CLASSIFICATION ACCURACY OF A SELF-REPORT PROFICIENCY ASSESSMENT FOR SPANISH-ENGLISH BILINGUAL SPEAKERS

by

Laura Catarina Herrera, B.S.

A thesis submitted to the Graduate Council of
Texas State University in partial fulfillment
of the requirements for the degree of
Master of Arts
with a Major in Communication Disorders
May 2019

Committee Members:

Maria Resendiz, Chair

Amy Louise Schwarz

Maria Diana Gonzales

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplication Permission

ACKNOWLEDGEMENTS

I would like to thank the Communication Disorders department at Texas State University for supporting my desire to learn and in supporting the completion of a thesis study. I would especially like to acknowledge Dr. Maria Resendiz (my thesis advisor and committee chair), Dr. Amy Louise Schwarz, and Dr. Maria Gonzales (thesis committee members) for their continued commitment to me and my research. I would like to extend an acknowledgement to Idalia Penaloza (graduate student) and Eric Rodriguez (undergraduate student) for volunteering to complete item-by-item reliability.

Finally, I would like to acknowledge my family, especially my parents, for their unconditional support. You all have raised me to strive and achieve my dreams. Thank you for everything you all have done and especially with the sacrifices made so I could complete a master's thesis.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF ABBREVIATIONS

Abbreviation **Description**

WMLS-III - Woodcock-Muñoz Language Survey III

ASHA - American Speech-Language-Hearing Association

SLP - Speech-Language Pathologist

LUQ - Language Use Questionnaire

L1 - first language

L2 - second language

BVNT-NU - Bilingual Verbal Naming Test Normative Update

BTLPT - Bilingual Target Language Proficiency Test

CEFR - Common European Framework of Reference for Languages: Learning, Teaching, Assessment

CALP - cognitive academic language proficiency

LEAP-Q - Language Experience and Proficiency Questionnaire

WIAT-III - Wechsler Individual Achievement Test

RPI - Relative Proficiency Index

LR+ - positive likelihood ratio

LR- - negative likelihood ratio

Austin ISD - Austin Independent School District

DME - direct magnitude estimation

I. LITERATURE REVIEW

The language and cognitive functions in bilingual individuals vary compared to monolingual counterparts based on factors such as vocabulary knowledge, lexical access/retrieval, and executive control (Bialystok & Craik, 2010). The deviations in language and cognitive abilities by bilingual speakers need to be assessed in both languages for professional, clinical, and research use. This thesis focuses on one factor, bilingual language proficiency.

The American Speech-Language-Hearing Association (ASHA) certifies speech-language pathologists (SLPs) to evaluate and treat individuals from birth to geriatric populations in communication (speech and language) and swallowing disorders. ASHA, as of the end of 2018, represented 191,904 personnel including SLPs, audiologists, speech, language, and hearing scientists, and SLP assistants (ASHA, 2018a). From these individuals, only 6% were identified as bilingual service providers (ASHA, 2018a). ASHA is currently attempting to increase the number of bilingual service providers, which includes bilingual SLPs, to meet a growing population of speakers who speak a language other than English (Ryan, 2013).

Standardized norm-referenced assessments, such as the Woodcock-Muñoz Language Survey-III (WMLS-III; Woodcock, Alvarado & Ruef, 2017) can be used to determine proficiency of Spanish-English bilinguals. In fact, the WMLS-III (Woodcock et al., 2017) is so highly regarded that it has been used as the reference or "gold standard" test in studies validating computerized neuropsychological tests (Holliday, Navarrete, Hermosillo-Romo, Valdez, Saklad, Escalante, & Brey, 2003), validating language instruments used for English language learners (Pray, 2005), and in comparing level of

bilingualism to self-report fluency levels in Spanish and English (Gasquoine, Croyle, Cavazos-Gonzalez, & Sandoval, 2007).

The WMLS-III (Woodcock et al., 2017) takes two to three hours to administer and is priced at over $1000, making it an inefficient and uneconomical option for determining proficiency in clinical practice. For example, ASHA does not require bilingual SLPs to take a standardized norm-referenced test, such as the WMLS-III (Woodcock et al., 2017), to demonstrate proficiency in Spanish and English for clinical service delivery. Instead, ASHA allows bilingual SLPs to self-report whether they believe they are sufficiently bilingual in more than one language for purposes of diagnosing and treating individuals who speak more than one language (ASHA, 2018b).

The Language Use Questionnaire (LUQ; Kiran, Peña, Bedore, & Sheng 2010), is a self-report instrument that gathers information about language history, current language use, and proficiency across a variety of dimensions (e.g., casual vs. informal settings, speaking vs. listening) for the languages spoken by an individual. Although the LUQ (Kiran et al., 2010) has been used to determine the proficiency levels of individuals for treatment and research purposes (Gray & Kiran, 2012), the accuracy of the LUQ in identifying language proficiency in comparison to a "gold standard" test, such as the WMLS-III (Woodcock et al., 2017), has not been examined. The current study fills this gap in the literature by determining the level of agreement between bilingual adults' criterion-referenced self-assessment of their proficiency in Spanish and English using the LUQ (Kiran et al., 2010) compared to their proficiency scores in Spanish and English achieved on the standardized norm-referenced WMLS-III (Woodcock et al., 2017).

To provide a context for this study, a discussion of the various factors of bilingualism that can influence a bilingual individual's proficiency profile will be examined. These factors are then considered in terms of development and use of current standardized norm-referenced assessments and criterion-referenced assessments. Furthermore, a discussion about ASHA's current standards for identifying bilingual practitioners will be discussed. Finally, an in-depth explanation of the two assessments compared in this study—the LUQ (Kiran et al., 2010) and the WMLS-III (Woodcock et al., 2017) —will be provided along with a thorough explanation of a classification accuracy study.

Bilingual Proficiency Assessment

A bilingual individual's neurological framework varies in language and cognitive processing compared to the monolingual individual (Bialystok & Craik, 2010). Furthermore, bilingual individuals vary in levels of first language (L1) and second language (L2) proficiency (Valdés & Figueroa, 1994). Proficiency can be measured using one or a combination of the following: (a) current language use (e.g., speaks Spanish 70% of the time and English 30% of the time) (De Houwer, 2017), (b) age of first exposure to the language(s) (e.g., first exposed to Spanish at birth, first exposed to English at age 4) (Perani, Abutalebi, Paulesu, Brambati, Scifo, Cappa & Fazio, 2003), (c) overall ability, (d) listening ability, (e) speaking ability, (f) reading ability, and (g) writing ability in the language(s) (Cummins, 2000). Variability exists in the factors listed above for bilingual individuals, which creates variations in how proficiency in L1 and L2 are measured and determined.

