Language Teaching Research Quarterly

2019, Vol. 11, 31–42

A Review of Standardizing an English Second Language Test-Item through Reverse Engineering

Zahra Foghahaee

Department of Humanities, Islamic Azad University Ayatollah Amoli International Branch (AAIB), Iran

Received *18 November 2018* ***Accepted*** *15 July 2019*

Abstract

Reverse engineering (RE) can play an important role in the re-designing tests in L2 English. It can also enrich the aim of teaching the same as raising children through academic achievement. In addition, it can play a key role in helping students understand how much their test is valid by using Standard reverse engineering (SRE). This paper is a literature review seeks to make a general second language test standardized through SRE by Popham's Model. According to this model, a test becomes standardized by considering the constructs of a standardized test through five steps respectively. This model evaluates participants' English test as foreign language proficiency. In order to explain it more, one example of writing the test in the site of the State Education Department Office of Bilingual Education in New York is explained in detail. This study can be used as a guide to other second language testers who wish to standardize their tests.

Keywords: *Reverse Engineering (RE), Standard Reverse Engineering (SRE), Popham's Model, English Second Language Testing*

Introduction

This paper is essentially a literature review with the purpose of exploring the constructs of a standardized test to assess participants' English test as foreign language proficiency. To fulfill this aim different processes are used and said in the following to design a standard test, respectively:

Candidates' communicative competence is checked by using the test that is designed not only on the base of psychometric conditions, but also on the base of its validity, reliability, and

phdazaduniversity@yahoo.com

construct. The reason of why it is said that the test must be valid is if the test is more valid for its intended purpose, it will be able to accept some changes in its content and represent the construct of language assessment literacy better. On the other hand, the test with a high degree of validity has a high assessment, too. In order to achieve this goal, the reverse engineering RE process is done to secure not only to make the test valid but also to standardize it (Firoozi & Razipour, 2014).

Firoozi and Razipour (2014) believe that because those institutions with the responsibility of designing and administering a test are neither available nor clear what the base of a test specification is, RE is adopted to make obvious the content structure of the test. According to Walters (2010), RE also lets the team of test developments clarify what they are trying to measure as a correct part of producing a feasible spec or test specification. Despite critiquing existing specs and tests, the number of studies on RE in language testing and teaching is few; while the learners' tendency for learning the standard form of English as a second language (ESL) increases more than before. That is why this is considered as a cause to improve English language proficiency. The reverse process of writing skill in language proficiency is used to make a standard test. It means, this process works on a mandate to evaluate it and tries to use previous methods for assessing the test rather than to use dimensions of assessment-literacy education in the context of the second language (L2) testing. As a result, the aim of this study is to increase the second language learners' proficiency by using of Popham's Model to make the test standardized through RE.

Assessment

According to Firoozi and Razaipour, assessment is not made automatically through the teaching process, because it needs experts to give feedbacks and constant professional developments. So, in order to do that, it is necessary to be aware of principles and procedures. Doing assessment in this way is known as Language assessment literacy (LAL). In general, education assessment and language assessment literacy LAL is done in the field of language testing Inbar-Lourie (2008). They also add to make the educational process better when it is possible to assess test takers' communicative competence; because according to Balk and Wiliam (2012), one of the most important ways in which learning and teaching activities can be defined or changed to modify learners' needs better is to use assessment, for example, through administration the tasks in the classroom in groups or pairs, not only do the students have time enough to create opportunities for production and learning, but also they can become aware of their problems through communications and try to solve them for their other negotiations that they have in the future (Fulcher, 2012), their own assessment is useful. According to Popham (2009), without the effect of a good assessment in a test, taking place of a good education will be impossible, because a good assessment can make specific changes in bad teaching. Even Firoozi and Razaipour suggest that the evidence indicates that implementing good assessments make positive changes in bad teaching. In comparison with other areas of language education, there is an indirect relationship between the literature on LAL and its vital importance in language education and assessment.

Validity

Birjandi and Seyyedrezaei (2015) believe that validity is one of the vital features of a good test, like IELTS and is considered as a center for many studies. In addition, in the field of testing they have different definitions for test validity are (a) validity is defined on the base of two different models separately: traditional and model views; but constructing most of written English language tests are done on the base of traditional ones; (b) Hughes (1990) claims that in testing and assessment, validity is traditionally understood to show whether a test measures accurately what it is intended to measure; and (c) Henning (1987) says test validation measures how much the item is appropriate to measure what is supposed to measure. So, Fulcher believes that the concept of validity is to lead our work in testing and assessment that among many phrases it can be considered as the central part of testing. On the other hand, the aim of validity is not only to create the tests that have a strong relationship between their inferences and decisions but also it has a positive influence on people and institutions.

