ANTICIPATING THE SIDE EFFECTS: A REVIEW OF A REFUGEE ENGLISH PLACEMENT TEST IN THE U.S.

Amy Nicole Soto
Eötvös Loránd University, Budapest
amynsoto4323@gmail.com

Abstract: Critical language testing considers the cultural, social, and political context of tests (Kramsch, 1993; Pennycook, 1994; Shohamy 1998). Existing research examines the role of testing at the national (or international) level in immigration, citizenship, and asylum (e.g. Kunnan, 2009; Saville, 2009; Shohamy & Kanza, 2009). However, as Bachman (2005) criticizes, critical language testing research insufficiently acknowledges the link between consequences and the validity of inferences made from tests. This paper reviews an English placement test used at the local level in the United States by a refugee resettlement agency where test takers are from low literacy and low education backgrounds. Methods include observations, interviews, and collection of testing materials (i.e., test scores and placement decisions). Validity is analyzed through the theories of Messick (1989) and Bachman and Palmer (2010) to explore the discrepancy between functionality and lurking potential washback. Monitoring washback is important because tests can become de facto policies which shape programs (King & Bigelow 2017, Shohamy 2014). Findings indicate that validity of the literacy and oral proficiency constructs must be improved to mitigate possible negative consequences, which stem from defining self-sufficiency and the prerequisite language proficiency for it (Shohamy 2013).

Keywords: construct validity, consequential validity, placement test, refugee, test review

1 Introduction

In recent years, a new test taker population has arisen from global human movement which existing research and practices in testing do not fully address. According to the United Nations High Council for Refugees (UNHCR) *Global Trends: Forced Displacement in 2017* report, 68.5 million people were forcibly displaced worldwide as a result of persecution, conflict, or generalized violence. Of this 68.5 million, 25.4 million were considered refugees under the UNHCR mandate and the United Nations Relief Works Agency for Palestine mandate. Ultimately, resettlement was the solution for 102 800 refugees. For many refugees, learning a new language, or improving it, is essential in the resettlement process. Previous research examines the role of language testing in immigration, citizenship, and asylum in countries around the world (e.g. Blackledge, 2009; Cooke, 2009; De Jong et. al. 2009; Eades, 2009; Gysen et. al. 2009; Kunnan, 2009; McNamara, 2009; Saville, 2009; Schüpbach, 2009; Shohamy & Kanza, 2009; Zabrodskaja, 2009). These studies respond to the call of critical language testing issued by researchers such as Pennycook (1994), Kramsch (1993), and Shohamy (1998). Critical language testing is influenced by the ideas of Foucault and is defined by Shohamy (1998) as follows, "Critical language testing assumes that the act of testing is not neutral. Rather, it is both a product and an agent of cultural social, political, educational and ideological agendas that shape the lives

of individual participants, teachers, and learners" (p.331). Shohamy (1998) claims that testing is not an isolated practice to measure skills or knowledge, but rather a part of a complex network of societal interactions which influence individual stakeholders. Critical language testing deviates from, and challenges, psychometric traditions and favors more interpretive methods of investigating language testing in reference to the context in which the test operates (Shohamy 1998, p.332). Following Messick's (1989) unified concept of validity, Shohamy (1998) presents several examples such as gatekeeping tests in Australia and US President Bill Clinton's call for "raising standards" (Shohamy 1993, 1996) which explore and emphasize the consequential validity of language tests in larger societal and political contexts.

Few studies investigate ideas of critical language testing at the local rather than national level. King and Bigelow (2016) investigate the local administration of the WIDA Access Placement Test[1] (W-APT) to meet federal requirements of United States (U.S.) K-12 schools to ensure that English language learners have equal access to "meaningful education". Although the requirements come from the national level, King and Bigelow (2016) find that the local administration of the W-APT operates as language policy in unintended and unexpected ways to try to meet the general, theoretical goal of providing "meaningful education" to English language learners. Such findings demonstrate that monitoring consequential validity is important because tests can become de facto policies which then switch to forces that shape the curriculum and structure of programs (King & Bigelow, 2016; Shohamy, 2014). In addition, King and Bigelow (2016) argue that W-APT was the most widely used, but understudied English language test used in the U.S. at that point in time which means the de facto policies it created were not based on a test that stood the scrutiny of research and validation.

King and Bigelow (2016) highlights that failure to properly research and validate a test used at the local level can lead to larger scale issues. Shohamy (1998) promotes the use of various research and data collection tools (e.g. interviews, questionnaires, observations and document analysis, etc.) to describe language test use, impact, and consequences. However, according to Bachman (2005), addressing test use, impact, and consequences is necessary but not sufficient to evaluate a test. Bachman (2005) criticizes critical language testing approaches for not acknowledging the possible links between consequences and the validity of inferences made from tests and, in doing so, failing to provide suggestions on how to anticipate and avoid negative consequences. Negative consequences may be the result of validity issues in the test, which merit investigation.

The present study fills a gap in the research by reviewing an English placement test used on the local level at a refugee resettlement agency. This review is unique due to the specific refugee context with a low education and low literacy test taker population. The aim of this review is not only to provide information about this context for further research but also to investigate links between the validity of inferences made from the test and consequential validity. Reviewing local English placement tests is important to investigate how local agencies attempt to meet federal requirements and support refugees in the resettlement process through English language education. Beyond the role of the test in the local context, these tests need to be reviewed in order to investigate larger implications concerning consequential validity. Improving

[1] Formerly World-Class Instructional Design and Assessment

the quality of tests used by local agencies can promote more successful refugee resettlement and benefit society as a whole.

1.1 Specific context of test review

The U.S. Department of State sets federal requirements for refugee resettlement and funding comes from federal and local grants, but resettlement is administered by local agencies. The U.S. Department of State has partnerships with nine domestic resettlement agencies that have the prerequisite knowledge and resources to resettle refugees. The nine domestic agencies place refugees in approximately 190 communities across the U.S. Each of the nine domestic agencies has local affiliates that monitor community resources and provide support services during the resettlement process (e.g. interpreters, housing, language classes, medical care, employment services, etc.).

This test review was conducted at a local affiliate refugee resettlement agency in the Midwest of the U.S. from 2015 to 2018. In 2015, the agency in question resettled approximately 400 refugees per year. The agency functions as a "U.S. tie site" meaning that it only resettles refugees who already have friends or family members in the area that can provide support. The main populations it resettled at the time of the study were from Southeast Asia, Bhutan, Iraq, Somalia, Congo, Eritrea, and Iran (U.S. Department of State Bureau of Population, Refugees, and Migration, n.d.).

As one of the first steps in the resettlement process, refugees meet with members of the Refugee Learning Center (RLC) of the agency. The RLC is staffed by a full time English language teacher, a coordinator, one to two interns (depending on availability), and volunteers. The RLC offers courses on English Language Learning (ELL), Cultural Orientation, and Job Readiness. The lead English instructor teaches half of the ELL courses and the other half are taught by volunteers. All students have two lessons per day Monday through Thursday and job training or other lessons on Friday. The RLC uses an English exam to place newly arrived refugees in appropriate level English classes. There are usually four levels of ELL courses.

3 Method

3.1 Research Questions

This case study of an English placement test in a refugee resettlement agency set out to answer these research questions:

(1)What are the test format, purpose, use, scoring method, and score reporting of the English placement test at a local refugee resettlement agency in the U.S.?
(2) What is the reliability and validity, in terms of content and construct validity, of the English placement test at a local refugee resettlement agency in the U.S.?

