Macrothink Institute™

International Journal of Learning & Development
ISSN 2164-4063
2012, Vol. 2, No. 1

Do High-Stakes Test Really Address English Language Learners' Learning Needs? – A Discussion of Issues, Concerns, and Implications

Jinyan Huang (Ph.D.)

Associate Professor of TESOL and Assessment, Niagara University

E-mail: jhuang@niagara.edu

Turgay Han

English Lecturer, Kafkas University

Kelli Schnapp

BA in TESOL Graduate, Niagara University

Accepted: February 3, 2012 Published: March 3, 2012

Doi:10.5296/ijld.v2i1.1472 URL: http://dx.doi.org/10.5296/ijld.v2i1.1472

Abstract

The No Child Left Behind Act (NCLB) has emphasized the role of high-stakes tests (HSTs) designed to increase accountability for schools and improve student achievement. Under NCLB, English language learners (ELLs) must be included in such tests. Educators must then make critical decisions concerning how to include ELLs in such HSTs in ways that the tests are fair and also address their learning needs. Factors to consider include the selection of appropriate testing accommodations and the accurate interpretation of test results. This paper discusses key issues and major concerns about ELLs and HSTs. It also proposes solutions to both expected and unexpected problems.

Introduction

The number of English language learners (ELLs) has more than doubled since 1980s and has recently grown significantly at American schools (U.S. Department of Education, 2008). Research in both second language education and English for Academic Purposes (EAP) has begun to show that these ELLs' insufficient English language proficiency, coupled with

their other learning challenges has prevented them from achieving satisfactory learning outcomes (Coltrane, 2002; DiCerbo, 2000; Huang, Clarke, Milczarski, & Raby , 2011; Huang, Cunningham, & Finn , 2010; Huang, Smith, & Smith, 2011; Solorzano, 2008).

Further, since the No Child Left Behind Act (NCLB) of 2001 was proposed and signed by President Bush, there has been a key issue involving the inclusion of ELLs in high-stakes tests (HSTs) (Coltrane, 2002; DiCerbo, 2000). The NCLB requires that all students from Grades three through eight must be tested yearly in reading and math. They will have to reach their grade level by 2014. Schools must pass these standardized tests; otherwise corrective actions must be taken, which may include school-wide restructuring or transferring students to another school (Coltrane, 2002). If a school continues to fail these HSTs, eventually it will be closed. It seems that the NCLB is punitive to schools rather than assisting them in improvement. For example, it has caused nearly 40 percent of the nation's schools to be labeled "failing," and by 2014, over 90 percent of the schools will be declared to be "failing" (Campbell, 2009).

There have been on-going debates since the implementation of the NCLB. Criticism of the NCLB emphasizes that ELLs should not have to undergo the same HSTs as regular English speaking students. However, promoters of the NCLB argue that anything less than 100 percent will hurt the children in the long run. In order to succeed, the "poor and minority" students need to be on the same page as everyone else (Paley, 2007).

For example, Berger (2006) described a situation where there was an ELL who had older siblings who spoke English clearly and another student who could not speak English well and whose parents spoke their native language and could not help with their school work. Berger (2006) argued whether or not either student should be included in HSTs. The debate was about whether or not the first child should be tested at the same level as the second, or if they both should be exempt from test taking because they are from other countries. Berger (2006) argued that immigrant children should not be tested because it brings down the schools improvement scores. This topic is highly debated throughout the nation with different scenarios. The legislation needs to decide when the child has enough knowledge of the English language to be tested at the same level as their NE speaking peers.

Obviously, these debates raise issues and concerns about ELLs and HSTs. Do HSTs really address ELLs' learning needs? Therefore, it is important to discuss these issues and concerns, and propose solutions to the identified problems.

Issues and Concerns

Many issues are raised concerning ELLs taking these HSTs. HSTs are actually assessments that test the students, teachers, and administrators. These tests hold everyone accountable for the students' performance. They may be used to determine promotion to the next grade level or whether or not they will graduate. HSTs are meant to raise standards for student learning. ELLs may be challenged to meet higher levels of academic achievement (Coltrane, 2002). Problems and concerns about the reliability, validity, and fairness of assessing ELLs arise because the high-stakes standardized tests that most states currently

employ were developed for the assessment of Native English (NE) speakers, but not for the ELLs (DiCerbo, 2000).