Construct Validity

According to Messick (1989), scoring is considered an aspect of construct validity and used to measure the results of attributing three major categories of validity to show how these three dimensions interact with each other and the unitary nature of validity. Weir (2005) defines these three categories in three dimensions in detail: social, context, and cognitive processing. Weir adds construct validity does not only reflect the underlying traits of communicative language ability, but also it is the result of trait, context, and score. On the other words (Heaton, 1990), those underlying traits of communicative language ability belong to theories of language behavior and learning, the theories of certain learning, and the theories of improving acquisition and skill are measured by construct validity. For example, a speed test of a short passage of reading comprehension doesn't have a good measurement of reading ability and thus has a low construct validity; but a reading comprehension test on the base of the communicative approach is not matched to the course syllabus, although it is made up of multiple-choice items, it won't have construct validity.

Reverse Engineering

Davidson and Lynch (2002) were the first experts who used the term RE which is made up of a test-item specification from a test that exists. According to Walters (2010), the whole process of test creation is (a) to analyze one item or a set of test items or tasks; and (b) to make a test specification or a set of test items or tasks. In addition, RE is a new concept of language testing and assessment. Davidson and Lynch (2002) believe that it has a process of analytical mechanism to create a test which considers actual questions in a test, and infers the guiding language that brings the process ahead to reconstruct the components of an actual test specification. But, Davidson and Lynch (2002) defines RE as it is used to produce test specification when there is no existence of it. From the test items or tasks, we conclude what the

blueprint might be. According to Firoozi and Razaipour, because it is unclear whether the LAL test is designed on the base of test specification, the RE approach is used to uncover the content structure of a test. An RE approach is a mechanism via which the components of a test specification on the basis of the actual test are constructed. Only those constituents of test spec are analyzed that mainly inform the constructs of the item and task writing (Davidson & Lynch, 2002). In order to be more familiar with RE, different types of it are explained below.

Straight RE

In the process of producing equivalent test questions, guiding language will be inferred (Fulcher & Davidson, 2007) if the existing items are not changed (Khorambin, 2016).

Historical RE

Fulcher and Davidson believe that the reason of the process of historical RE occurs among several existing versions and archives of a test if, on the basis of Khorambin's point of view, the existence of different versions of items are confirmed and through using of historical RE it can be understood how and why the tests change.

Critical RE

Fulcher and Davidson define critical RE as it is able to think, ask about, and analyze each item in the test independently to understand whether the following questions match the definition of critical RE (a) are the items going to be tested; and (b) are the items used to make changes in the test design and also on the base of Davidson and Lynch's point of view (2002) are able to measure the thing that we want to measure it? In addition, Birjandi and Seyyedrezaei believe that critical RE is used for constructing the validity of General Writing Test Task (GWTT), for example, IELTS. This means that it is not important how we take a look at this process, whether we are looking at it critically, straightly or historically.

Parallel RE

On the base of definitions on external influences of the mandate, the meaning of parallel RE is explained by Fulcher and Davidson is considered as a tool to determine parallelism when teachers are asked to produce tests. Khorambin adds that it is possible to have a set of external standards outside of the classroom, for example The Common European Framework. Maybe, the teachers become encouraged to design such a test on the base of external standards at the same time without consulting with other teachers to measure the same thing and showing straight RE on the samples and then compare the resulting specs. So, as the meaning of parallel RE shows separate sets of test specifications are compared to each other to see the partial degree of two sets of test specifications measure the same skill.

Test Deconstruction RE

Elatia (2003) analyzed the history of a major national language test, she defined the term "test deconstruction" for the first time. This type of RE takes a look beyond the test setting to discover larger realities (Khorambin, 2016), for example, (a) why our particular test setting is very comfortable to the students and causes them to continue the test?; (b) what the role of close inferential reading is to the school setting, and (c) Do educators use RE deconstruction to produce difficult items and spread out the students' ability to have a bell-shaped curve? As a result, applied deconstruction RE, examining the philosophical, ideological, and governmental-policy assumptions underlie the Baccalaureat Exam in France (Walters, 2010, p. 321).