(3) What is the consequential validity and impact of the English placement test? Does the English placement test contribute to meeting the overall goals of the local refugee resettlement agency for its students?

The researcher addressed these questions through a test review with two rounds of data collection.

3.2 Data collection

The researcher had personal experience and knowledge from volunteering full-time as an English language teacher in the RLC two years prior to the test review. The initial data collection at the RLC took place in 2017. Due to the sensitive context of the RLC in which refugees came from backgrounds with little knowledge of technology, diverse religious and cultural beliefs, and trauma, the researcher took no video or audio recordings. The researcher only used written methods of data collection. The data included a written questionnaire filled out by the lead English instructor via email, an in-person follow-up interview with the lead English instructor and the RLC coordinator after completion of the written questionnaire, observations of test administration, a copy of the placement test, and personal communication (email and phones calls) with the English instructor and the RLC coordinator.[2] First, the lead English instructor completed the written questionnaire. Then the researcher observed a test administration and took notes on procedures, oral instructions, location, and other observations about the student's actions while completing the test. Next, the researcher conducted a follow up interview to clarify responses on the questionnaire and ask additional questions about test use and impact.

The second round of data collection took place in 2018. The second data collection followed a needs analysis methodology outlined by Serafini et al. (2015) which emphasizes the need for mixed-methods, pilot studies, multiple sources of information, and triangulation of sources to ensure validity. An inductive and qualitative approach was used to collect data using multiple methods including review of teaching materials, in-person observations, NA questionnaires, focus groups, interviews, and analysis of test scores and placements. The sources of information included teaching materials, current teachers, the RLC coordinator, other members of the RLC (citizenship teachers or other support staff), copies of completed placement tests, and current students. For the purposes of this test review, the needs analysis questionnaire and focus group data will not be discussed.

Participants were informed of the purpose of the research and risks associated with participating. All efforts were made to ensure the confidentiality and anonymity of participants. No identifying information was included in the final data. All efforts were made to remove identifying data from collected materials (i.e. NA questionnaires). Any materials containing identifying data (i.e. placement tests, course enrollment lists, etc.) remained at the refugee resettlement agency. At the end of the study, the refugee resettlement agency was given access to all relevant data and results.

[2] See appendix A for the basic format of the written questionnaire.

The researcher observed one session of each of the four classes taught in the RLC and one administration of the English placement test at which three students were tested and took field notes following the observations. To prepare for observations, the researcher completed a materials review of the books used in each course to identify the main topics and to analyze the amount of speaking, reading, writing, and listening tasks. For each observation, the researcher took notes on how many students attended, the course materials used, the activities, the instructions given by the instructor, and the type of engagement required from students (i.e. speaking, writing, listening, and reading). In addition, the researcher reviewed the English placement exam and score of every student that was in the program at that time (n=36).[3]

After the English placement exam, the researcher observed as the tests were graded by the English teacher using a think-aloud protocol. During this, the researcher asked the teacher for explanations about the grading of the exam and how the result is used for placement decisions.

The semi-structured interviews were conducted individually with program administrators and language teachers. The interviews collected information about the goals of the RLC, purpose of the English courses, use of the English placement exam, and problems the RLC faces. After the test review was completed, the results were shared with the stakeholders and suggestions for improvement were provided.

It was not possible to collect a complete set of test score data. Many of the tests of students placing into the lower levels did not have any answers for sections C, D, and E. The data for section E, the oral section, only included the total score for all 10 questions instead of individual scores for each item.

4 Results and Discussion

After data collection, the researcher coded all notes from class observations, test observations, the written questionnaire, and interviews based on the categories of test format, purpose, use, scoring method, score reporting, validity, and goals of the RLC.

4.1 Research Question One

The researcher used the coded notes along with a logical analysis of the test, test scores, and placement decisions to answer research questions one.

4.1.1 Test Purpose and Use

This section summarizes the coded notes of the written questionnaire and follow-up interview with the RLC coordinator and lead English teacher. The test is used to place students into one of four levels of English Language Learning (ELL) courses offered by the RLC once they arrive at the resettlement agency: A-True Beginner, B-Low Beginner, C-High Beginner/Intermediate, and D-

[3] The data set was incomplete because there were a few missing exams for students in the program.

Intermediate/Advanced. The descriptions of each level and distinctions between them are highly fluid due to the "revolving door" nature of ELL courses (Finn, 2010, pp.589-590)[4]. New students enter courses after the placement test each Friday and students leave class as they gain employment. Attendance can also be irregular due to other student obligations which are part of the resettlement process, such as medical appointments, appointments with social workers, legal appointments, etc. Therefore, considerable variation of student proficiency is possible within each level. According to the current lead instructor's response to the written questionnaire, at the time of the study general descriptions of each level are as follows:

- True Beginner: "Cannot read or write in English. Might not know the alphabet or numbers. Might be able to respond to *What is your name?*"
- Low Beginner: "Can read and write simple sentences in English. Can respond to *What is your name?* Can ask and answer personal questions about themselves. Know alphabet and numbers…Can mostly respond to close-ended questions."
- High Beginner/Intermediate: "Can read simple stories and write in English. Can respond to questions about what they did/are going to do today, yesterday, and tomorrow…Can respond to more open ended questions."
- Intermediate/Advanced: "Can respond to many open-ended questions…Can read short stories by themselves and can write paragraphs. This level is the most varied depending on who is currently in the class."

4.1.2 Test format

This section describes the copy of the placement test with some sample items. The test contains five sections which assess students' reading, writing, listening, and speaking skills. Sections A and B were created by the RLC, whereas sections C, D, and E are from *Interactive English* resources provided for free by the non-profit Intercambio Uniting Communities. A brief description of each section is as follows:

(1) **Section A:**
Picture identification. This section contains 10 pictures in color with a blank space provided to the left of each picture. The test takers' task is to write at least one word for each picture.
Instructions provided: Look at the picture and write the word.

(2)**Section B:**
Numbers: This section contains ten questions which alternate between two question types. For five questions the number is provided numerically and the test taker must fill in the blank with the vocabulary word for the number. For the other five questions the vocabulary word for the number is provided and the test taker must fill in the blank with the numeric representation. Below is an example of two possible test items:

[4] The 'revolving door' is a concept discussed as a central challenge in adult education settings across populations of learners. It is mainly used to refer to community classes offered at little to no cost to participants in efforts to increase literacy rates. In such classes student attendance is not stable due to other obligations to family, works, etc.

- One = _____
- 17 = _____

Instructions provided: Read and write the number.

(3)**Section C:**

Reading Test: This section contains 27 multiple choice questions. There are three possible options to choose from for each question. Some questions require test takers to fill in the blank with the missing word from the sentence. Other questions require test takers to respond to questions or sentences as they would in normal conversation. Below are a few examples multiple choice questions:

- Have they ever _____ their house?
    a. Paint
    b. Painting
    c. Painted
- Did you pay the bill?
    a. Yes, I paid them.
    b. Yes, I paid it.
    c. No, I didn't pay them.
- I am hungry.
    a. You should eat something.
    b. Should you eat.
    c. You don't have to eat.

Instructions provided: Read each item and circle the correct answer.

(4) **Section D:**

Listening Test: This section contains 21 questions. For each question there are three options to choose from. Students must listen to a portion of dialogue and select the correct response.

Instructions provided: Listen and circle the correct answer.