Standardized norm-referenced assessments can be used to determine the proficiency profile of bilingual individuals. Unfortunately, these assessments are time consuming and costly. For instance, the Bilingual Verbal Naming Test Normative Update (BVNTU; Muñoz-Sandoval, Cummins, Alvarado, & Ruef, 2005) is a standardized norm-referenced assessment of verbal knowledge that costs about $555 and must be administered with another bilingual assessment to determine proficiency in both languages. The single administration of the BVNTU takes approximately 30 minutes to complete, and when combined with another standardized assessment will take additional time (Texas Statewide Leadership for Autism Training [TSLAT], 2015). Researchers and SLPs need an efficient and cost-effective assessment to accurately and efficiently measure proficiency of bilingual individuals.

ASHA Standards for Proficiency

Professionals who work with monolingual and/or bilingual individuals are just some of the individuals who need to be assessed in language proficiency to determine their ability to provide adequate and ethical services. Various professional agencies require different standards for a professional to be considered bilingual for their field of work. For instance, teachers in the state of Texas are required to complete the criterion-referenced proficiency test, Bilingual Target Language Proficiency Test (BTLPT), as part of their bilingual educator requirements in order to be certified as a bilingual teacher (Arroyo-Romano, 2016). The BTLPT takes 5 hours and costs $116 to complete, which adds to the many previous finances made towards their bilingual education certification (Arroyo-Romano, 2016; Texas Educator Certification Examination Program (2019). Conversely, ASHA has bilingual practitioners self-identify if they are a qualified

bilingual service provider by taking into consideration the following criteria: (a) have near-native or native proficiency in the other language, (b) are knowledgeable of language development in monolinguals and bilinguals, (c) can administer and analyze an evaluation in the other language, (d) can provide treatment in the other language, and (e) are aware of cultural differences (ASHA, 2018b; Cornish, 2011).

Proficiency Assessments & Correlational Studies

The two types of proficiency assessments discussed in this section include standardized norm-referenced assessments and criterion-referenced assessments. Standardized norm-referenced proficiency assessments are one means of determining proficiency of bilingual speakers in their L1 and L2. As Hulstijn (2015) explains, it is arguably difficult to compare language proficiency between languages that are based in different syntactic forms (e.g., subject-verb-object versus subject-object-verb sentence formation), vocabulary use (e.g., multiple words for one object based on context of use), and speech production (e.g., pronunciation of single phonemes). Hulstijn (2015) suggests that the best possible solution to measure proficiency is by using standardized norm-referenced assessments that are "designed to tap roughly the same language proficiency component in each language and compare bilinguals' performance to the performance of native-speaker reference groups in each language" (p. 139-140). A widely used standardized norm-referenced assessment that attempts to meet this need is the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR; Council of Europe, 2001). The CEFR is designed to determine an individual's proficiency level for academic purposes and is currently available in multiple languages (Council of Europe, 2001; Hulstijn, 2015).

The CEFR (Council of Europe, 2001) is a highly regarded proficiency assessment, but it is not as widely popular in the United States. Currently, the WMLS-III (Woodcock et al., 2017) standardized norm-referenced test is often used to assess proficiency levels for Spanish-English bilinguals (e.g., Mahon, 2006; Pray, 2005). The WMLS-III (Woodcock et al., 2017) takes 2-to-3 hours to administer, which is rarely feasible in clinical practice given the caseloads of speech-language pathologists. The WMLS-III (Woodcock et al., 2017) includes the following six levels of proficiency: (a) Initial Development, (b) Early Development, (c) Continuing Development, (d) Emerging Proficiency, (e) Proficient, and (f) Advanced Proficient.

The WMLS-III (Woodcock et al., 2017) is designed to assess individuals' cognitive academic language proficiency (CALP; Cummins, 2000), which is a decontextualized and cognitively demanding form of language often used in academic settings (Aukerman, 2007). The six levels of proficiency in the WMLS-III provide information on academic language proficiency and the possible academic success of the individual (Woodcock et al., 2017). The WMLS-III only considers two levels comparable to a native speaker, Proficient and Advanced Proficient (Woodcock et al., 2017).

However, many scholars contend that native-like speakers are not the only group of bilinguals who should be considered proficient because near native-like proficiency is the highest level of proficiency that can be attained by sequential L2 speakers (Hyltenstam & Abrahamsson, 2000; Morgan-Short, Steinhauer, Sanz, & Ullman, 2012; White & Genesee, 1996). Sequential L2 speakers are those who learn their second

language after mastering their first language. Following precedent set by authors, the current study will classify near native-like speakers as proficient as well.

In the WMLS-III (Woodcock et al., 2017) individuals identified as having Emerging Proficiency are described as demonstrating an "understanding of the specialized academic language…but still [require] some instructional scaffolding in the classroom environment for effective learning" and their "receptive and expressive proficiency is near the level of average native-language peers" (Woodcock et al., 2017, p. 62). In order to ensure that all proficient bilingual individuals on the bilingual continuum (Cummins, 2000) were captured in this study, Emerging Proficiency was included in the proficient category.

Criterion-referenced self-report proficiency assessments were developed to meet the need for an efficient and cost-effective assessment of bilingual individuals. Two commonly used criterion-referenced language self-reports are the LUQ (Kiran et al., 2010) and the Language Experience and Proficiency Questionnaire (LEAP-Q; Marian, Blumenfeld & Kaushanskaya, 2007). These criterion-referenced language self-reports both assess contributing factors to proficiency (e.g., education level, exposure to the languages in the domains of reading and speaking) and confidence in specific language domains (e.g., reading and speaking).

There are also differences between the LUQ (Kiran et al., 2010) and the LEAP-Q (Marian et al., 2007). While the LUQ (Kiran et al., 2010) has individuals rate their proficiency in reading and writing in both languages, the LEAP-Q (Marian et al., 2007) only has individuals evaluate language preferences in reading and writing. The LUQ (Kiran et al., 2010) has individuals consider all language domains for both languages

whereas the LEAP-Q (Marian et al., 2007) has individuals provide preferences for either language by percentage values. Kiran and Iakupova (2011) contend that criterion-referenced self-reports, such as the LEAP-Q (Marian et al., 2007), do not capture the full extent of an individual's language use and history.

The LUQ (Kiran et al., 2010) captures the language profile of a bilingual individual in the domains of overall language abilities, listening, speaking, reading, and writing (Gray & Kiran, 2012; Kiran & Iakupova, 2011). It is a sensitive criterion-referenced self-report proficiency assessment for individuals who speak the following language combinations: English-Spanish, English-Hindi, English-Russian, English-Mandarin, English-Kannada, English-Turkish and English-Arabic (Kiran & Iakupova, 2011). The LUQ (Kiran et al., 2010) takes approximately 20 minutes to complete and is freely available online, making it a more time-efficient assessment and cost-effective, compared to the aforementioned "gold standard" counterpart, the WMLS-III (Woodcock et al., 2017). For these reasons, the LUQ (Kiran et al., 2010) is being used in clinical and research settings to determine a bilingual individual's proficiency profile (e.g., Gray, 2017; Gray & Kiran, 2012; Kiran & Iakupova, 2011).