In educational assessments, as argued by Huang (2008, 2009, 2011, 2012), reliability, validity, and fairness are the three major indicators of quality. A high-quality assessment, therefore, should be reliable, valid, and fair (AERA, APA, NCME, 1999; Popham, 2008, 2011).

Educational assessments are consistent or reliable when they produce results that would remain constant on repeated trials (NCATE, 2002). The *Standards of Educational and Psychological Testing* (AERA, APA, NCME, 1999) also uses the word "consistency" to define reliability, indicating that reliability is the consistency of a test "when the testing procedure is repeated on a population of individuals or groups" (p. 25).

Assessments are accurate when they measure what they purport to measure (NCATE, 2002). Accuracy is closely related to the statistical term "validity." The *Standards of Educational and Psychological Testing* (AERA, APA, NCME, 1999) indicates that validity is "the most fundamental consideration in developing and evaluating tests… Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (pp. 8-9).

Assessments need to be reliable and valid in order to be fair (Huang, 2008, 2009, 2011, 2012). Fairness has been the priority in educational assessments during the past few decades (Cole & Zieky, 2001). Educational organizations, institutions, and individual professionals should make assessments as fair as possible for test takers of different races, genders, and ethnic backgrounds (AERA, APA, NCME, 1999).

In educational assessments, the term "fairness" has a broad meaning. It is defined and can also be used in many different ways. As described in the book of *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) and explained by Huang (under review), the term "fairness" has the following four principal definitions. First, it is interpreted as the absence of "bias", which means that test scores earned by students of different identifiable subgroups should not have different meanings. Assessment bias, as defined by Popham (2008, 2011), refers to "qualities of an assessment instrument that offend or unfairly penalize a group of students because of students' gender, race, ethnicity, socioeconomic status, religion, or other such group defining characteristics" (p. 73).

Second, fairness is interpreted that all examinees should be treated fairly during the testing process itself (AERA, APA, & NCME, 1999). Fair treatment of all examinees requires considerations of the assessment context and purpose, how the assessment results are used, and how to assure that all examinees have comparable opportunities to demonstrate their skills and abilities that are measured (AERA, APA, & NCME, 1999). For example, all examinees should be given appropriate testing conditions and equal opportunity to prepare for a test. Further, fairness requires that the marking of all examinees' work should be accurate and consistent, and the reporting of their results accurate and fully informative.

Third, fairness is interpreted that the outcomes of examinee subgroups defined by race, ethnicity, gender, disability, or other characteristics should be equal. However, the idea that fairness requires equal overall passing rates across examinee subgroups is not generally accepted in the professional testing literature (AERA, APA, & NCME, 1999). The more

commonly-accepted view is that "examinees of equal standing with respect to the construct that the test is intended to measure should on average earn the same tests score, irrespective of group membership" (p. 74). Since examinees' levels of the construct are not measured perfectly, this requirement is rarely amendable to direct assessment (AERA, APA, & NCME, 1999). It seems that unequal outcomes across examinee subgroups have no direct relationship with fairness. However, these outcome differences can be further investigated for a testing alternative that minimizes unequal outcomes across examinee subgroups (AERA, APA, & NCME, 1999).

Finally, fairness is interpreted that each examinee has had an equal opportunity to learn (AERA, APA, & NCME, 1999). It is believed that low scores obtained by examinees may have resulted in part from their lacking of the opportunity to learn; and adequate opportunity to learn is clearly relevant to some uses and interpretations of an assessment, although opportunity to learn generally plays no role in determining whether a test is fair or not (AERA, APA, & NCME, 1999).

Further, Huang (under review) indicates that guidelines for fair large-scale standardized assessment practices have been developed and implemented in the United States. The *Code for Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2004) is a guide for professionals who provide and use tests that are fair to all examinees regardless of "age, gender, disability, race, ethnicity, national origin, religion, sexual orientation, linguistic background, or other personal characteristics" (p. 2). It is directed primarily at professionally developed tests (e.g., large-scale standardized tests such state tests) used in formally administered testing programs. It discusses the roles of test developers (i.e., people and organizations that construct tests and those that set policies for testing programs) and test users (i.e., people and agencies that select and administer tests, commission test development services, or make decisions on the basis of test scores).