(5) **Section E:**

Oral Evaluation: Students must respond orally to a series of 10 questions asked by the test proctor.

Instructions provided: Proctor tells student that the will ask them some questions.

4.1.3 Test procedure

This section summarizes the data collected from the written questionnaire, follow-up interview, and the notes from the observation of the test administration. The English teacher or the RLC coordinator administers the test as needed each Friday when new refugees arrive. The test is not timed. Students take the test in a classroom at the RLC with ideally one student at each table. Attempts are made to create a comfortable, anxiety-free test environment and to control for any factors of the test setting that might affect student performance. As the EL instructor states:

> Since some students have not been invited into the education system in their home countries, their response or affective filter might be high or low depending on past experiences with schooling or tests. The teacher tries to make it [the test] a light hearted experience.

Studies have shown that it is important to take into account the influence of psychological trauma, including disorders such as Post Traumatic Stress Disorder, when making decisions about teaching and assessment in refugee populations. The physical space in the classroom or test setting can influence student performance due to affective attributes. Rooms without windows and doors may make test takers feel enclosed, while open doors, windows, and the ability to leave the classroom can instill a sense of freedom and empowerment (Isserlis, 2000).

For reading, writing, and listening the test has a paper and pencil format in which the expected response of the test taker must be marked on the test itself. Sections A, B, and C are taken all at once, then a short break is provided, followed by the listening section. After the listening section, tests are collected and the proctor takes students one at a time into another room for the speaking portion of the exam. Directions for each section are written and given orally with motions for true beginners. If test takers struggle visibly to complete the test, the test administrator will prompt the student by asking the questions orally, indicating where to write answers, or guiding the student through a more basic task such as saying the alphabet, counting, or naming items around the classroom. This is to ensure that students with limited exposure to formal education or testing environments still have an equal opportunity to demonstrate their English language ability.

4.1.4 Scoring method and score reporting

This section summarizes the data from the written questionnaire, follow-up interview, notes from the observation of test administration, and note from the think-aloud protocol conducted during the grading of the exam. The placement test is a norm referenced test which uses an interval scale for measurement. In sections A, B, C, and D each question is worth one point. For section A, there are many possible answers and a minimum of one correct answer is sufficient to get the question correct. For sections B, C, and D there is only one correct answer for each question. Each response in section E is scored on a scale of 0-2 according to the following criterion:

- 0=*did not understand the question and/or could not give a response*
- 1=*gave a response, but was not correct according to the language focus in italics*
- 2=*gave a correct response according to the language focus in italics*

There are no specific cut points to distinguish between course levels when making placement decisions. .The test aims to spread out learners by proficiency level and match them to the most appropriate level course offered. However, interpretation of test scores is highly variable due to the fluid environment of the RLC in which volunteers and students move into and out of the program unpredictably. According to the current instructor, score interpretation can also vary due to the rolling schedule of courses. The instructor ultimately uses the test to make a decision about which courses students should be placed into, based on not only the test score, but also the

perception of the current level of students in each course, at what point in the session the student enrolls, and the book being used for each course. Students who can do some of sections A and B and answer questions 1-2 of section E (the speaking portion) usually go into B class. If students are unable to meet the aforementioned standards they are placed into A class. Students who complete sections A and B with high scores are placed in level C or D based on their scores on section D and E of the test. Placement in C or D level depends how the correct number range for reading and speaking matches up with the book the instructor is using for that class (books are switched throughout the year so placement decisions can change). After the instructor grades the tests and makes a placement decision, the instructor enrolls students in courses and hands out class schedules.

4.2 Research Question Two

According to Bachman (1990), the reliability of tests concerns the consistency of measurement, which can be evaluated with statistical analysis and logical analysis. The researcher used logical analysis to identify test items, instructions, and factors in test administration which could result in inconsistent measurement. Due to previously mentioned issues in the data set it is not possible to calculate the internal consistency reliability of the test or to estimate the accuracy of individual scores in making appropriate placement decisions. Inter-rater reliability statistics are not informative because the test is generally only scored by the EL instructor.[5]

To answer research question two, validity is defined as the trustworthiness or plausibility of interpretations and uses of test scores (Messick, 1989; Kane, 2006). Messick (1989) proposed a unified view of validity which emphasized the necessity of integrating elements of content validity, criterion-related validity, and construct validity. The present study will review validity in terms of content validity and construct validity. Content validity is based on expert judgements and concerned with relevance to the target domain and how well items or task content cover the domain (Messick, 1989). Construct validity is based on evidence related to the interpretation of scores and is challenged by two main threats: construct under-representation and construct-irrelevance (Messick, 1989). Criterion validity is not considered in the current paper because its analysis only works when the test is used to predict future performance and a clear definition of successful performance is available; no such definition is available in the scope of this research (Kane, 2004, p.137).

The current review follows Bachman and Palmer's (2010) method to evaluate content validity through description and comparison of tasks. Tasks are defined by three key criteria "(1) closely associated with, or situated in specific situations, (2) goal-oriented, (3) involve active participation of language users" (Bachman & Palmer, 2010, p.59). It is necessary to describe language use tasks and assessment tasks in order to determine if performance on assessment tasks can generalize to situations outside the language test (Bachman & Palmer 2010, p.59). In this framework, interpretations about language ability made from assessment tasks should generalize to target language use tasks. The target language use (TLU) domain is defined as the setting

[5] The Learning Center Coordinator administers the test in the event that the Refugee English Language Instructor is absent.

outside the test where the test-taker would perform language use tasks. The current test review will analyze generalizability in terms of two TLU domains: the language teaching domain and domains outside the classroom, which will be describe in detail in the results.

The researcher analyzed notes from the observation of the test and review of the content of the test to summarize the test tasks. After class observations, the researcher used teaching materials and observations from class to summarize the language teaching domain. Then the researcher compared the TLU tasks from the teaching domain and domains outside the classroom with the test tasks to evaluate content validity.

To explore construct validity the researcher analyzed notes from the observation of the test, test scores, placement decisions, the written questionnaire, interviews, and notes from the think-aloud procedure during the grading of the exam and placement decision.

4.2.1 Reliability

The reading and the listening portions of the test have misleading items which may impact the reliability of the test. For the reading section of the test, 5 out of 27 items were judged by the test reviewer to have more than one possible correct answer in the options provided for selection. Consider the following examples:

- _____ three oranges on the table.
    a. Is there
    b. There's
    c. There are
- I work at a restaurant. I am a _____.
    a. Cook
    b. Teacher
    c. Nanny
- I am hungry.
    a. You should eat something
    b. Should you eat.
    c. You don't have to eat.

In example (i) both answers (b) and answer (c) could be correct depending on whether the grader (and test taker) has prescriptive or descriptive view of grammar. Examples (ii.) and (iii.), could have multiple possible answers based on the background knowledge and logic of the test taker or grader. Although for (ii.), answer (a.) is correct in the test-setting, if the test taker draws on real-world knowledge, for instance countries where teachers work multiple jobs, answer (b.) could also be correct.

In the listening section 4 out of 21 items were judged by the reviewer to have more than one possible answer among the options presented. Consider the following example:

- Test taker hears: When do you read?

a. I always read in bed after I go to sleep.
b. I read in bed while I go to sleep.
c. I always read in bed before I go to sleep.

Presumably, (c.) is the correct answer, but it is possible to imagine a scenario in which (b.) could also be correct.