Although a classification accuracy study between a criterion-referenced self-report language survey and a standardized norm-referenced test of proficiency has not been conducted, researchers have conduced correlational studies with contradicting results (Gaffney, 2018; Gollan, Wissberger, Runnquvist, Montoya, & Cera, 2011; Marian et al., 2007). For example, the criterion-referenced LEAP-Q (Marian et al., 2007) was externally validated by correlational analysis with subtests from standardized norm-referenced assessments including the Woodcock-Johnson Tests of Achievement

(Woodcock, McGrew, & Mather, 2001), the Woodcock-Muñoz Tests of Achievement (Muñoz-Sandoval, Woodcock, McGrew, & Mather, 2005), and the Peabody Picture Vocabulary Test in Spanish (Dunn, Padilla, Lugo, & Dunn, 1986) and English (Dunn & Dunn, 1997). These comparisons have demonstrated a positive correlation between the criterion-referenced self-report and the standardized norm-referenced assessments for global proficiency ratings (Gollan et al., 2011; Marian et al., 2007), but mixed results for proficiency ratings in the individual language domains, such as listening, speaking, reading and writing (Gaffney, 2018).

Delgado, Guerrero, Goggin, & Ellis (1999) completed a correlational analysis of proficiency between self-ratings and a standardized norm-referenced assessment. Participants completed self-ratings on a five-point Likert scale for Spanish and English in areas of overall fluency, speaking, listening, reading and writing (Delgado et al., 1999). The self-ratings were then correlated with scores from the Woodcock-Muñoz Language Survey (Woodcock & Muñoz, 1993). Delgado et al. (1999) found correlations between self-rating and standardized norm-referenced assessment scores in Spanish, but not in English. Because of the significant correlations in Spanish, but not in English, results from Delgado et al. (1999) further corroborate the variation in correlational studies.

Classification Accuracy

Given that the LUQ (Kiran et al., 2010) is used to review aspects of language proficiency (Kastenbaum et al., 2018) and is a criterion-referenced assessment, classification accuracy of the self-report is needed to justify its use for identifying proficient and non-proficient bilinguals. Classification accuracy aids in determining "whether the external evidence suggests that a change to a different diagnostic tool

should be considered" for screening, diagnosis, and/or differential diagnosis (Dollaghan, 2007, p. 81). In a classification accuracy study, all participants complete two assessments, the "gold standard" and the new test, typically to determine diagnostic accuracy. From the results collected from both tests, participants fall into one of four groups: (a) both tests agree on the presence of a diagnosis (true positive), (b) both tests agree on no presence of a diagnosis (true negative), (c) only the "gold standard" identifies the presence of diagnosis (false negative), or (d) only the new test identifies the presence of diagnosis (false positive). For this study, proficiency (proficient or non-proficient) will be used in place of diagnosis.

Many scholars report likelihood ratios when evaluating the results of classification accuracy studies (see Dollaghan, 2007 for a discussion). The magnitude of the positive likelihood ratio indicates the level of confidence in the result that an individual who tests as having a diagnosis on the criterion-referenced test has a diagnosis. The magnitude of the negative likelihood ratio indicates the level of confidence in the result that an individual who tests as not having a diagnosis on the criterion-referenced test does not have a diagnosis (Sackett, Haynes, Guyatt, & Tugwell, 1991; Sackett, Straus, Richardson, Rosenberg, & Haynes, 2000). Sackett et al. (1991; 2000) offer benchmarks for interpreting likelihood ratios for diagnostic accuracy studies. In the current study, these benchmarks were applied to identify the likelihood of an individual having proficiency or not having proficiency in a language. Table 1 provides the benchmarks and adapted descriptions used to interpret likelihood ratios in the current study based on benchmarks established by Sackett et al. (1991; 2000).

Table 1: Likelihood Ratio Interpretation

LR+		LR-	
Value	Indication	Value	Indication
≥ 10	Most likely is proficient	≤ 0.10	Most likely is not proficient
3	Suggestive but unable to determine proficiency	≤ 0.30	Suggestive but cannot rule out proficiency
1	Unable to determine proficiency	1	Unable to rule out proficiency

Summary & Research Question

The variability in a bilingual individual's proficiency profile needs to be captured for research, treatment and for ASHA's standards for ethical practice. ASHA currently has bilingual practitioners self-identify their proficiency in a second language for practice, which is a time-efficient and cost-effective method of assessment. However, the effectiveness of a criterion-referenced self-identification of proficiency has yet to be determined. Standardized norm-referenced and criterion-referenced assessments have been identified in the literature, specifically the criterion-referenced LUQ (Kiran et al., 2010) and the standardized norm-referenced WMLS-III (Woodcock et al., 2010). These assessments were developed to identify the bilingual profile in overall proficiency, listening, speaking, reading and writing. The current study is a classification accuracy study that will test how well the criterion-referenced LUQ (Kiran et al., 2010) accurately classifies proficient and non-proficient Spanish-English bilinguals in the area of overall language abilities and specific language domains (listening, speaking, reading, and writing) in Spanish and English compared to the standardized norm-referenced WMLS-

III (Woodcock, et al., 2017), the "gold standard" in the current study. The specific research question is:

Will the criterion-referenced LUQ (Kiran et al., 2010) accurately classify proficient and non-proficient Spanish-English bilinguals in Spanish and English in both formal and informal contexts in the areas of (a) overall proficiency, (b) listening, (c) speaking, (d) reading, and (e) writing, when compared to the WMLS-III (Woodcock et al., 2017) standardized norm-referenced test?

The research hypothesis is that likelihood ratios for both languages in each of the five areas listed above will approach the "most likely" confident ranges shown in Table 1. The null hypothesis is that the likelihood ratios for both languages in each of the five areas listed above will not approach the "most likely" confident ranges shown in Table 1. The results of the current study will offer preliminary evidence whether the criterion-referenced LUQ (Kiran et al., 2010) is a time-efficient and cost-effective alternative to the time-intensive and expensive standardized norm-referenced WMLS-III (Woodcock et al., 2017) assessment for determining proficiency in both Spanish and English in (a) overall proficiency, (b) listening, (c) speaking, (d) reading, and (e) writing.

II. METHOD

Participants

The minimum number of participants for a classification accuracy study of a screening study to have enough power to find an effect is 34 participants (Bujang & Adan, 2016). A total of 39 participants ranging in age from 20 to 44 years (*M*=24.64 years, *SD*=5.18) with 35 females participated in the current study. These participants met the following inclusion criteria: (a) Spanish-English bilingual (proficient and non-proficient) (b) in the process of completing a bachelor's degree or previously earned a bachelor's degree. Of the participants, 10% were undergraduate students, 87% were graduate students and 2% held a bachelor's degree and were not currently enrolled in a graduate program. These participants were recruited in two phases. In the first phase, 24 participants ranging in age from 21 to 44 years (*M*=25.5 years, *SD*=6.11, 21 females) were recruited to participate in a storybook translation study that included the LUQ (Kiran et al., 2010) and WMLS-III (Woodcock et al., 2017) assessments in the test battery (Schwarz, Resendiz, & Gonzales, in preparation). In the second phase, 15 participants ranging in age from 20 to 31 years (*M* = 23.33 years, *SD* = 2.85, 14 females) were recruited so that the current study would have enough power to conduct a classification accuracy test comparing the LUQ (Kiran et al., 2010) and the WMLS-III (Woodcock et al., 2017).