History of Writing

Ellis (1994) claims that writing is defined as transcribed speech and viewed as decontextualized one. In addition, Hyland (2002) defines writing is a text in which independent issues and various elements are gradually organized. The reason of why it is said that writing is seen as a communicative and social activities is because in a written text the writer tries to convince the reader by using of linguistic patterns with social constraints and choices, like the writer's goal, knowledge, and etc. As a result, Birjandi and Seyyedrezaei define a written text is viewed as a discourse. Hence, writing is seen as a social act because each writing explains contextual factors.

The process of this research does not focus only on the stages of the writing process, but also it works on writing and how these stages of writing process relate to different levels of language proficiency. According to Scardamalia and Bereiter; Eysenck and Keane (2005), the stages of the writing process are planned processes in which different acts are done from skilled to unskilled writing. In this way, there are two major strategies for the writing process: knowledge telling and knowledge transforming which are explained completely through an example in the current study. These two strategies occur mainly at the planning stage and identify more skilled writers from less skilled ones. In the former, the writer's role in planning the writing process is a little and most parts of planning of the writing are done mainly on the base of generating content from those resources that the writer remembers them. Those resources are on the base of content, task, and genre. In the latter, the writer considers both the complexity of tasks, content, audience, register, and other relevant factors in written communication.

The Aim of Standardizing the Test

In order to enhance improvement in instruction as well as raising children in academic achievement, Ravitch (2004) describes various aims of standards not only are used to standardize the test content or curriculum but also they are used to standardize the students' performance through administering an achievement test. According to Ravitch (2004), the former relates to what the content a student should know, what skills he is able to display, and the latter relates to different levels of mastery of content and skills.

The History of Standardizing the Test

Different movements are done in general education influence learners on the field of teaching English language learners (ELLs). Even intergovernmental institutions, like the New York State Education Department Office of Bilingual Education (2004) influences the field of teaching English to language learners ELLs and creates standards for ELLs. The advantages of these standard forms of English language (a) have a positive effect on the instruction to ELLs; (b) highlight the special needs of ELLs; and (c) make educators responsible to improve it Goertz, Duffy, and Menken (2001, 2008). Unfortunately, using the standard forms of English language is criticized because of, for example, the fear of government imposition of education and moral values on educational systems and cultures which are considered as the root of local communication Ravitch (2004). Kraft (2001) believes that standard forms of the English Language don't improve the level of student achievement or instruction but promote school status. Other experts such as Valencia, Valenzuela, Sloan, and Foley (2001) add this function becomes more limit when they judge these types of standard forms are used for expanding educational and economic barriers to minorities; while in fact it is emphasized a lot if standard of implementations do not support their effects on educational resources to fulfill the English learners specific needs they will be penalized.

Tools for Assessing/Standardizing Mandate

Davidson and Lynch (2002) define mandate as a combination of social forces which decides what will be tested and what the content of the test is. According to Fulcher (2012), the term '*mandate*' is also used to describe where the test purpose comes from and can be seen in either internal or external institutions. The former is used a lot by teachers and school administrations. The latter which is used in this study considered by a group of people who don't have enough knowledge about the ecology of local learning.

On the base of Walter's point of view, although defining the domain of a test standardizing is difficult, it is not clear to say a single standard may have the content elements from more than one content area. Nevertheless, through comparing those tests to standard ones, Davidson and Lynch divide the items of these exams into four different levels: high, moderate, low, and no alignment; while Resnick, Rothman, Slattery, and Vranek (2002), to define different dimensions of a standard test are (a) to compare the test content with the content of standard test; (b) to match the cognitive response that is given on the base of identity, selection, or analysis to the response of the relevant standard; and (c) torelate the issue (thesubject matter mastery) to standard(s) which must be found among the items of the test.

How to define Item Specification

According to Davidson and Lynch (2002), a test item specification (or test spec) is a list that test-item writers use it to produce parallel test tasks effectively. Walter adds because in this way the process of designing test-item specification is a hand-craft, not only are teachers and item designers in their best way are always able to distinguish the goals of a specific test, but also they

are able to use a specific method for designing different test tasks and goals of teaching. So, this task differentiates from designing the test randomly. Through random task process, it is possible that the items are designed with either the same or different skills; but through designing the test spec by hand, the expert of designing a test on the base of a list of test-item specifications chooses those items that are related to that field of study and puts them beside each other to make a test. In order to make stronger the role of test specification through test creation and its administration, Davidson and Lynch (2002) suggest that the effective aim of a test specification is to make consistency and fairness among test items and so the teacher's role becomes colorless in administering the test. There are different models to define the process of test specification, like: Popham's (1978) Assumption for the model of the test-item specification is defined by Davidson and Lynch in this study. His supposition has five main components:

General Description Section (GD)
The students give a summary in the form of a paragraph. It is used for the same text which has different outlines.