A statistical analysis of test taker responses to each of the ambiguous items detected would be necessary to make claims about how these items affect the reliability of scores. Nevertheless, based on a logical analysis it is possible the scores would be affected in some manner. Further statistical analysis of reliability was not possible due to the lack of item-level data.

In terms of sections A and B, observations of the test administration show that the reliability is affected by interventions of the test administrators in response to the perception that a student struggles to complete the test. Studies have shown test administrators going "off-script" in this manner can increase the difficulty of test items by introducing more complex language structures into the assessment process. The subjectivity of test administrator input and the unpredictability of how "off-script" interactions will affect test difficulty negatively impacts test reliability (King & Bigelow, 2017, pp.11-12).

4.2.2 Validity

It is useful to discuss the validity of the test in terms of alignment with each of the four levels of ELL courses offered at the RLC separately because the validity of the test varies for each level. There is a difference in the content validity and the weight each construct tested has in making placement decisions for each level.

The main construct measured to determine placement in level A-True Beginner and level B-Low Beginner is literacy. On the theoretical basis that performance reflects ability this portion of the test concerning the literacy construct predicts students' basic literacy in a sufficient manner to make placement decisions in the context of the RLC (McNamara, 2006, p.41-42). No student can be placed into level C or D if they cannot read and write basic numbers and words. If the target language use domain is the classroom then, in terms of performativity, the test task appears to measure the ability to write, as intended, because it requires test takers to write. The task in section B also appears to measure the ability to read as intended because it requires students to read numbers in order to be able to write the corresponding numeral. It seems to measure numeracy (the ability to interpret numbers) and writing as intended because it requires test takers to interpret the number, recall the corresponding vocabulary word, and write the word. Unsuccessful completion of tasks on sections A and B is evidence supporting an inference of low literacy ability and therefore precludes placement in level C or level D. Figure 1 and Figure 2 show how the combined scores from section A and section B, which intend to measure the literacy construct, relate to placement decisions.

Figure 1 illustrates a de facto cut-point, around a score of 12, used to make placement decisions between the lowest level A and the upper levels (B and C). There is only one outlier in the test score data where a test taker scored above the de facto cut-point, but was still placed in level A.

Based on the format of sections A and B one would expect a high probability of simultaneous floor and ceiling effects in score distribution. Figure 1 shows this prediction is partially born out. In terms of floor effects, out of 33 test takers, 21 were placed in the lowest level, 12 of which had a combined score of 0 for sections A and B. Above the de facto cut-point, placement decisions do not reflect scores. This ceiling effect indicates that the measurement of literacy ability may no longer have an effect on placement decisions. This ceiling effect necessitates using the speaking section to make placement decisions between level B and C. Figure 3 illustrates how oral proficiency scores relate to placement decisions. As seen in Figure 3 below, test score data provides evidence for a de facto cut-point that is used to make placement decisions between level B and level C.

The scores on the reading and listening sections would not provide useful information for placement decisions between level A and level B because basic literacy is a prerequisite to be able to respond on those sections of the test.