Participants reported their language history and language ability using the LUQ (Kiran et al., 2010). Language history included the following areas: exposure (hearing, speaking and reading) across age intervals, levels of confidence (hearing, speaking, and reading) across age intervals, daily use, family language history, education, and language

ability rating (Kastenbaum et al., 2018). The language ability rating required the participants to rate themselves in the areas of (a) overall ability, (b) speaking in casual conversations, (c) listening in casual conversations, (d) speaking in formal situations, (e) listening in formal situations, (f) reading, and (g) writing (Kastenbaum et al., 2018). Participants reported a range in age of first exposure (hearing, speaking, and reading) and in confidence levels (hearing, speaking and reading) in Spanish and English.

The standardized scores from the WMLS-III (Woodcock et al., 2017) indicate the predicted performance on academic tasks in each of the language domains assessed.

The language history gathered from the LUQ (Kiran et al., 2010) and the standardized scores from the WMLS-III (Woodcock et al., 2017) represent a range in proficiency across participants.

Materials

Language Use Questionnaire (LUQ). The LUQ (Kiran et al., 2010) was used as the criterion self-report measure for bilingual individuals to rate their proficiency for Spanish and English. Individuals rated themselves on a 5-point Likert scale. The number one on the scale signifies non-fluent and the number five on the scale signifies native fluency. The individuals rated themselves on the scale in the following language domains for both languages: (a) overall ability, (b) speaking in casual conversations, (c) listening in casual conversations, (d) speaking in formal situations, (e) listening in formal situations, (f) reading, and (g) writing (Kiran et al., 2010).

Woodcock-Muñoz Language Survey III (WMLS-III). The WMLS-III (Woodcock et al., 2017) is a standardized norm-referenced assessment used to evaluate language proficiency in Spanish and English. The WMLS-III (Woodcock et al., 2017)

consists of eight subtests to measure the following areas of language: (a) overall proficiency, (b) listening, (c) speaking, (d) reading, and (e) writing (see Table 2). The WMLS-III (Woodcock et al., 2017) language proficiency score is calculated by entering the raw scores into the online scoring system (https://www.wjscore.com) which calculates the standardized score and language proficiency.

Table 2: Language Domains Measured by Subtests

Language Domain	Woodcock-Muñoz Language Subtests
Listening	Test 1: Analogies Test 2: Oral Comprehension
Speaking	Test 3: Picture Vocabulary Test 4: Oral Language Expression
Reading	Test 5: Letter-Word Identification Test 6: Passage Comprehension
Writing	Test 7: Dictation Test 8: Written Language Expression

Individuals' language proficiency may fall into one of the six predetermined levels of language proficiency: (a) Initial Development, (b) Early Development, (c) Continuing Development, (d) Emerging Proficiency, (e) Proficient, and (f) Advanced Proficient. Only two of the six levels are considered proficient as determined by the authors of the WMLS-III (Woodcock et al., 2017): Proficient and Advanced Proficient. For the current study, individuals who were identified as having Emerging Proficiency were also included in the proficient group (see Table 3). As previously explained, bilingual individuals who acquire their L2 after acquiring their L1 can achieve near native-like proficiency as the highest level of proficiency in their L2 (Hyltenstam & Abrahamsson, 2000; Morgan-Short et al., 2012; White & Genesee, 1996). The WMLS-III (Woodcock et al., 2017) classifies Emerging Proficiency as an equivalent to near

native-like proficiency. Therefore, individuals were included in the proficient group if they were identified by the WMLS-III (Woodcock, et al., 2017) as Advanced Proficient, Proficient, or Emerging Proficiency.

Table 3: Proficiency Categories for Current Classification Accuracy Study

Classification Accuracy Study	Classifications of Proficiency Applied to the LUQ	WMLS-III Classifications for Proficiency
Non-proficient	1 – Non-proficient	Initial Development
	2	Early Development
	3	Continuing Development
	4	
Proficient	5 – Native-like proficiency	Emerging Proficiency
		Proficient
		Advanced Proficient

Woodcock et al. (2017) did not complete a classification accuracy study to determine classification accuracy of the WMLS-III (Woodcock et al., 2017). The strong reliability and validity values that assessment has obtained within each test cluster when compared to other available standardized assessments that determine language proficiency make the WMLS-III a good "gold standard" for the current study (Woodcock et al. 2017). Reliabilities for each test cluster (see Table 4) of the WMLS-III (Woodcock et al., 2017) were calculated using the split-half procedure. Internal reliability for each test was calculated by first computing the raw scores for the odd and even numbered items. Correlations were then calculated between the odd and even raw scored items (Woodcock et al., 2017). To calculate the test cluster reliability, a correlation between each test that forms the cluster was conducted. Table 4 is an adaptation of the reliability results for the language domains (test clusters) used in the study as calculated by Woodcock et al. (2017) for the 20+ age group. As Woodcock et al. (2017) reports, correlations are generally higher when completed within language domains compared to

correlations completed between subtests/clusters of different language domains. The differences in the correlational relationships between subtests indicate that the "WMLS-III clusters measure distinct language domains" (Woodcock et al., 2017, p. 117) which indicate a good "gold standard" for language domains.

Table 4: WMLS-III Reliability Statistics

Language Domain	Correlation (r) Results
Overall Proficiency	0.97
Listening	0.89
Speaking	0.94
Reading	0.95
Writing	0.88

Reliability and validity of the standardized norm-referenced assessment were previously determined by the authors of the test (Woodcock et al., 2017). The reliability coefficients reported for the WMLS-III (Woodcock et al., 2017) are above 0.80, which is adequately reliable (Webb, Shavelson, & Haertel, 2006). There is concurrent validity for the WMLS-III (Woodcock et al., 2017), when compared to other assessments, such as the Wechsler Individual Achievement Test (WIAT-III; Wechsler, 2009).

Procedure

The procedure required that each participant complete the criterion-referenced self-report LUQ (Kiran et al., 2010) and the standardized norm-referenced WMLS-III (Woodcock et al., 2017), which in this study is considered to be the "gold standard" assessment. Examiners were undergraduate and graduate students who were either monolingual English speakers or bilingual Spanish-English speakers. All examiners were trained by professors and graduate students to administer both assessments. For

each participant, the LUQ (Kiran et al., 2010) was administered first, followed by the WMLS-III (Woodcock et al., 2017) so that the participants self-ratings on the LUQ (Kiran et al., 2010) would not be influenced by the participants' experiences taking the WMLS-III (Woodcock et al., 2017).