Prompt Attribute Section (PA)
It talks about those stimuli that the candidate sees or hears.

Response Attribute Section (RA)
It explains what the candidate does to respond to those stimuli that he sees or hears.

A Sample Item Section (SI)
First, one item as a sample of test tasks is shown to writers and teachers. The details of this example are explained exactly in three different sections and are told them how candidates must answer those items.

Specification Supplement Section (SS)
It is related to the information that the teacher or item writer uses, but this information is not easily classified like other components of Popham's Assumption.

The aim of RE is to make similar items as a test and the aim of standard reverse engineering (SRE) is to compare the result of the RE process which is done in the test to make it standardized and checks how much it becomes close to its standard published version. This comparison is done in order to be sure how well the students could take the test. In order to discuss it more, one example that occurs in the State University of New York is explained in detail below. Popham's Model is used to make the test standardized.

The participants are English teachers who teach the English Language as a second language to foreigners. On the base of their extracted experiences in teaching from demographic questionnaires, they are divided into three groups: preservice, new in-service and old in-service. As the names of these groups show, their experiences in teaching increase from the first group to

the third one. It means, the first group of teachers has experience in ESL teaching in classrooms for two semesters. It is clear that they are dependent on participants and can't be controlled strictly. The mean of the second group of teacher's ESL teaching is 0.5-1.5 years, and the amount of the third group of teachers in teaching ESL in classes is about 12.5 years. That is why there are three small workshop groups with different background characteristics. Each of these workshops are made up of four subparts. The process is done on the first and the second group while the third group who is the most experienced in English writing texts revises their notes. It is supposed in the first part of all of these workshops that Popham's Model rather than Davidson and Lynch's Model is used because it is supposed that it is easier than their model. The reason of why Popham's Model is easier than Davidson and Lynch's is it explains all the details through the writing process as the same as a tutorial class. His Model makes the students able through reading comprehension to notice to those details and sequences are used in the text naturally and to guess what the whole passage is by taking a look to one of the test-items without taking a look to the text. In the second stage of this Model, the students' language proficiency as their second language proficiency is measured in four skills (a) listening, (b) reading, (c) writing, and (d) speaking. This measurement is compared with the standard form of language proficiency on the site of the New York State Education Department (2006). In order to check the degree of students' understanding of the text, two multiple choice items with one stem and Distracters are used. Each of these two items is answered on the base of the sample on that site. At the beginning of the workshop, one item is asked from the teachers of Grades 2 to 4 but to the rest of the workshop is asked from those teachers who have the Grades of 9 to 12. The domain of measurement of the first item is asked belonged to check how much the candidate's total comprehension degree of his/her own narration text is. Narration is about walking in woods up to the time raining interrupted their walking there. Because the first item is asked from those teachers who have the Degrees of 2 to 4, concentration is put only on present tense the only linguistic features are used by candidates in their narrations are (a) verbs, (b) adjectives, and (c) nouns; but the exceptions of the second item asked from the teachers with Grades 9 to 12 is more, they must use expanded domains of linguistic rules through their narrations. So, the second item is a passage made up of three to four paragraphs in where the complete and the short form of past tense are used. The second item is asked to measure the degree of students' total comprehensions of air condition. From the comparison is made between these three groups of candidates, it becomes clear that experienced teachers' group would better facilitate the SRE process. In the second workshop, unlike the first workshop, the participants (preservice group with Grades of 9 to 12) try to work on skill before considering specification properties. The participants (in-service group) talk about the process of spec components continuously while at the end of the text, the students must complete multiple choice item tests correctly by comprehending the meaning of that item. In the next step, the participants are asked to write a text on the base of their own understanding to show how much they understood the text, and then the quality of students' writing is checked by using of specification of RE through multiple choice test. In continuous, the experts of reverse engineering describe the students' skill that is

going to be tested. At the third step, comparing students' knowledge in language creation with the New York Site occurs through standardizing reverse engineering (SRE) process, through which the experts of RE find those details have influence on SRE process, for example, having the students' tendency to answer one item sooner than the other one will make some changes in his/her own decision making. These changes will be obvious through the interaction process. Through this comparison, it becomes clear how many students' language is similar to standardized language. Theoretically, in each test it is assumed that there are subfields of standard learning indicate the learner's performance clearly. Each of it describes the specific and learning task as evidence of progress toward achieving a learning or standard goal. In the fourth part of the test, the students compare the reverse engineering of their handwriting with the standard version of writing exists on New York Site to see similarities and differences between them.