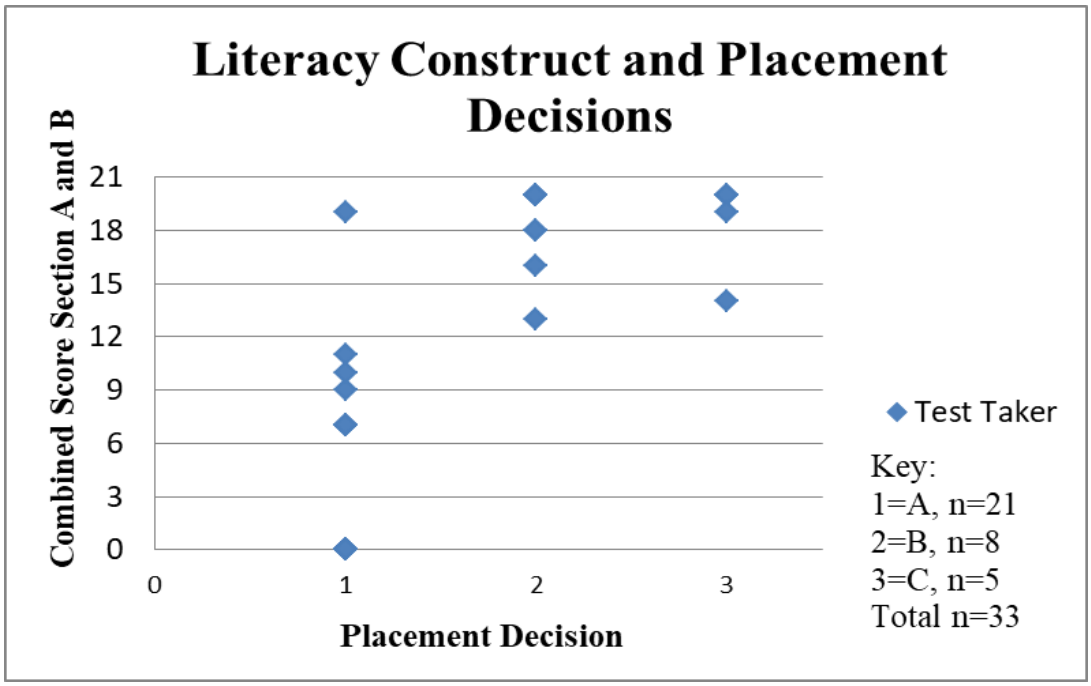

Figure 1. Literacy construct and placement decisions

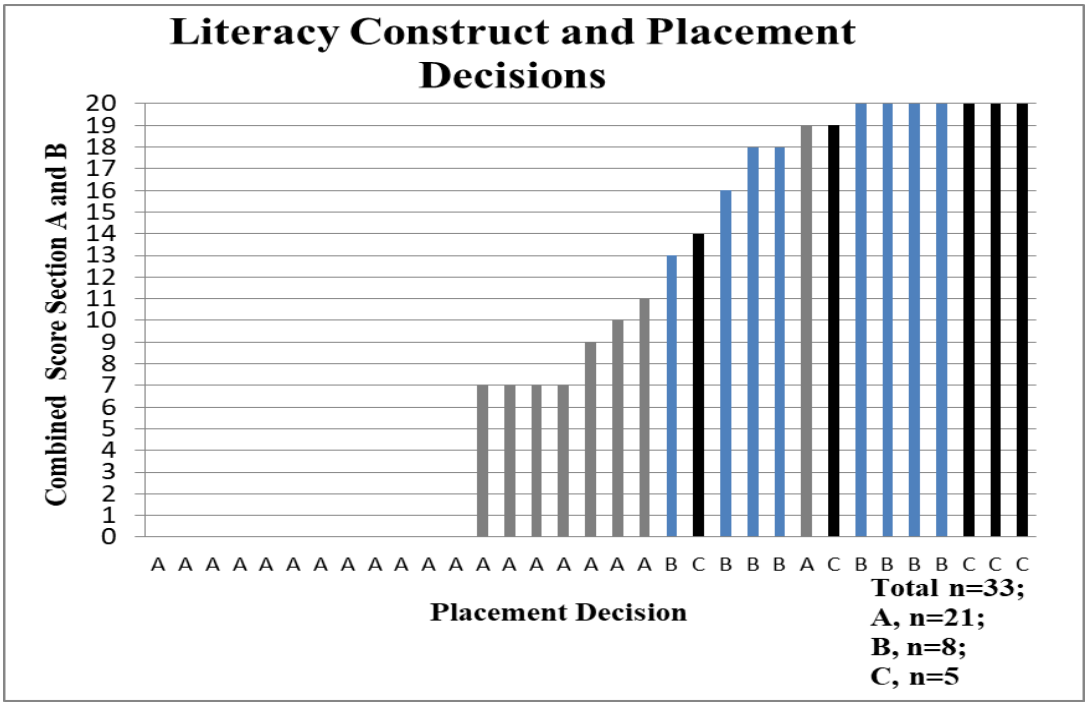

Figure 2. Literacy construct and placement decisions

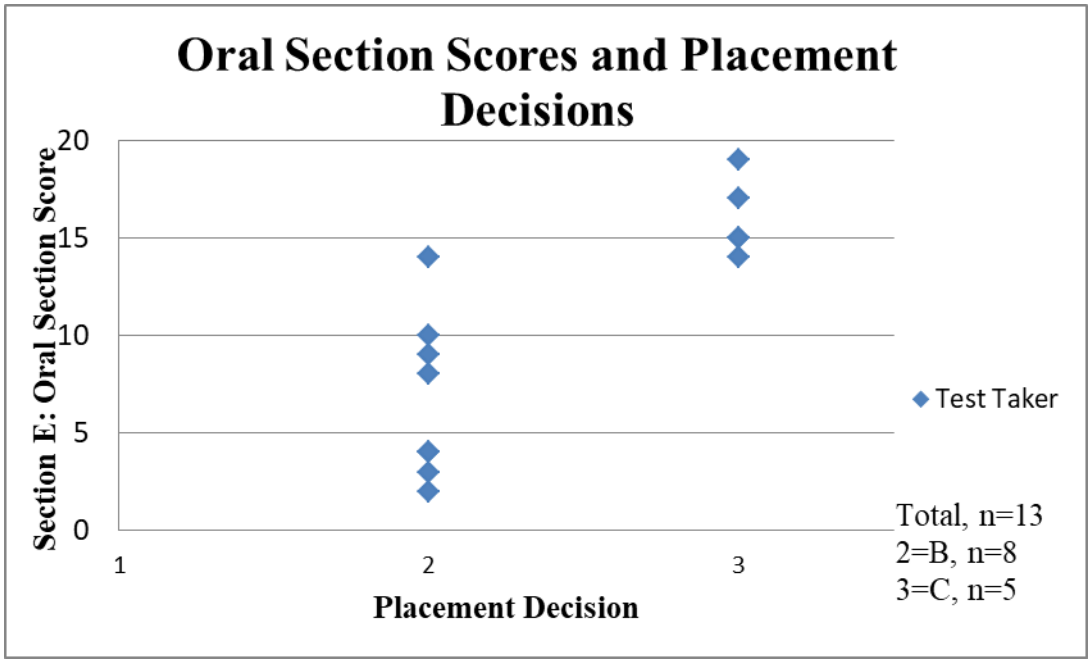

Figure 3. Oral section scores and placement decisions, n=13

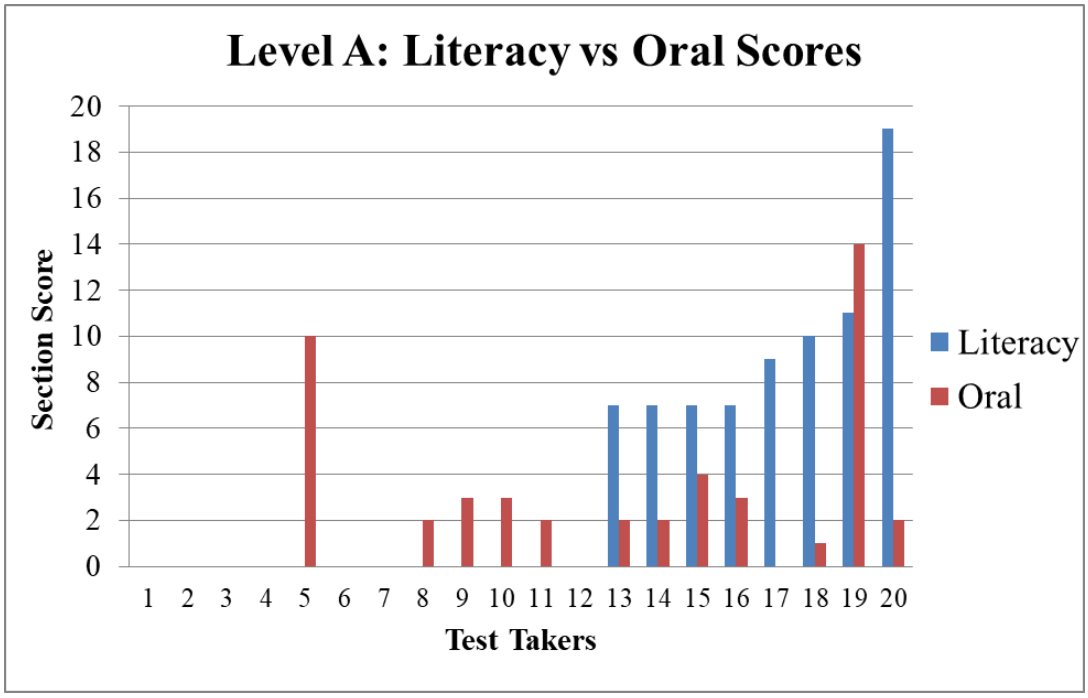

Figure 4. Level A: literacy vs oral scores

	Test Taker (TT) 1, 2, 3, 4	TT 5	TT 6	TT 7	TT 8	TT 9	TT 10	TT 11	TT 12
Literacy Score	0	0	0	0	0	0	0	0	0
Oral Score	0	10	0	0	2	3	3	2	0

	TT 13	TT 14	TT 15	TT 16	TT 17	TT 18	TT 19	TT 20
Literacy Score	7	7	7	7	9	10	11	19
Oral Score	2	2	4	3	0	1	14	2

Table 1. Level A: literacy vs oral scores data

The de facto cut points show sections A and B which measure literacy and section E which measures oral skills predominantly determine placement decisions. Due to this, groups B and C are fairly homogenous; lower level groups, however, are not. Figure 4 above demonstrates how literacy and oral proficiency scores fluctuate in level A, meaning students do not have homogenous literacy and oral proficiency. Observations in the lowest level class showed that while some students could write words and answer basic questions, others were still learning to

hold a pencil and say the alphabet. This could have implications for classroom instruction, materials, and student improvement (or the ability to advance through the levels of the program). King and Bigelow (2017) emphasize that test scores which conceal qualitative differences between students in the same score range can lead to placement errors. To resolve such placement errors King and Bigelow (2017) propose using several modalities to assess proficiency, as the RLC English placement test does (pp.25-26). However, interviews with the RLC coordinator and lead teacher revealed that the main objectives for the courses are not only to improve overall proficiency, but also to develop the literacy skills of students with no, low, or emerging literacy. Checking multiple modalities is not sufficient in the context of placement decisions for the RLC because there is no valuable information provided upon which to make gradient inferences about literacy due to the under specification of the literacy construct.

Data suggest that the literacy construct is defined by performance rather than an evaluation of the skills and strategies necessary to complete each assessment task successfully. In the literature of educational, ethnographic, psychological, and psychometric approaches to investigating literacy, it is accepted that literacy is not dichotomous. There is a spectrum of literacy which should be reflected in assessment (i.e., Maddox & Esposito, 2011). The floor and ceiling effects seen in Figure 2 above suggests that the inferences made from the test become dichotomous when the test items are unable to reflect the spectrum of the literacy construct.