The *Code* provides guidance separately for test developers and test users in four critical areas: a) developing and selecting appropriate tests; b) administering and scoring tests; c) reporting and interpreting test results; and d) informing test takers (Joint Committee on Testing Practices, 2004). The *Code* is intended to be consistent with the previously mentioned book, *the Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Research has examined the reliability, validity, and fairness issues of assessing ELLs (Abedi, 2002; DiCerbo, 2000; Huang, 2008, 2011, 2012; Huang & Foote, 2010; Rivera & Vincent, 1997; Wolf, Farnsworth, & Herman, 2008). Including ELLs in the HSTs, as argued by DiCerbo (2000) assessment raises many questions about whether or not the assessment tools are valid, reliable, and appropriate for the assessment of ELLs. For example, "when accommodations are permitted, is the test still valid for the intended purpose? Does the test accurately measure the test takers' knowledge in the content area being tested? Does the performance by ELLs with accommodations compare equally to the performance by native-English speaking test takers?" (p. 3).

Using existing data from several locations across the U.S., Abedi (2002) examined the impact of students' language background on the outcome of achievement tests. The results of the analyses indicated that students' assessment results might be confounded by their language background variables. ELLs generally perform lower than non-ELL students on reading,

science, and math. Moreover, the level of impact of language proficiency on assessment of ELLs is greater in the content areas with higher language demand. For example, analyses showed that ELL and non-ELL students had the greatest performance differences in the language-related subscales of tests in areas such as reading. The "gap between the performance of ELL and non-ELL students was smaller in science and virtually nonexistent in the math computation subscale, where language presumably has the least impact on item comprehension" (p. 231). The results also indicated that test item responses by ELLs generated low reliability. Further, the correlation between HSTs scores and external criterion measures was significantly larger for the non-ELL students than for the ELL students. These results suggest that "language factors may be a source of construct-irrelevant variance in standardized achievement tests (Messick, 1994) and may affect their construct validity" (p. 232).

If the construct validity of an assessment is a major concern, its fairness becomes a question (Huang, 2012). One may raise the following two questions easily regarding these HSTs taken by the ELLs: a) what is actually being assessed by these HSTs? and b) are they measuring ELLs' academic knowledge and skills, or primarily just their language skills? When ELLs take these HSTs, the results tend to reflect their English language proficiency and may not accurately assess their content knowledge or skills; therefore weakening the validity of these tests. If ELLs are not able to demonstrate their knowledge due to their linguistic difficulty, the test results will not be a valid reflection of what the students know and can do (Coltrane, 2002).

Wolf et al. (2008) further discussed these validity issues. The NCLB Act has made a great impact on states' policies in assessing the ELLs. The legislation requires states to develop or adopt sound assessments to validly measure the ELLs' English language proficiency. However, due to the lack of available resources states face considerable challenges in validating their current assessment and accountability systems for the ELLs. Considering the significant role of assessments in guiding decisions about organizations and individuals, as argued by Wolf et al. (2008), "it is of paramount importance to establish a valid assessment system" (p. 80).

Furthermore, the administration of the HSTs might cause concerns about the appropriateness and validity of the assessment of ELLs. For example, these HSTs are administered in English, which places ELLs at a disadvantage and raises questions as to how the tests should be interpreted. Students being tested in their native language may be more appropriate since many ELLs are enrolled in a bilingual education classroom. In these classrooms, students are instructed in their native language for some content-area instruction. These students may demonstrate their subject-area knowledge more effectively in their native language. Tests in other languages are rarely provided and even having test accommodations in their native language is prohibited. Test items may contain references to ideas or events that are unfamiliar to ELLs because they have not been exposed to similar concepts in their native culture and have not lived in the United States for a long period of time (Coltrane, 2002).

In the past few years, Huang (2008, 2011, 2012) has been investigating the fairness concerns about the assessment of ELLs in large-scale standardized tests. By examining both the reliability and validity of assessing ELLs' writing in large-scale assessments, Huang (2008,

2011, 2012) has continuously provided empirical evidence for the inequitable assessment practices of the ELLs.

The research in the area of ELLs and HSTs has provided important implications for research and practices (Coltrane, 2002). For example, how should those professionals responsible for selecting the HSTs carefully examine how closely a test reflects the curriculum and standards being used in their state or district? How should teachers of ELLs be involved in the decision-making process regarding which tests are to be used and which accommodations and modifications should be selected? Further, how should HSTs data be interpreted and used for decision-making?

Implications for Research and Practices

Researchers in this area need to understand the contexts of large-scale assessments that are intended to hold schools accountable for what students know and can do on the basis of their performance on assessments. In order to find new directions for linking assessment and schooling practices furthering the education of the ELLs, researchers need to ask themselves the following questions: Who are the ELLs? What are their expectations? How do HSTs affect school achievement? And how should HSTs address ELLs' learning needs (Dura'n, 2008)?