Participants were instructed to complete all sections of the LUQ (Kiran et al., 2010) which included ratings of exposure, confidence, language history, education, family history, language use, and language ability rating (see Table 5). Completion of the LUQ (Kiran et al., 2010) took up to 20 minutes to complete. Following completion of the LUQ (Kiran et al., 2010), participants were administered by a trained examiner both the Spanish and English subtests of the WMLS-III but not in a particular order.

Table 5: LUQ Language Ability Ratings based on a 5-point Likert-type scale.

			Mean	**Min. & Max.**	**SD**
Overall Ability		Spanish	3.79	2-5	1.15
		English	4.72	3-5	0.51
Listening					
	Casual	Spanish	4.13	1-5	1.19
		English	4.95	4-5	0.22
	Formal	Spanish	3.72	1-5	1.23
		English	4.87	4-5	0.34
Speaking					
	Casual	Spanish	3.64	1-5	1.37
		English	4.82	3-5	0.45
	Formal	Spanish	3.18	1-5	1.45
		English	4.74	3-5	0.55
Reading		Spanish	3.82	1-5	1.09
		English	4.85	3-5	0.49
Writing		Spanish	3.26	1-5	1.35
		English	4.82	3-5	0.45

* 1 = Non-fluent, 5 = Native fluency; Min. = minimum; Max. = maximum.

The WMLS-III (Woodcock, et al., 2017) consists of eight subtests in each language that assess proficiency in the areas of (a) speaking, (b) listening, (c) reading, and (c) writing. Examiners followed the protocol of completing testing in a quiet environment with the stimulus book, protocols, computer, and headphones. The audio portion of the WMLS-III (Woodcock et al., 2017) was administered with over-the-ear headphones. After the completion of either the Spanish or English WMLS-III (Woodcock et al., 2017), participants were instructed to complete the assessment in the other language. Completion of the Spanish and English WMLS-III (Woodcock et al., 2017) took subjects one to three hours, depending on language proficiency. The examiners completed testing following the guidelines in the Examiner's Manual of the WMLS-III (Woodcock, et al., 2017). The average calculated standardized scores and the corresponding proficiency level from the participants can be found in Table 6.

Table 6: WMLS-III Standards Scores and Proficiency Level

		Mean	**Proficiency Level**	**Min. & Max.**	**SD**
Overall Ability	Spanish	40.67	Continuing Development	0-94	37.01
	English	87.64	Proficient	52-100	11.43
Listening	Spanish	41.87	Continuing Development	0-96	37.00
	English	86.15	Proficient	43-99	13.66
Speaking	Spanish	25.10	Continuing Development	0-87	31.91
	English	73.94	Emerging Proficiency	18-99	21.43
Reading	Spanish	54.97	Emerging Proficiency	2-98	36.95
	English	88.10	Proficient	53-100	12.05
Writing	Spanish	48.61	Continuing Development	1-99	34.86
	English	92.94	Proficient	58-100	8.68

* Min. = minimum; Max. = maximum.

Reliability

Inter-rater reliability was conducted by two Spanish-English bilingual graduate and undergraduate research assistants. The graduate research assistant reported Spanish as her native language. She was exposed to Spanish from birth and felt comfortable speaking Spanish from the time she started talking. The graduate research assistant's first exposure to English was at 5 years old. She began speaking English at 9 years old and felt comfortable speaking English at the age of 13. The undergraduate research assistant reported being exposed to Spanish and English at birth. He started speaking Spanish at age 3; English was his dominant language by 5 years old. He experienced some language loss in Spanish at age 5 and began feeling comfortable speaking Spanish again at the age of 14.

The research assistants were trained by the author to calculate the raw scores following the WMLS-III Examiner Manual scoring protocol (Woodcock et al., 2017). Ten percent of the assessments (n=4) were randomly selected to complete inter-rater reliability. The randomly selected assessments were used to complete item-by-item analysis for each subtest of the WMLS-III (Woodcock, et al., 2017) for Spanish and English. The overall reliability was a Kappa value of .97, which meets the benchmark of .90 (McHugh, 2012). The percent agreement between the raters was 99.25%. After calculating inter-rater reliability, discrepancies in scoring between raters were identified and resolved by the two raters following the scoring protocol.

Analysis Plan

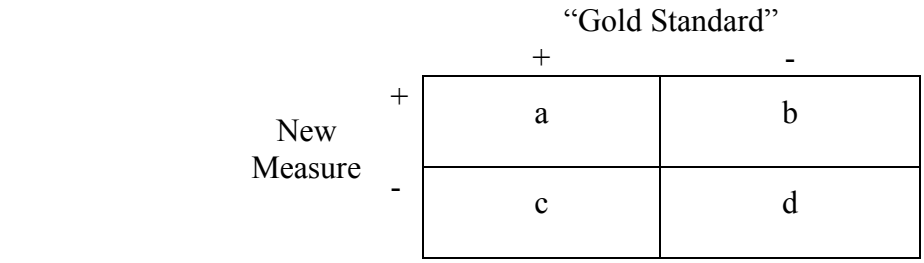
Four calculations are generally made in classification accuracy studies: (a) sensitivity, (b) specificity, (c) positive likelihood ratio, and (d) negative likelihood ratio to determine diagnosis of a disorder. In this study, classification accuracy was uniquely applied to determine proficiency levels. Sensitivity determines the screening measurement's accuracy in identifying individuals as proficient in comparison to the "gold standard" assessment. Specificity measures the screening's ability to accurately identify individuals who are not proficient in comparison to the "gold standard" (Dollaghan, 2007). Sensitivity and specificity can be ineffective due to their susceptibility to extraneous factors (e.g., base rate variations). Variations in the population that truly have the condition can easily affect sensitivity and specificity measurements which would lead to inaccurate results. For instance, samples with low base rates will always have sensitivity values lower than specificity. Likelihood ratios, on the contrary, are derived from both sensitivity and specificity and indicate the probability that the proficiency classifications of proficient and non-proficient are accurate.

Recall from Table 3 that proficiency categorization was determined for the current analysis plan. From the LUQ (Kiran et al., 2010), ratings of 1-4 were categorized as non-proficient; ratings of 5 were categorized as proficient (Kiran et al. 2010). The language domains of the LUQ (Kiran et al., 2010) that were utilized in classification analysis were the following: (a) overall ability, (b) speaking in casual conversations, (c) listening in casual conversations, (d) speaking in formal situations, (e) listening in formal situations, (f) reading, and (g) writing.

Proficiency levels of the WMLS-III (Woodcock et al., 2017) were calculated by following the scoring protocol from the Examiner's manual of the WMLS-III (Woodcock et al., 2017). The standardized scores and corresponding proficiency levels were calculated using the online scoring system associated with the WMLS-III (Woodcock et al., 2017). The online system then used the standardized scores to determine the proficiency level of the participant for each of the subtests of the assessment, each language domain assessed (i.e., listening, speaking, reading, and writing) and the participant's overall language proficiency for both Spanish and English. The three proficiency levels, (a) Initial Development, (b) Early Development, and (c) Continuing Development, were coded as non-proficient. The proficiency levels, (d) Emerging Proficient, (e) Proficient, and (f) Advanced Proficient, were coded as proficient (see Table 3).