The structure of the test assumes that basic literacy is a prerequisite for L2 reading ability, but lacks alignment indicative of the theoretical relation between the two. If the reading section was a continuation of the literacy sections, but represented higher literacy abilities, then one would expect reading scores to increase as placement decisions increase (assuming that the placement decisions are correct). However, Figure 5 below shows that there is overlap in the reading scores of students who were placed into B and C levels.

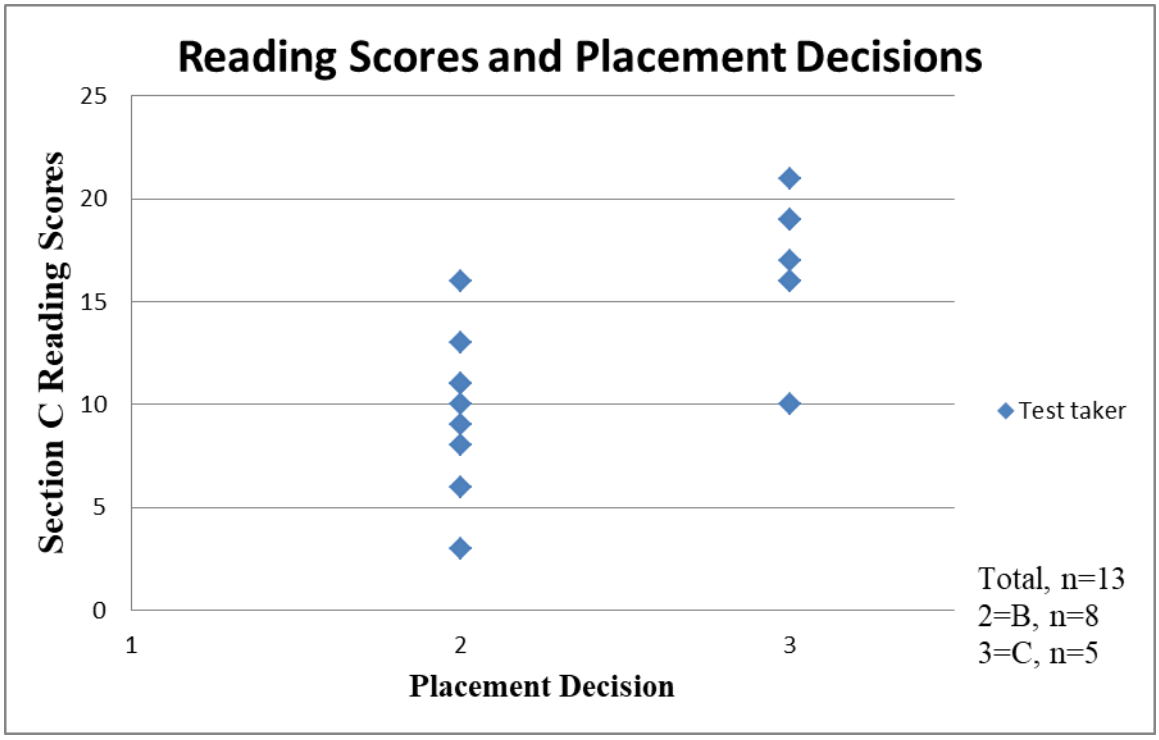

Figure 5. Reading scores and placement decisions, n=13

Despite small sample sizes and uncontrolled variables, the researcher proposes that the effects seen in the literacy sections and reading section scores may be attributed to construct under-representation issues which threaten the validity of inferences made by the test about literacy and reading ability.

Even with the assumption that literacy and reading are separate constructs which can be used to make inferences about ability, the reading portion of the test faces validity challenges. Kim (2015) defines L2 reading ability as strategic competence and language knowledge such as grammatical form, semantic meaning, and pragmatic meaning (p.230). Based on this definition, the reading section lacks construct validity because it does not provide meaningful information about strategic competence. The items on the reading section are only one sentence long which makes it difficult to measure strategic competence including attributes such as metacognitive and cognitive strategies. The materials in the reading section are not authentic because they do not reflect language use in the TLU domain. The reading section is similar to other reading tests such as cloze tests which require a fill-in-the blank response to input. The authenticity of such tests is questionable because seldom does performance in the TLU domain reflect this limited written response format.

Underrepresentation of the literacy construct leads to other issues in terms of predictive validity, authenticity, and content validity. According to interviews with the RLC coordinator and the lead English teacher, promoting self-sufficiency of students is the goal of the RLC. In line with the capabilities approach, literacy performance is essential for human development and well-being (Maddox, 2008, p.189). The test does not align with this goal because test performance does not necessarily generalize to the ability to use basic literacy skills in the real life TLU

domain to complete language use tasks necessary to be self-sufficient such as writing checks or filling out forms with personal information. Levels A and B of the ELL courses do cover the aforementioned real life TLU domain content, but the assessment tasks have a narrow focus in sections A and B on reading and writing of simple words. This underrepresentation of the literacy construct gives the test low content validity in terms of the classroom setting and the TLU domain (Bachman & Palmer, 2010; Lewkowicz, 2000, p.41).

The construct validity of the listening portion of the test is questionable because this portion does not only test listening, but also reading. Input is aural, but the expected response from the student is written. Students must be able to read the response options and understand the vocabulary in order to select the correct answer. This is problematic because a student could possess high listening abilities but score very low if their reading abilities are low. To better measure the construct of listening ability the test should incorporate portions in which the test taker's response to prompt input can written or oral. In addition, this section may test the student's knowledge of grammar because many responses involve selecting the response with the verb in the corresponding form in terms of tense, person, and aspect. Vocabulary and grammar knowledge are essential components of the listening ability, but it is not clear how the items on the listening portion of the test discriminate between test takers with high listening abilities and test takers with low listening abilities. Based on a logical analysis, there are no criteria on which to make claims about the difficulty levels for vocabulary and grammar included in test items. Therefore, little meaningful information for interpretation can be gleaned from test takers missing one item as compared to missing another item. It would be ideal to be able to rank the items in this section on a scale of difficulty determined by the predicted listening ability necessary to answer the item correctly.

Thus far, the discussion of test validity has focused on aspects of construct validity, content validity, predictive validity, and authenticity which are intrinsically linked to the test itself and the TLU domain. The validity of test use to make placement decisions must also consider the score interpretation. Due to the lack of cut scores to make placement decisions, the interpretation of scores is variable and subjective. Placement decisions are made not only based on scores but also the perceived proficiency of students who are currently in each level or the book currently being used for each level. Placements based on perceived proficiency can be inconsistent because they are subjective decisions made by the instructor. It is difficult to control for factors independent of language ability which might affect these decisions. An instructor's perception of student proficiency in each level could change each week based on student performance or response to instruction. It is problematic that the placement decisions rely heavily on the test administrator rather than the test itself.