The framework of published standardized test is (a) to achieve standard information in English needs, like listening, speaking, reading, and writing; (b) to understand how much they could support their inferences, they must take a look to the written text, like vocabularies and facts; (c) to convey information in terms of variety of organizational patterns and structures, for example, to use the knowledge of text structure that facilitates comprehension; and (d) to achieve information from the written text, for example, they use different learning strategies, like visualization by which those connections are made to read for details. In the end, based on Popham's Model, a questionnaire is distributed among participants and their ideas are asked about the process of improving the test into its standardized version by doing any changes if it is needed. The items in the questionnaire are (a) what your idea about using this method for standardizing a test is? Do you accept it or not and why; (b) whether it is necessary to revise your manuscript and why; and (c)whether you could answer the test better by using of SRE and why.

Complete Explanation of SRE Process

In the preservice group, the participants tried to complete the stems while they are trying to use the time that is put under focus in the test-item. It is called GD. Then, the students' understanding from the details of the text through the visualization process is checked. It is called PA. The blank of the stem is related to the target point. Here, they show how much their visualization is confirmed. In the next step, RA, through which the participants (in-service group) considers visualization as a strategy not as a skill because test concentration is put on the time of the beginning of the text. So, in order to get that time, not only students understand the sequential order of events, but also they must be able to revise it through different times. That is why they must use different information to be able to revise it.

Candidate's Answers to the Questions of Questionnaire about SRE

Walters believes that in order to be sure how much SRE will be beneficial for candidates and can improve their knowledge through the correction process, it is possible to ask them to answer the

items of a questionnaire. From their responses to items, it becomes clear that different teachers' groups had different ideas about it. For example, preservice group and in-service group believe that SRE can improve their writing skill, because it can solve their errors by giving them the correct form of structure. Not only can all these groups distinguish some differences and similarities between their performances and their own standards of reverse engineering, but also they can distinguish between relevant standards on New York State Site and the indicators of their performance. In order to explain their comprehension more about what the aim of using SRE is, it is better to expand their opinions into details. Preservice and in-service group believe that New York State standards are broader than their own reverse engineering standards. It means that state standards cover all four skills, while the candidates' own focus is only put on reading.

The other question asks them to evaluate published standards or to determine whether they should be revised. Different groups have different ideas about it. For instance, the new in-service group believes that it isn't necessary to revise it; but the in-service group suggests that it is necessary to revise it. According to the preservice group, it is necessary to revise it; although this group talks about speaking and reading corrections, there isn't said anything about the guidance of writing structures. So, the lack of them won't facilitate writing.

The last question of questionnaire asks candidates of these three groups to give their opinions about whether SRE process helps them to understand the nature of the test items and standards and whether their understanding would help them in their ESL classroom teaching and testing. Preservice group reports that SRE makes you sure that you are testing the skill that you want to test it. Similarly, the new in-service group says SRE can help you to make clear the nature of given test items. So, the speech of these two groups makes clear that SRE gives you the specific type of insight for assessing items.

Through comparison the students' writing with its standard form, it becomes obvious the degree of candidates' knowledge to answer the question. The older in-service group also believes that SRE is helpful in understanding the nature and purpose of test items. By taking a look at groups' answers, it becomes clear that the range of answers and their interpretations are from general into specific details and interpretations of test items. For more explanations, a few numbers of teachers' answers inspected (a) without using SRE are not obvious; (b) through a discussion that a multiple choice test can be created under an hour; while without using it, its creation will be impossible; (c) through using discussion, the range of students' abilities becomes obvious.

Results and Discussion

According to Walters, not only SRE can be used for assessing language skills, but also it can be used for teacher’s classroom-based on assessments. Moreover, the impact of RE on language testing education is very useful in serious investigations because it is a process of testing and analyzing a system or a device to identify, understand, and write its functionality (Bani & Tutunji, 2012); while Firoozi and Razaipour believe that it has not been discovered completely.

Even many experts in this field of the study believe that analyzing the test items on the base of one specific model is ambiguous and will not be done easily. The implications for future research can be, for example, based on Walters's idea about either depending on the test-takers on the study or on the field of the second language, like How practicing with SRE will have an impact on testing of teachers' classroom and teaching practices? It is assumed that the earlier versions of RE are useful in increasing language-teacher or test awareness of assumptions implicitly in given test items. In addition, a similar effect on assessment literacy might obtain with SRE which are extracted from the participants' answers to the items of the questionnaire. As an example, during a test administration, from each individual's asking of their teacher, it becomes clear that (a) how well the test is assessed; (b) what the nature of the test is; and (c) what strategies are used by test-takers to answer the questions.