The LUQ (Kiran et al., 2010) requires participants to self-rate their listening and speaking skills in *casual* and *formal* contexts. When completing the LUQ (Kiran et al., 2010), participants self-defined the terms *casual* and *formal* in the contexts of listening and speaking, which can vary between each participant (Gaffney, 2018), thus allowing for a more subjective result. The WMLS-III (Woodcock et al., 2017) does not assess *casual* and *formal* conversational abilities. Rather than obtaining a score for *casual* speaking and *formal* speaking in Spanish and English, the WMLS-III (Woodcock et al., 2017) provides one subtest cluster score for speaking in Spanish and one subtest cluster score for speaking in English. Due to possible variations in defining *casual* and *formal* from the LUQ (Kiran et al., 2010), responses to both informal and formal contexts must be analyzed in comparison to the respective language domains of the WMLS-III

(Woodcock et al., 2017). Therefore, self-reports of listening in *casual* and *formal* conversations in Spanish were each compared to the WMLS-III language subtest for listening in Spanish (Woodcock et al., 2017). The same procedure was followed for listening in *casual* and *formal* conversations in English, speaking in *casual* and *formal* conversations in Spanish, and speaking in *casual* and *formal* conversations in English.



a indicating true positive
b indicating false positive
c indicating false negative
d indicating true negative

Figure 1: 2x2 Table

The coded proficiency outcomes for participants across the language domains were applied to a 2x2 classification accuracy table to calculate sensitivity and specificity (see Figures 1 and 2).

$$Sensitivity = \frac{a}{a + c}$$

$$Specificity = \frac{d}{b + d}$$

Figure 2: Sensitivity and Specificity Formulas

The calculations of sensitivity and specificity were used to calculate the positive and negative likelihood ratios for each language domain in Spanish and in English from

the same 2x2 table (see Figure 3) and were evaluated using the likelihood ratio benchmarks shown in Table 1.

$$LR+ = \frac{sensitivity}{1 - specificity}$$

$$LR- = \frac{1 - sensitivity}{specificity}$$

Figure 3: Positive and Negative Likelihood Ratio Formulas

III. RESULTS

Forty likelihood ratios were calculated comparing the criterion-referenced self-report LUQ (Kiran et al., 2010) and the standardized norm-referenced WMLS-III (Woodcock et al., 2017): (a) language: Spanish and English, (b) context: formal and casual, and the five areas of interest (c) overall proficiency, (d) listening, (e) speaking, (f) reading, and (g) writing. Table 7 summarizes the results that had the greatest values indicating most suggestive in proficiency identification.

Table 7: Sensitivity, Specificity, and Likelihood Ratio Results

		Sensitivity	**Specificity**	**LR+**	**LR-**
Spanish	Overall	0.76	0.91	8.41	0.25
	Formal Speaking	0.6	0.93	8.7	0.43
	Formal Listening	0.56	0.85	3.89	0.52
English	Overall	0.74	--	--	--
	Formal Speaking	0.85	0.8	4.26	0.18

Two results, those for Spanish overall proficiency and Spanish formal speaking proficiency, confirm the research hypothesis because the likelihood ratios approach the "most likely" proficiency categories shown in Table 1. Specifically, the positive likelihood ratios for Spanish overall (LR+ = 8.41, LR- = 0.25) and Spanish speaking in formal situations (LR+ = 8.70, LR+ = 0.43) approach the "most likely proficient" range. The LR+ for Spanish overall proficiency and Spanish formal speaking proficiency indicate individuals are accurately self-identifying as proficient in Spanish. A LR- of 0.25 for Spanish overall proficiency is suggestive; so, the LUQ (Kiran et al., 2010) is one piece of information that can be used in addition to other measures to accurately identify people who are not proficient in Spanish as non-proficient in Spanish (Sackett et al., 1991; Sackett et al., 2000). Table 8 displays the classification accuracy categories

participants fell into for Spanish overall proficiency. The results indicate that 23 of the participants accurately identified level of proficiency with 6 participants underestimating proficiency (false negative) or overestimating proficiency (false positive).

Table 8: Spanish Overall Proficiency 2x2 Table

		WMLS-III Results		
		Proficient	Non-Proficient	Total
LUQ Results	Proficient	13	2	15
	Non-proficient	4	20	24
	Total	17	22	39

The LR- of 0.43 for Spanish formal speaking proficiency is suggestive that the participants accurately identified their non-proficiency in this domain. The remaining results shown in Table 7 fall well within the "suggestive but unable to determine proficiency" range.

As shown in Table 7, the likelihood ratios for overall proficiency in English were not calculated because all participants were identified as proficient based on their scores on the WMLS-III (Woodcock et al., 2017), so sensitivity and specificity are reported for this outcome instead. Sensitivity for overall proficiency in English was 0.74, falling below the acceptable 0.95 sensitivity requirement for this sample size (n=39) (Hajian-Tilaki, 2014). The sensitivity value indicates that 74% of individuals accurately self-identified as proficient in English when compared to performance on the WMLS-III (Woodcock et al., 2017).

Table 9: English Overall Proficiency 2x2 Table

		WMLS-III Results		
		Proficient	Non-proficient	Total
LUQ Results	Proficient	29	0	29
	Non-proficient	10	0	10
	Total	39	0	39

Twenty-nine of the 39 participants accurately self-reported proficiency in English when compared to the standardized norm-referenced assessment. The remaining ten participants did not self-identify as proficient in English but performed within the proficient range in English on the WMLS-III (Woodcock et al., 2017).

See Appendix A for the results of the likelihood ratios examining the language domains (speaking, listening, reading, and writing) and overall proficiency in Spanish and English. See Appendix B for all likelihood ratio charts from all language domains.

IV. DISCUSSION

In this study, the classification accuracy based on proficiency was examined for the LUQ (Kiran et al., 2010) using the WMLS-III (Woodcock et al., 2017) as the reference or "gold standard" test. Likelihood ratios, sensitivity and specificity calculations that are typically used in classification accuracy studies to make diagnoses for disorders were uniquely applied to proficiency. Previous research has not compared the likelihood ratios of proficiency self-reports to a standardized proficiency assessment; rather, the previous studies have only looked at the correlational relationship between the two proficiency measurements in determining self-accuracy. Mixed findings in the correlational relationships did not provide definitive results (Gaffney, 2018).