Assessment professionals need to take into account more than the ELLs' scores on these HSTs. For example, they need to aware of the fact that the ELLs are not on the same page as their NE speaking classmates. Many factors affect their performance on the HSTs including the number of years they have lived in the United States and the curriculum they are learning at schools. The ELLs are not taught the same information as their NE peers, which could have a major impact on their HSTs scores.

Similarly, teachers need to understand the learning challenges faced by the ELLs. For example, the ELLs are on various sides of the English language spectrum. A teacher could have a student who was born in the United States, but has parents who speak their native language at home. The teacher could also have a student who just came to the United States a month ago and knows no English. This teacher is expected to teach all sides of the spectrum and get them all on the same page to take the HSTs after being in the United States for at least three years. Further, this teacher is expected to teach the discourse of these HSTs and test-taking skills. To raise their awareness and familiarize them with the formats of the high-stakes standardized tests (e.g., language and patterns of the test and useful test-taking skills) definitely help improve their performance on these HSTs.

In terms of assessment practices, it is suggested that test modifications, special accommodations, and alternative assessments be implemented for the ELLs. Test modifications, for example, can refer to modifications to the test itself as well as modifications to the test procedure.

Further, accommodations can be classified into four types: a) presentation, which allows for repetition, explanation from test administrators, the translation of texts into the students' native language, or an ESL/bilingual specialist as an administrator; b) response, which allows for an oral response from the student or for the student to respond in their native language; c) setting, which allows students to be administered the test in a group; and d) timing/scheduling, which allows students to have additional time or extra breaks. Setting and

timing/scheduling do not specifically address the ELLs' linguistic needs. However, presentation and response do address the linguistic needs of the ELLs (DiCerbo, 2000).

In addition, alternative assessments in the classroom can be implemented as a supplementary to HSTs in the assessment of ELLs. Alternative assessments refer to procedures and techniques which can be used within the context of instruction and can be easily incorporated into the daily activities of the school or classroom (Tannanbaum, 1996). Alternative assessments are beneficial to ELLs because they are then evaluated on what they integrate and produce rather than on what they are able to recall and reproduce. Alternative assessments focus on documenting individual student growth over time, rather than comparing students with one another. Further, it emphasizes students' strengths (what they know), rather than weaknesses (what they don't know). Consideration is given to the learning styles, language proficiencies, cultural and educational backgrounds, and grade levels of students (Tannenbaum, 1996).

Tannanbaum (1996) provided the following suggestions for alternative assessments. First, use non-verbal assessment strategies such as physical demonstrations to assess the ELLs. Physical demonstrations can express academic concepts without speech (using gestures). Students can perform hands-on-tasks or act out concepts (e.g., thumbs up or down for true/false statements). Pictorial products are also a non-verbal assessment, which is where the teachers can have students produce or manipulate drawings, dioramas, models, graphs, and charts (e.g., labeling maps).

Second, teachers can have their students perform oral presentations which would include interviews, oral reports, role plays, describing, explaining, summarizing, retelling, paraphrasing stories or text materials. Oral assessments should be conducted on an ongoing basis in order to monitor comprehension and thinking skills. In conducting interviews with ELLs with early stages of language development, teachers should use visual cues often and allow for a minimal amount of English in the response. Role plays can also be used as a presentation.

Third, oral and written products include content area logs, which encourage the use of metacognitive strategies when students read expository test (e.g., "What I understood/What I didn't understand). Reading response logs are used for students' written responses or reactions to a text. They may respond to questions that encourage critical thinking. Dialogue Journals provide a means of interactive, ongoing correspondence between students and teachers. Students determine the choice of topics. Beginners in the language can draw pictures. Audio and video cassettes can be made of student oral readings, presentations, dramatics, interviews, or conferences (with teacher or peers).

Finally, portfolios can be used to collect samples of student work over time to track the students' development. The materials to be included are audio and video recordings, writing samples, art work, conference or interview notes, checklists, and tests and quizzes. It is important for teachers to include more than one type of material in the portfolio in order to gain multiple perspectives on students' academic development.

References

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

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

Berger, J. (2006). Immigrant children shielded from state tests, but for whose protection? *The Washington Post*, 27 December 2006: B.7.

Campbell, D. (2009). Children left behind. *The State Hornet* 11 March 2009. Retrieved from http://www.statehornet.com.

