also directed her staff to mark nonstandard answers as correct, such as if a student offered "nurse" for an answer when asked to identify a picture of a doctor. We know of no data to indicate these practices are widespread. But to the extent they are, they not only compromise reliability—that is, consistency—but perhaps more important, these deviations also decrease validity because changing the questions changes what is being measured.

Again, these issues are not necessarily inherent to the CELDT itself, but they nonetheless threaten to compromise the data the CELDT provides. Specific recommendations for how to correct shortcomings of CELDT administration are beyond the scope of this paper. However, we would suggest that in addition to the administration materials that come with the CELDT to train testers, there must also be active monitoring by the state and districts to assure testing protocols are consistently followed. At present there does not seem to be sufficient attention to monitoring CELDT administration to assure uniform and consistent procedures. Monitoring is particularly important when there is high turnover among testers.

Because CELDT scores are used to make high-stakes decisions for individual students, the level of unreliability in the scoring of the CELDT might mean that a nontrivial number of students are receiving scores that are not indicative of their true abilities, thus increasing the likelihood that their academic placements will not be appropriate to their needs. Even though issues of inconsistent classification do not seem to be particularly problematic at the group level, these issues do become problematic when considering the likelihood that a *particular* student might be incorrectly rated or classified.

So given what we now know about the validity and reliability of the tool, we now consider the CELDT's suitability for carrying out its three purposes.

How Effectively Does the CELDT Carry Out Its Three Required Purposes?

Purpose #1: Identify students who are English learners.[6] For those students who do take the CELDT, it can be an adequate tool for identifying

which students are English learners only if it validly and reliably distinguishes between EL and non-EL students. On this count, the CELDT falls short. The CELDT probably misclassifies substantial numbers of students. CTB/McGraw-Hill (2005) suggests that an appropriate classification rate is 80%, meaning that by "industry standards," 20% of students could be misclassified. But the CELDT might not meet even this standard, since, as we have seen, the misclassification rate—when CELDT scores are compared to those of "expert raters"—could be as high as 30%. But again, it is hard to know what the potential misclassification rate is since we cannot know with certainty whether the validity of the CELDT should be measured against the standard of the experts or the validity of the expert judgments should be measured against the standard of the CELDT.

If students are indeed being misclassified, we cannot say for certain which of the factors we have identified, if any, may be responsible, such as faulty administration of the tool, something inherent in the instrument itself, or the more fundamental problem that there is no generally agreed upon definition—with clear criteria—for what constitutes English proficiency.

Purpose #2: Determine the English proficiency level of students who are English learners. If the CELDT is not sufficiently valid and reliable for classifying individual students as ELs, it is even less valid and reliable for determining a student's level of English proficiency, since this level of classification requires finer distinctions to be made than a simple EL/FEP decision. Indeed, the data indicate that the correct classification rate might be as low as 40% (CTB/McGraw-Hill, 2005). Again, it is not entirely clear why there is lack of precision in determining the various performance levels from non-English proficient to fluent English proficient; inconsistent administration, something inherent in the CELDT itself, or the absence of generally agreed-upon definitions of language proficiency levels—or some combination of these—may play a role.

As discussed in the previous section, there are inconsistencies between raters when scoring the same written sections of the CELDT. Scores are fairly consistent if we use as a criterion that they be no more than 1 score-point apart on the scoring rubric. Using this as the standard, written sections of the CELDT (the only ones for which there are published interrater reliability studies; we have no interrater reliability data for the speaking, listening, and reading sections) are consistently scored in 95-99% of cases. These data suggest that we can be fairly confident of English proficiency rating within a range of two levels; for example, a student is *either* Beginning *or* Early Intermediate or *either* Intermediate *or* Early Advanced. This is a lower level of precision than is acknowledged when CELDT scores are reported and used. However, it might be adequate if used as a more general guideline for any one student's English language proficiency, rather than as a definitive indicator.