References

Bani Younis, M. & Tutunji, T. (2012). Reverse engineering course at Philadelphia University in Jordan. *European Journal of Educatneering Education*. *37*(1), 83-95.

Birjandi, P., & Seyyedrezaei, S. H. (2015). An approach to test validation of general writing task two of IELTS through reverse engineering. *International Journal of Basic Science & Applied Research. 4*(sp), 23-38. ISSN: 2147-3749.

Davidson, F., &Lynch, B. K. Testcraft: A teachers' guide to writing and using language test specifications (2002). New Haven, CT: Yale University Press.

Elatia, S. (2003). *History of the Baccalaureat: A study of the interaction between education legislation, government policy, and language theory in the National Language Examination*. Unpublished doctoral dissertation, University of Illions at Urbana-Champaign.

Ellis, R. (1994). The study of second language acquisition. Oxford: Oxford University Press.

Eysenk, M., & Keane, M. (2005). Cognitive psychology (5Th ed.). Hove: Psychology Press.

Firoozi, T. & Razaipour, k. (2014). How the knowledge base in language assessment is measured: A reverse engineering approach. ISSN: 1755-9715.

Fulcher, G. (2012). Practical language testing. First published in Great Britain in 2010: Hodder Education.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: Language testing and assessment: An advanced resourse book. Taylor & Francies.*

Goertz, M., & Duffy, M. (2001). Assessment and accountability systems in 50 states, 1999-2000. *Philadelphia: Consortium for Policy Research in Education*. University of Pennsylvania. Retrieved July, 28, 2009, from http://www. eric.ed.gov/ERICOcs/data/ericocs2sql/content_storage_01/0000019b/80/16/df/13.pdf.

Heaton, J. B. *Writing English language tests* (1990). Salman-e Farsi: Payeh-e danesh (the 1th ed.).

Henning, G. (1987). A guide to language testing. Cambridge. MA: Newbury House.

Hughes, A. (1990). Testing for language teachers. Glascow: Cambridge University Press.

Hyland, K. (2002). Teaching and researching writing. London: Longman.

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base. *Language Testing*, *25*(3), 385-402.

Khorambin, S. (2016). Reverse engineering in language testing, science and promoting teacher quality. *International Journal of Modern Language Teaching and Learning. 1*(5),182-187. ISSN: 2367-9328.

Kraft, N. P. (2001). Standards in teacher education: A critical analysis of NCATE, INTASC, and NBPTSCA conceptual paper/review of the research. Paper presented at American Educational Research Association Conference, Seattle, WA.

Menken, K. (2008). *English Learners left behind: Standardized testing as language policy*. Clevendon, UK: Multilingual Matters.

Messick, S. (1989). Educational measurement. New York: Macmillan.

New York state Education Depatment (2006). NYSESLAT sampler 2006 grades 9-12. Orlando, FL: Harcourt Assessment. Retrieved on 31 July 2009 from http://www.emsc.nysed.gov/osa/nyseslat/sampler06/9-12bk.pdf.

New York State Education Department Office of Bilingual Education (2004). The teaching of language arts to limited English proficient/English language learners: Learning standards forEnglish as a second language. Albany, NY: Author. Retrieved from http://www.emsc.nysed.gov/biling/resource/ESL/standards.html.

Popham, W. J. (1978, 2009). Transformative assessment. Aexandria, VA: ASCD.

Ravitch, D. (2004). The language police: How pressure groups restrict what students learn. New York, NY: Vintage Books.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). Benchmarking and alignment of standards and testing (CSET ech. Rep. No. 566). Los Angeles, CA: Center for the study of evaluation.

Valencia, R. R., Valenzuela, A., Sloan, K., & Foley, D. E. (2001). Let's treat the cause, not the symptoms: Equity and accountability in Texas revisited. *Phi Delta Kappan, 83*, 318-326.

Walters, F. S. (2010). Cultivating assessment literacy: Standards evaluation through language test specification reverse engineering. *Language Assessment Quarterly*. http://dx.doi.org/10.1080/15434303.2010.516042.

Weir C. J. (2005). Language testing and validation. New York: Plgrave Macmillan.

Wiliam, D. (1996). Meaning and consequences in standard setting. Assessment in Education: *Principle, Policy and Practice 3, 3*, 287-308.