In this study, a majority of Spanish-English bilingual individuals correctly self-identified their proficiency in Spanish overall and in the area of formal speaking in Spanish, corroborating part of the hypothesis from Delgado et al. (1999). Delgado et al. (1999) found that participants who were skilled in *both* Spanish and English accurately assessed their overall proficiency and specific language skills (listening, speaking, reading, and writing) in Spanish. However, in English, participants only accurately assessed their reading and writing skills (Delgado et al., 1999). The Delgado et al. study (1999) and the current study both share similar participant self-report measurements in both Spanish and English for overall proficiency and within the language domains (i.e., speaking, listening, reading and writing). However, in the current study, participants additionally reported a *range* in their Spanish and English exposure and confidence, falling along various points of the bilingual continuum (Cummins, 2000). The additional language self-report information from participants in this study (i.e. confidence in

abilities) depicted a greater variance in bilingual individuals than was reported in Delgado et al. (1999).

Implications

The results from the current study make two important contributions to the literature. First, the outcomes will contribute to our growing professional knowledge in identifying bilingual practitioners. Second, the findings from this study contribute to our current knowledge in measuring language proficiency.

ASHA currently has practitioners self-identify their overall proficiency for the purpose of seeking to treat clients as a bilingual SLP. As the results of this study show, ASHA's method of determining proficiency of bilinguals for practice appears to have some validity. The standardized approach to assess a bilingual individual's proficiency may not be necessary as the outcomes from this study suggest; rather, the method of testing language proficiency by standardized assessment is arguably inefficient and not cost effective. The results of the current study suggests that the method ASHA currently uses to identify practitioners as bilingual is effective.

One distinguishing factor between this study and ASHA's current practice is the influence of monetary gain. The participants in this study were not financially motivated to rate themselves as more or less proficient for personal gain. Texas teachers are paid a higher salary and/or stipend if they are qualified as a bilingual educator. For instance, Austin Independent School District (Austin ISD) provides a $1500 stipend for first-year bilingual educators compared to their monolingual teacher counterparts (Austin ISD, 2018). ASHA members who are bilingual but are not considered "native-like" may be arguably motivated to self-report higher levels of proficiency due to a possible monetary

gain at a place of employment. Previous research has discussed the motivation of financial gain on accurate self-assessment, which was shown to have no effect on the phenomenon of individuals who inflate their skills on a self-report (Trofimovich, 2016). However, financial gain cannot be ruled out as a non-contributing bias; financial gain may affect one's motivation to overestimate their language proficiency in the field of speech-language pathology.

Another differentiating factor between this study and that of ASHA's current practice is the amount of insight into an individual's proficiency. ASHA currently has bilingual practitioners self-report their overall proficiency in consideration of the areas of "vocabulary, word-meaning, phonology, grammar, and pragmatics" (Cornish, 2011, p. 16). This global outlook on proficiency does not fully capture the possible variation in language abilities across the domains of speaking, listening, reading and writing. In this study, the language domains were additionally considered. The highly likely results in overall proficiency and the suggestive findings in formal speaking in Spanish with suggestive findings in formal listening in Spanish and overall proficiency and formal speaking in English indicate that individuals self-identify proficiency overall and within certain language domains. These results suggest that ASHA should possibly consider incorporating language domains into the self-report measurement for bilingual practitioners.

As mentioned previously, the findings from this study will have an impact on our current growing knowledge of language proficiency. Researchers have defined proficiency by various dimensions including: native-like (or near native-like) proficiency (Hyltenstam & Abrahamsson, 2000; Morgan-Short et al., 2012; White & Genesee, 1996),

conversational proficiency, academic proficiency (Aukerman, 2007; Cummins, 2000), and L2 proficiency (Luo, Luk, & Bialystok, 2010). Recall from the literature review that the LUQ and the WMLS-III have different definitions of proficiency (Kiran et al., 2010; Woodcock et al., 2017). The LUQ (Kiran et al., 2010) has individuals determining "native-like" proficiency as the highest rating while the WMLS-III (Woodcock et al., 2017) assesses an individual's academic proficiency.

As argued by Tremblay (2011), most standardized proficiency assessments examine proficiency for an educational application and do not assess the other forms of proficiency. This could possibly lead to identifying an individual as being non-proficient for academic proficiency and possibly fall short in identifying their proficiency in comparison to other definitions, such as native-like proficiency. The WMLS-III (Woodcock, et al., 2017) is an example of a standardized assessment that examines proficiency for academic purposes without regard to varying proficiency definitions and the influence of proficiency in various environments. Participants that self-reported proficiency as native-like but were determined to be non-proficient by academic standards, may have native-like proficiency that could not be determined by the standardized norm-referenced academic assessment.

The mismatch in proficiency definitions (i.e., native-like and academic) between the self-report and the standardized assessment is a factor that needs to be considered for research and clinical use. As Cummins (2000) argues, there is a distinct difference between native-like proficiency and academic proficiency. Native-like proficiency includes language dimensions such as "conversational fluency and pronunciation" (p. 53) that are mastered early in life; whereas academic proficiency entails knowledge in "low

frequency vocabulary, complex grammatical structures, and greater demands on memory, analysis, and other cognitive processes" (p. 36) that requires years of special instruction to master (Cummins, 2000). The current requirement by ASHA has practitioners self-identify as either near native-like or native-like in their L2 proficiency (ASHA, 2018b). ASHA needs to clearly define the standards for near native and native-like proficiency to ensure ethical standards are being met by bilingual practitioners.

Limitations

There are three major limitations to this study. First, this study was not conducted with practicing bilingual SLPs. This study was designed to answer a question regarding the accuracy of ASHA's current practice in identifying bilingual SLPs but instead, undergraduate and graduate students—non-practitioners—were utilized as the participants. Since these participants are not currently practicing professionals, they may not have had the insight practicing SLPs may have when identifying proficiency for practice.

Second, the WMLS-III's (Woodcock, et al., 2017) norming population had a wide age range (up to 90 years in age), but the standardized scores were normed for 3 years, 0 months to 22 years, 11 months (Woodcock et al., 2017). The test manual did not contain an explanation of how participants who were outliers to the age limit were entered into the online scoring system in order to calculate the standardized proficiency scores for the current analysis. In this study, participants' ages were modified to match the age limit of the WMLS-III (Woodcock et al., 2017) scoring system.

Lastly, the number of male to female participants in this study was a limitation. Out of the total participants, only 10% (n=4) were male. Furnham (2001) described

various studies in the literature that have examined the effects of gender upon self-reports in intelligence domains. The literature found that males would generally report themselves to have a higher intelligence compared to female counterparts and females would underestimate their intelligence (Furnham, 2001). Furnham (2001) concludes possible gender bias in self-report measurements that were unable to be considered in this study. Gender influences were variable, but it is a factor that could affect how individuals report their abilities and warrants consideration in the future.

Future Directions

There are many considerations for how proficiency will be assessed and discussed in future research. First, the study needs to be replicated with practicing bilingual SLPs. A factor that is not discussed often when identifying bilingual practitioners is financial motivation. Bilingual practitioners can receive a financial gain in practice compared to their monolingual counterparts. As previously discussed, financial gains in this study were not a contributing variable, therefore this motivation can be ruled out. However, when self-identifying as bilingual for ASHA or for any other profession seeking bilingual individuals, there is a financial incentive. In consideration of the different professional populations that will receive monetary incentives for being bilingual (e.g., speech-language pathologist, teacher), it is necessary to study the effects of financial incentive on self-report measures of proficiency.