4.3 Research Question Three

4.3.1 Consequential Validity and Impact of the test

Based on the mission statement of the RLC and the interviews with the RLC coordinator and the lead English teacher, the purpose of the test is to place students into the proper level of ELL classes in order to facilitate the acquisition of skills and knowledge necessary for becoming

self-sufficient in the U.S.. This aligns with the general goal of the RLC to help students as they transition to life in the U.S. The RLC operates in a context of limited resources and competing interests.[6] There are only a limited number of spots available in the program and student eligibility can change based on their employment status or the employment status of a family member.[7] For many refugees in transition to the U.S., ELL skills are not the most urgent need. Therefore, it is not possible for some students to fully participate in the program as they are transitioning to a new culture and trying to meet other basic needs. Teachers within the RLC must be flexible and willing to adapt in order to function in a constantly changing environment and teach in a manner which is sensitive to a spectrum of student needs. The observation of the test administration provided evidence that the RLC used elements of dynamic assessment, such as providing assistance or rephrasing instructions. The written questionnaire and think-aloud protocol during the grading of the test show the fluid nature of score interpretation. These reflect the overall flexibility and adaptability which are necessary in the context of a non-profit ELL program serving a refugee population.

The test has a potential for positive consequences for society and within the educational system using the test because it represents a commitment to helping refugees become part of their new communities. In the RLC this means prioritizing individual student progress and self-paced learning over following a set program structure or guidelines, as can be seen through the flexible use of the exam.

Based on interviews conducted with the RLC coordinator and EL instructor, from the perspectives of the administrators, the test serves its purpose to help make placement decisions. This test review suggests there are validity issues in the placement test which could have broader implications for the quality of the RLC and success of its students. Even decades later, Messick's philosophical discussion of test validity is relevant to address the discrepancy between functionality and lurking potential washback of this test. Messick argues:

> Using test scores that 'work' in practice without some understanding of what they mean is like using a drug that works without knowing its properties and reactions. You may get some immediate relief, to be sure, but you had better ascertain and monitor the side effects. (1989, p.8)

Although the English Placement test is useful to help make decisions in the RLC, it is necessary to consider the impact of validity issues on the practices in the program. If these practices of the program are the side effects of the test, Messick (1989) argues they can lead to social consequences for all stakeholders. It is important to monitor these social consequences considering that tests can become de facto policies which then switch to forces that shape the curriculum and structure of programs (King & Bigelow, 2017; Shohamy, 2014).

---

[6] The RLC is funded through a variety of grants and private donations. The RLC participates in the Voluntary Agencies Matching Grant Program which pushes for participants to reach self-sufficiency within 120-180 days of arrival in the US without access to public cash assistance (Giossi 2016, p.1).

[7] Once students are employed or a member of their household is employed they are no longer eligible for services from the RLC, but are encouraged to participate in other ELL resources at little or no cost offered by other organizations.

The test manifests a tension found in many adult ESL programs in the interaction of two main competing goals: literacy and oral proficiency. In the RLC the ultimate goal is self-sufficiency, yet it is debatable what levels of literacy and oral proficiency are necessary to attain such a goal. Defining self-sufficiency in an ESL program for refugee populations and assessing literacy through the English Placement test is the first step into a larger discourse about the societal values of literacy, multilingualism, self-sufficiency, and social mobility (Ahmed, 2011; Maddox & Esposito, 2011; Shohamy, 2013).

6 Conclusion

Due to the lack of item-level data it was not possible to do statistical tests for reliability. At this point in time, it is not possible to determine if the test is reliable. Mainstream validation considers reliability as a prerequisite for claims about validity (see Kane, 2004). Therefore, comments concerning validity in this paper should be considered as questions for future research rather than strong claims.

According to the findings of this study, placement decisions are made primary based on inferences concerning student literacy and oral proficiency abilities. According to Bachman (2005), validity issues may be linked to the aforementioned social consequences of the test which were hypothesized from a critical testing perspective. Pending future research investigating the consequences of the tests in depth, following Bachman's (2005) suggestions, the researcher provides the following suggestions and guidance to improve possible issues with construct validity and mitigate possible negative consequences.

The construct validity issues in this placement exam reveal a lack of consideration for theory or research during the test development process. It is difficult to create and define constructs to assess literacy and oral proficiency because there is a lack of research on the second language acquisition processes of low education and low literacy learners (Bigelow & Tarone, 2004; Young-Scholten, 2013, 2015). In the RLC, the determining factor for placement in level A is the scores on sections A and B which represent the literacy construct, which implies that the primary focus of level A is literacy acquisition. Only placement decisions for higher levels rely on inferences about listening, reading, and oral abilities. This shows a disjunction in the alignment of the course levels suggesting that oral proficiency skills can develop while courses focus on literacy acquisition. The emphasis on literacy skills as a gateway to courses for developing higher oral proficiency skills could explain possible issues with many students never advancing out of A and B level.[8] Inability to gain oral proficiency and literacy skills has social consequences as previously discussed. Exploring this topic merits further investigation because it is unclear whether the alignment issues concerning literacy and oral proficiency stem from the test or from the curriculum.

More evidence is needed to support inferences concerning literacy ability and oral proficiency ability. Specifically, evidence is needed on the relationship between the two abilities in order to create test items that can be used to make valid inferences and appropriate placement

[8] This commentary is made on the basis of experiences serving as an intern and teacher in the Refugee Learning Center and personal conversations with long standing staff members and teachers.

decisions. The main question for future research pertaining to this context, with a low literacy and low education population, is whether simultaneous literacy and L2 acquisition is possible. If not, is a certain level of L2 oral proficiency a prerequisite to literacy acquisition? Is a certain level of literacy necessary for the acquisition of L2 oral proficiency in instructional contexts (Bigelow et al. 2006)?

Questions about the relationship between oral proficiency and literacy should be explored and the answers to such questions may be found in literacy research, psycholinguistic research on how literacy affects the mind, or second language acquisition research on phonological processing (see Ahmed, 2011; Fracasso et al. 2016; Huettig & Mishra, 2014; Vinogradov & Bigelow, 2011). Understanding the developmental stages and processes of second language acquisition and L2 literacy acquisition for low educated and low literacy populations is necessary to theoretically define constructs in a manner which can accurately assess learners.

To improve the reviewed placement test, a new section could be developed that can make gradient inferences about literacy abilities. This section could focus on assessment tasks that reflect how literacy ability is used to complete language use tasks in the TLU domain. In the absence of research pertaining to L2 literacy acquisition, the author suggests searching for answers and methods from research pertaining to L1 acquisition, L1 reading, and reading education. L1 reading research has shown that phonological awareness is an early predictor of reading ability (Hogan, Catts, and Little, 2005). Speech pathologists administer tests of phonological awareness to help predict future reading ability in children and recommend intervention if necessary. Tests of phonological awareness may be fruitful to assess lower levels of literacy ability that the current test is not sensitive enough to measure. It is possible that these tests would not be applicable in the L2 setting due to a lack of oral proficiency in English of test takers, but tests could be developed for this specific L2 context.

Pending the development of new tests, the reading and listening sections of the reviewed test should be removed. The test score data and de facto cut-points show that sections A, B, and E (literacy and oral sections) are sufficient to make placement decisions. Due to the current satisfaction of the administrators with the placement test, reducing the number of sections will maintain the status quo, but also make the testing process more efficient for both administrators and test takers.

Acknowledgements: Thank you to Dr. Sun-Young Shin of Indiana University Bloomington and Dr. Senyung Lee of Northeastern Illinois University for feedback on early versions of this paper. Thank you to the editors for feedback. All remaining errors are my own.

References

Ahmed, M. (2011). Defining and measuring literacy: Facing the reality. *International review of education, 57*(1-2), 179-195.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford university press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1–34.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Bigelow, M., Delmas, R., Hansen, K., & Tarone, E. (2006). Literacy and the processing of oral recasts in SLA. *Tesol Quarterly*, *40*(4), 665–689.