Purpose #3: Assess the progress of English learner students in acquiring listening, speaking, reading, and writing skills in English through time. The CELDT was not originally designed to assess year-to-year progress except within the test form administered to each of four grade spans (K-2, 3-5, 6-8, and 9-12). Technically, the different forms of the test (K to grade 2, grades 3-5, etc.) are not "vertically equated," which means that language proficiency ratings

based on performance at one grade span (e.g., K-2) cannot be compared to ratings based on performance at another grade span (e.g., 3-5). Title III of NCLB, however, requires states to provide information on students' growth in English proficiency. Therefore, in the 2006-2007 Edition (Form F) CTB/McGraw-Hill (2007) introduced a new "common scale" that allows for interpretation of scores from one year to the next across grade spans. Because the common scale is relatively new, it is too early to tell how valid and reliable an assessment of progress the CELDT will turn out to be. At the level of the *individual* student, we are again concerned about how much confidence we can have in reports of growth through time when an incorrect language proficiency designation for individual students in any one year might be as high as 60%.

Other purposes of the CELDT. Originally, CTB/McGraw-Hill (2005) had identified two additional purposes of the CELDT over and above the three purposes mandated by the state. One purpose was to help determine the readiness of students for various instructional options. This purpose was problematic because the CELDT is not designed to be a formative test, meaning that it cannot diagnose what skills a student needs to work on. Thus, we support CTB/McGraw-Hill's decision to remove it as one of the stated purposes of the CELDT.

The other purpose originally listed by CTB/McGraw-Hill was to evaluate program effectiveness using aggregate data. Although CTB/McGraw-Hill stopped making this claim, we argue that providing some overall gauge of student performance at a classroom, grade, school, or district level might be the purpose for which the CELDT is best suited. As we discussed above, some validity and reliability criteria are different depending on whether we are making judgments about individuals or groups. Assuming sufficient numbers of test takers—and assuming inconsistencies in scores and ratings are random—the CELDT can provide useful overall data about how well *groups* of students are performing with respect to their language proficiency and skills, as measured by the CELDT.

Further, the CELDT is used throughout California, allowing it to be used for comparing schools and districts across the state. In fact, this may be its biggest strength because the CELDT appears to be a sufficiently reliable indicator of school- and districtwide progress, assuming a large enough population of ELs. One potential problem with using CELDT data to compare schools or districts, however, is that the CELDT is unlikely to be administered or scored uniformly at all sites. Thus, some part of a school or district's CELDT scores may be influenced by the administration and scoring of the test more than the actual English proficiency level of the students.

Given the challenges the CELDT faces in carrying out its three purposes, educators and policy makers must consider both the appropriate uses of the CELDT and how we might improve our ability to assess the English proficiency of students validly and reliably.

Recommendations for Use of the CELDT

The main strength of the CELDT is that, in principle, it is administered to all ELs in California, providing researchers and policy makers a valuable

source of cross-school and cross-district data. The ubiquity of the CELDT in California schools allows researchers, for example, to examine which school or district factors or practices lead to higher proficiency levels so that this information can be shared and used in improving outcomes (e.g., Gold, 2006). Policy makers can also use the data to assess the effects of various educational programs or make decisions about resource allocation. Further, we hope that the CELDT's new common scale (CTB/McGraw-Hill, 2007) will provide longitudinal data on how effectively schools and districts are meeting the needs of their EL students.

However, we have two principal concerns about the CELDT. First is whether it is a sufficiently valid and reliable measure of *individual* students' English proficiency. Second, and related to the first, is whether it should be used in isolation to make high-stakes decisions about students' initial language proficiency classification. Although an EL classification can help ensure that students of limited proficiency receive needed services and supports, an EL classification may not be helpful and, in fact, may be harmful when applied inappropriately.