Cole, N. S., & Zieky, M. J. (2001). The new faces of fairness. *Journal of Educational Measurement, 38*, 369-382.

Coltrane, B. (2002). English language learners and high-stakes tests: An overview of the issues. ERIC Educational Resources Information Center. ERIC Clearinghouse on Languages and Linguistics. 15 April 2009. Retrieved from http://www.eric.ed.gov.

DiCerbo, P. A. (2000). What are the critical issues in wide-scale assessment of English language learners? ERIC Educational Resources Information Center. National Clearinghouse for Bilingual Education. The George Washington University Center for the Study of Language and Education. 15 April 2009. Retrieved from http://www.eric.ed.gov.

Dura'n, R. P. (2008). Assessing English language learners' achievement. *Review of Educational Research, 32*, 292-327.

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing. *Assessing Writing.*

Huang, J. (under review). Reliability, validity and fairness in educational assessments. *International Journal of Applied Educational Studies.*

Huang, J., Smith, A., & Smith M. (2011). Teacher perceptions of strategies for improving ESOL students' academic English skills: A K-12 perspective. *The Canadian and International Education Journal, 40*(3), 61-80.

Huang, J. (2011). Generalizability theory as evidence of concerns about fairness in large-scale ESL writing assessments. *TESOL Journal*, 2(4), 423-443.

Huang, J., Clarke, K., Milczarski, E., & Raby, C. (2011). The assessment of ESOL students with learning disabilities: Issues, concerns, and implications. *Education, 131*(4), 732-739.

Huang, J., Foote, C. (2010). Grading between the lines: What really impacts professors' holistic evaluation of ESL graduate student writing. *Language Assessment Quarterly: An International Journal, 7*(3), 219-233.

Huang, J., Cunningham, J., & Finn, A. (2010). Teacher perceptions of ESOL students' greatest challenges in academic English skills: A K-12 perspective. *International Journal of Applied Educational Studies, 8*(1), 68-80.

Huang, J. (2009). Factors affecting the assessment of ESL students' writing. *International*

Journal of Applied Educational Studies, 5(1), 1-17.

Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments? – A generalizability theory approach. *Assessing Writing, 13(3),* 201-218.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC: American Psychological Association.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

National Council for Accreditation of Teacher Education (2002). *Professional Standards for the Accreditation of Schools, Colleges, and Departments of Education.* Washington, D.C.: Author. Retrieved from http://www.ncate.org/2000/unit_stnds_2002.pdf.

Paley, A. (2007). No child's target is called out of reach: Goal of 100% proficiency debated as congress weighs renewal. *The Washington Post, 14* March 2007: A.1. 30 March 2009 Retrieved from www.thewashingtonpost.com.

Popham, W. J. (2008). *Classroom assessment: What teachers need to know* (5th Ed.). Pearson Education, Inc.

Popham, W. J. (2011). *Classroom assessment: What teachers need to know* (6th Ed.). Pearson Education, Inc.

Rivera, C., & Vincent, C. (1997). High school graduation testing: Policies and practices in the assessment of English language learners. *Educational Assessment, 4*(4), 335-355.

Solorzano, R. W. (2008). High states testing: Issues, implications, and remedies for English language learners. *Review of Educational Research, 78*, 260-379.

Tannanbaum, J. E. (1996). Practical ideas on alternative assessment for ESL students. ERIC Educational Resource Information Center. 15 April 2009. Retrieved from http://www.eric.ed.gov.

U.S. Department of Education. (2008). Retrieved June 26, 2008 from the World Wide Web: http://www.ed.org.

Wolf, Farnsworth, & Herman. (2008). Validity issues in assessing English language learners' language proficiency. *Educational Assessment, 13*(2), 80-107.

About the Authors:

Jinyan Huang (Ph.D.) is an associate professor and Ph.D. faculty in the College of Education at
Niagara University. His research centers on the following areas: a) ESOL students' learning challenges and coping strategies; b) Factors that affect ESOL students' learning outcomes; c) Reliability, validity, and fairness issues of ESOL assessments; and d) The use of assessment data for supporting ESOL leadership and policies.

Turgay Han is an English lecturer at the Faculty of Letters of Kafkas University, Turkey. He is also a Ph.D. candidate at the Faculty of Letters of Atatürk University, Turkey. His areas of scholarship include assessing language skills, examining score variability and reliability of ESOL writing, and rater training.

Kelli Schnapp is a recent graduate in TESOL from Niagara University. Previously she worked on projects dealing with ESOL learning and assessment issues.