Bigelow, M., & Tarone, E. (2004). The role of literacy level in second language acquisition: Doesn't who we study determine what we know? *TESOL quarterly*, *38*(4), 689–700.

Blackledge, A. (2009). "As a country we do expect": The further extension of language testing regimes in the United Kingdom. *Language Assessment Quarterly*, *6*(1), 6–16.

Cooke, M. (2009). Barrier or entitlement? The language and citizenship agenda in the United Kingdom. *Language Assessment Quarterly*, *6*(1), 71–77.

De Jong, J. H., Lennig, M., Kerkhoff, A., & Poelmans, P. (2009). Development of a test of spoken Dutch for prospective immigrants. *Language Assessment Quarterly*, *6*(1), 41-60.

Eades, D. (2009). Testing the claims of asylum seekers: The role of language analysis. *Language Assessment Quarterly*, *6*(1), 30–40.

Finn, H. B. (2010). Overcoming barriers: Adult refugee trauma survivors in a learning community. *TESOL Quarterly*, *44*(3), 586–596.

Fracasso, L. E., Bangs, K., & Binder, K. S. (2016). The contributions of phonological and morphological awareness to literacy skills in the adult basic education population. *Journal of Learning Disabilities*, *49*(2), 140–151.

Giossi, T. (2016, December 29). About the Voluntary Agencies Matching Grant Program. Retrieved April 20, 2017, from https://www.acf.hhs.gov/orr/programs/matching-grants

Gysen, S., Kuijper, H., & Van Avermaet, P. (2009). Language testing in the context of immigration and citizenship: The case of the Netherlands and Flanders (Belgium). *Language Assessment Quarterly*, *6*(1), 98–105.

Hogan, T. P., Catts, H. W., & Little, T. D. (2005). The relationship between phonological awareness and reading: Implications for the assessment of phonological awareness. *Language, speech, and hearing services in schools*, *36*(4), 285–293.

Huettig, F., & Mishra, R. K. (2014). How literacy acquisition affects the illiterate mind–a critical examination of theories and evidence. *Language and Linguistics Compass*, *8*(10), 401–427.

Isserlis, J. (2000) *Trauma and the adult English language learner.* Center for Adult English Language Acquisition.

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, *2*(3), 135–170.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement*, 4th ed. (pp. 17–64), American Council on Education and Praeger.

Kim, A. Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing*, *32*(2), 227–258.

King, K.A. & Bigelow, M. (2017). The language policy of placement tests for newcomer English learners. *Educational Policy, 32*(7), 936–968.

Kramsch, C. (1993). *Context and culture in language teaching*. Oxford University Press.

Kunnan, A. J. (2009). Testing for citizenship: The US naturalization test. *Language Assessment Quarterly*, *6*(1), 89–97.

Lewkowicz, J. A. (2000). Authenticity in language testing: some outstanding questions. *Language testing*, *17*(1), 43–64. Maddox, B. (2008). What good is literacy? Insights and implications of the capabilities approach. *Journal of Human Development*, *9*(2), 185–206.

Maddox, B., & Esposito, L. (2011). Sufficiency re-examined: A capabilities perspective on the assessment of functional adult literacy. *Journal of Development Studies*, *47*(9), 1315–1331.

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly: An International Journal*, *3*(1), 31–51.

McNamara, T. (2009). Australia: The dictation test redux? *Language Assessment Quarterly*, *6*(1), 106-111.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, *18*(2), 5–11.

Pennycook, A. (1994). *The cultural politics of English as an international language*. Longman.

Saville, N. (2009). Language assessment in the management of international migration: A framework for considering the issues. *Language Assessment Quarterly*, *6*(1), 17–29.

Serafini, E. J., Lake, J. B., & Long, M. H. (2015). Needs analysis for specialized learner populations: Essential methodological improvements. *English for Specific Purposes*, *40*, 11–26.

Schüpbach, D. (2009). Testing language, testing ethnicity? Policies and practices surrounding the ethnic German Aussiedler. *Language Assessment Quarterly*, *6*(1), 78–82.

Shohamy, E. (1993). *The power of tests: The impact of language tests on teaching and learning*. The National Foreign Language Center at Johns Hopkins University.

Shohamy, E. (1994). The use of language tests for power and control. In J. Alatis, (Ed.), *Georgetown University round table on language and linguistics* (pp.57–72). Georgetown University Press.

Shohamy, E. (1998). Critical Language Testing and Beyond. *Studies in educational evaluation*, *24*(4), 331–345.

Shohamy, E. (2013). The discourse of language testing as a tool for shaping national, global, and transnational identities. *Language and Intercultural Communication*, *13*(2), 225–236.

Shohamy, E. (2014). *The power of tests: A critical perspective on the uses of language tests*. London, England: Routledge.

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing, 13*, 298–317.

Shohamy, E., & Kanza, T. (2009). Language and citizenship in Israel. *Language Assessment Quarterly*, *6*(1), 83–88.

United Nations High Commissioner for Refugees. (2017). *Global Trends: Forced Displacement in 2017*. https://www.unhcr.org/5b27be547.pdf.

U.S. Department of State Bureau of Population, Refugees, and Migration (n.d.). *The Reception and Placement Program.* Retrieved November 1, 2019, from https://2009-2017.state.gov/j/prm/ra/receptionplacement/index.htm.

Vinogradov, P., & Bigelow, M. (2010). Using Oral Language Skills to Build on the Emerging Literacy of Adult English Learners. CAELA Network Brief. *Center for Adult English Language Acquisition*.

Young-Scholten, M. (2013). Low-educated immigrants and the social relevance of second language acquisition research. *Second Language Research*, *29*(4), 441–454.

Young-Scholten, M. (2015). Who are adolescents and adults who develop literacy for the first time in an L2, and why are they of research interest? *Writing Systems Research*, 7(1), 1–3, DOI: 10.1080/17586801.2015.998443.

Zabrodskaja, A. (2009). Language testing in the context of citizenship and asylum: The case of Estonia. *Language Assessment Quarterly*, *6*(1), 61–70.

APPENDIX

CCNEK Language Assessment Questionnaire

1. What is the main purpose of this assessment?
2. If this assessment is a placement test, what are the different levels students can be placed into? Please provide a brief description of each level.
3. What is the test setting like?
    a. Where is the test administered? Please describe the test location.
    b. Who is present during the test? (test proctor and student or test proctor and multiple students?)
    c. Who proctors the test? (the same person every time or different people?)
    d. What time is the test taken? (in the morning, in the afternoon, in the evening, etc)
    e. Are there any other factors about the setting of the test that might influence student performance in your opinion?
4. Is the test taken all in one sitting or divided into sections with pauses between each?
5. What instructions are given before each section of the test? (Only those written on the test? Or additional oral or written instructions?)
    a. Section A: Look at the picture and write the word
    b. Section B: Read and write the number
    c. Section C: Reading Test
    d. Section D: Listening Test
    e. Section E: Oral Evaluation
6. How is each section of the test score?
    a. Section A: Look at the picture and write the word
    b. Section B: Read and write the number
    c. Section C: Reading Test
    d. Section D: Listening Test
    e. Section E: Oral Evaluation
7. How are scores interpreted?
    a. Holistically or section by section?
    b. How are scores used to place students into each level? (Is it strictly a scores 10-20 go into level one and scores 20-30 into level two? Or would poor/good performance on one section of the test automatically place test takers into a certain level?)
8. Are certain test takers given any test accommodations? (such as a scribe, a translator, or extra time, etc)