Just as the CELDT is used in combination with other criteria to make reclassification decisions, it should similarly be one of several measures used to make *initial* classification decisions. Other measures could include local assessments, parent opinion and consultation, and teacher professional judgment. These measures, of course, present their own serious challenges, not the least of which are nonstandardization and the logistical challenges of trying to do all this at the beginning of the school year. At the same time, they provide some level of redundancy that may help increase the reliability and validity of EL designations (see Abedi, 2008, for an overview). For initial classification, it may be prudent to classify students provisionally, based on their CELDT scores, but allow that EL classification status to be adjusted during a certain period (e.g., 30-60 days), during which other measures and observations can be gathered to support or counterweigh the provisional identification.

We fully acknowledge the challenges that accurate second language assessment and classification pose and do not wish to offer glib prescriptions. Nonetheless, it would seem that a policy requiring the CELDT to be used in conjunction with other measures would provide a needed check and lower the probability of initial misplacement. Indeed, when the educational fates of about 1.5 million EL students hinge on their EL status, it is imperative to identify them correctly and provide them with appropriate educational services.

We also do not wish to appear to be singling out the CELDT for undue criticism. The issues we have discussed pertaining to the CELDT also apply to other measures of English language proficiency used by educators in other states. Developing valid and reliable measures of English learners' language proficiency and using assessment data in productive ways represent substantial challenges that many researchers and educators around the country are actively addressing. Interested readers should consult a very informative publication by Jamal Abedi and colleagues (2007) that further describes these challenges.

We cannot offer a quick or easy solution; to a large extent, the issues we have identified with respect to the CELDT reflect the state of knowledge and

practice in defining and measuring language proficiency. But we do urge caution in using and interpreting CELDT data as the search continues for creating measures and tools to help teachers understand children's developing language skills.

Acknowledgments

The authors would like to thank Jamal Abedi, Alison Bailey, Lauri Burnham-Massey, Anne Davidson, David Dolson, Jeanette Ganahl, Norm Gold, Jim Grissom, Ed Haertel, Robert Linquanti, Lily Roberts, Robin Scarcella, and Richard Schwarz for their help, suggestions, and feedback. We are especially indebted to Alison Bailey and Robert Linquanti for their generous feedback and assistance.

Authors

Katie Stokes-Guinan is a doctoral student in Psychological Studies in Education at Stanford University. Her research interests center on the education of linguistic minority students. She previously worked as the director of programs and quality control for a family resource center focused on Latino literacy and parent involvement.

Claude Goldenberg is a professor of Education at Stanford University. His research focuses on improving the academic achievement of language minority students. He is conducting studies on vocabulary and reading comprehension, effective classroom environments, and language and literacy development among Mexican children.

Notes

[1]Before the 2009-2010 school year, students in kindergarten and grade 1 were assessed only on their listening and speaking skills. Starting in 2009-2010, the State Department of Education has been directed by the legislature to implement an assessment of reading and writing in kindergarten and grade 1 in order to conform to NCLB Title III assessment requirements. However, to our knowledge the CDE has not yet determined whether to include CELDT reading and writing scores to identify ELs in kindergarten and 1st grade.

[2]CTB/McGraw-Hill (2005a) originally listed five functions of the CELDT: the three listed here, which were mandated by the state, and two additional ones offered by CTB/McGraw-Hill. We address these two functions later in the paper.

[3]The experts included 35 individuals identified by CTB/McGraw-Hill, the California Department of Education and the San Joaquin County Office of Education who met criteria that included proficiency in English, knowledge about English language development among English learners, and at least one additional qualification that could include experience as a certified teacher, CLAD or BCLAD certification, familiarity with California's English Language Development Standards, familiarity with the California Content Standards in Language Arts, or experience at an assessment center. Participating experts all participated in an orientation session before conducting their student assessments (CTB/McGraw-Hill, 2005a).

[4]We would also note that there was a correlation of .60 between CELDT scores and expert rater scores (CTB/McGraw-Hill, 2005a). However, a correlation tells us only *relative* agreement between two sets of scores; it does not tell us anything about *absolute* agreement between them. A correlation of .60 indicates that, to a moderate degree, if a student's CELDT score tended to be high (or low), then the rater's agreement for the same student also tended to be high (or low). Absolute agreement is perhaps a more important indicator in this case because decisions are made about educational programs for students based on presumed *absolute* judgments about their language proficiencies, not whether their proficiency levels tend to be high or low.

[5]Our thanks to Robert Linquanti for helping us understand this point.

[6]The CELDT is administered to all students whose parents indicate on the Home Language Survey that a language other than English is spoken in the home. However, anecdotal evidence from teachers and school administrators and a study by Abedi, Lord, and Plummer (1997) indicate that many parents do not indicate the use of a non-English language at home. When this happens, schools are authorized by law to administer the CELDT to students whom they believe to be ELs, but in practice it is not clear that many do. Thus, an unknown number of English learners are never assessed with the CELDT or officially identified as ELs. This is a larger policy issue beyond the scope of this paper.

References

Abedi, J. (Ed.). (2007). *English language proficiency assessment in the nation: Current status and future practice*. Davis: University of California, Davis.

Abedi, J. (2008). Classification system for English language learners: Issues and recommendations. *Educational Measurement: Issues and Practice, 27*(3), 17-31.

Abedi, J., Lord, C., & Plummer, J. (1997). *Language background as a variable in NAEP mathematics performance* (CSE Tech. Rep. No. 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological measurement*. Washington, DC: AERE, APA, NCME.

Bialystok, E. (2001). *Bilingualism in development: Language, literacy, and cognition*. Cambridge, England: Cambridge University Press.

California Department of Education. (2002). *English-language development standards for California public schools: Kindergarten through grade twelve.* Retrieved December 14, 2010, from http://www.cde.ca.gov/be/st/ss/index.asp

California Department of Education. (2005). *Home language survey.* Retrieved November 24, 2009, from http://www.cde.ca.gov/ta/cr/el/documents/hlsform.doc

California Department of Education. (2008). *Understanding and using 2009-*

10 individual results. Retrieved October 13, 2008, from http://www.cde.ca.gov/ta/tg/el/documents/celdt09astpkt1.pdf

California Department of Education. (2009). *Explaining 2008-09 summary results to the public*. Retrieved December 4, 2009, from http://www.cde.ca.gov/ta/tg/el/documents/celdtrptrslts0809.pdf

Callahan, R. M. (2005). Tracking and high school English learners: Limiting opportunity to learn. *American Educational Research Journal, 42*(2), 305-328.

CTB/McGraw-Hill. (2005). *Technical report for the California English Language Development Test (CELDT) 2003-2004 Form C*. Report submitted to the California Department of Education on January 31, 2005. Retrieved March 1, 2008, from http://www.cde.ca.gov/ta/tg/el/documents/techrpt3.pdf

CTB/McGraw-Hill. (2007). *Technical report for the California English Language Development Test (CELDT) 2006-07 Edition (Form F)*. Report submitted to the California Department of Education in May 2007. Retrieved May 23, 2009, from http://www.cde.ca.gov/ta/tg/el/documents/formftechreport.pdf

Educational Data Systems. (2009). *About CELDT*. Retrieved November 24, 2009, from http://www.celdt.org/about

Gold, N. (2006). *Successful bilingual schools: Six effective programs in California*. San Diego, CA: San Diego County Office of Education.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65-110). Westport, CT: American Council on Education/Praeger.

Linquanti, R., & George, C. (2007). Establishing and utilizing an NCLB Title II accountability system: California's approach and findings to date. In J. Abedi (Ed.), *English language proficiency assessment and accountability under NCLB Title III: A national perspective*. Davis: University of California.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

No Child Left Behind Act of 2001 (NCLB), Pub. L. No. 107-110, § 115 Stat. 1425. (2002).

Williams, T., Hakuta, K., Haertel, E., et al. (2007). *Similar English learner students, different results: Why do some schools do better? A follow-up analysis, based on a large-scale survey of California elementary schools serving low-income and EL students*. Mountain View, CA: EdSource.