Second, the study should be replicated with modified factors to determine influence upon the results. For instance, the classification accuracy of this study with likelihood ratios can be completed with a more diverse group of participants (e.g., gender, age, educational background). The current study included participants with

various bilingual abilities. However, there are many other variables that warrant consideration, such as majority of participants being female in this study in comparison to a participant pool with equal gender representations (e.g., Furnham, 2001). Furthermore, standardized norm-referenced tests other than the WMLS-III (Woodcock et al., 2017) could be used as the reference or "gold standard" test for future classification accuracy studies. By using the same procedure and data analysis of the current study with these suggested modifications, future studies can additionally address self-report biases, such as over and under estimation.

Furthermore, other self-report measurements should be considered for future classification accuracy analysis in determining an individual's proficiency. One self-report method that has been used in the field of speech-language pathology is direct magnitude estimation scaling (DME) (e.g., Weismer & Laures, 2002). DME has been advocated for use in linguistic research (Sorace, 2010), and is an option for which validation studies should be completed in the future. DME has individuals create their own scale for judgement or self-reported tasks. DME is a more sensitive measure to "gradience in [syntactic] acceptability judgements" (p. 67) and is not restricted to "an n-point rating scale with 'anchored' extremes" (p.59), like a Likert scale (Sorace, 2010). Arguably, DME may be appropriate in developing a language questionnaire for individuals to use in identifying their language proficiency. A strong level of awareness may not be required to complete DME since there are no restrictions on the number of levels assigned to degree of proficiency. These aspects of DME should be considered in a future classification accuracy study that may or may not compare to the LUQ (Kiran et al., 2010) rating as well.

Finally, accent perception of the L2 needs to be addressed as a possible bias in the future development of self-report measurements. As discussed earlier, Cummins (2000) identified that proficiency can be measured by different language variables (i.e., current language use, age of first exposure, listening, speaking, reading, and writing) but did not include accent perception. Trofimovich (2016) analyzed how perceptions of accent influence pronunciation and comprehensibility using self-reports and listener judgments. That study found that L2 speakers were most likely to be inaccurate in self-assessment of their accent and how comprehensible they believed themselves to be (Trofimovich, 2016). ASHA considers accents/dialects to be a cultural difference. Considering the perception of accent/dialect, a bias in speaking ability and overall proficiency from an accent needs to be considered when asking individuals to identify their proficiency for future research.

Conclusion

Spanish-English bilingual individuals are able to self-identify their overall proficiency in Spanish and speaking in formal situations in Spanish. These findings suggest that the method of self-report that ASHA currently uses to identify bilingual practitioners may be valid, which is essential in ensuring an ethical workforce. In addition, these findings contribute to the growing knowledge of bilingualism and the need to continue further exploration into proficiency standards.

APPENDIX SECTION

APPENDIX A

		LUQ Proficient (n)	WMLS-III Proficient (n)	Sensitivity	Specificity	LR +	LR -
Spanish	Informal Listening	22	18	0.77	0.62	2.04	0.35
	Formal Listening	13	18	0.56	0.85	3.89	0.52
	Informal Speaking	15	10	0.8	0.75	3.31	0.26
	Formal Speaking	8	10	0.6	0.93	8.7	0.43
	Reading	13	20	0.65	1	-	0.35
	Writing	10	16	0.76	0.9	3.35	0.64
English	Informal Listening	37	37	0.94	0	0.94	-
	Formal Listening	34	37	0.89	0.5	1.78	0.21
	Informal Speaking	33	34	0.91	0.6	2.27	0.14
	Formal Speaking	30	34	0. 85	0.8	4.26	0.18
	Reading	35	39	0.89	-	-	-
	Writing	33	39	0.84	-	-	-

APPENDIX B

Table 10: Informal Listening in Spanish

		WMLS-III Results		
		Proficient	Non-proficient	Total
LUQ Results	Proficient	14	8	22
	Non-proficient	4	13	17
	Total	18	21	39

Table 11: Informal Listening in English

		WMLS-III Results		
		Proficient	Non-proficient	Total
LUQ Results	Proficient	35	2	37
	Non-proficient	2	0	2
	Total	37	2	39

Table 12: Formal Listening in Spanish

		WMLS-III Results		
		Proficient	Non-proficient	Total
LUQ Results	Proficient	10	3	13
	Non-proficient	8	18	26
	Total	18	21	39

Table 13: Formal Listening in English

		WMLS-III Results		
		Proficient	Non-proficient	Total
LUQ Results	Proficient	33	1	34
	Non-proficient	4	1	5
	Total	37	2	39

Table 14: Informal Speaking in Spanish

		WMLS-III Results		
		Proficient	Non-proficient	Total
LUQ Results	Proficient	8	7	15
	Non-proficient	2	22	24
	Total	10	29	39

Table 15: Informal Speaking in English

		WMLS-III Results		
		Proficient	Non-proficient	Total
LUQ Results	Proficient	31	2	33
	Non-proficient	3	3	6
	Total	34	5	39

Table 16: Formal Speaking in Spanish

		WMLS-III Results		
		Proficient	Non-proficient	Total
LUQ Results	Proficient	6	2	8
	Non-proficient	4	27	31
	Total	10	29	39

Table 17: Formal Speaking in English

		WMLS-III Results		
		Proficient	Non-proficient	Total
LUQ Results	Proficient	29	1	30
	Non-proficient	5	4	9
	Total	34	5	39

Table 18: Reading in Spanish

		WMLS-III Results		
		Proficient	Non-proficient	Total
LUQ Results	Proficient	13	0	13
	Non-proficient	7	19	26
	Total	20	19	39

Table 19: Reading in English

		WMLS-III Results		
		Proficient	Non-proficient	Total
LUQ Results	Proficient	35	0	35
	Non-proficient	4	0	4
	Total	39	0	39

Table 20: Writing in Spanish

		WMLS-III Results		
		Proficient	Non-proficient	Total
LUQ Results	Proficient	7	3	10
	Non-proficient	9	20	29
	Total	16	23	39

Table 21: Writing in English

		WMLS-III Results		
		Proficient	Non-proficient	Total
LUQ Results	Proficient	33	0	33
	Non-proficient	6	0	6
	Total	39	0	39

Table 22: Overall Proficiency in Spanish

		WMLS-III Results		
		Proficient	Non-proficient	Total
LUQ Results	Proficient	13	2	15
	Non-proficient	4	20	24
	Total	17	22	39

Table 23: Overall Proficiency in English

		WMLS-III Results		
		Proficient	Non-proficient	Total
LUQ Results	Proficient	29	0	29
	Non-proficient	10	0	10
	Total	39	0	39

REFERENCES

American Speech-Language-Hearing Association [ASHA]. (2018a). Demographic profile of ASHA members providing bilingual services, year-end 2018 [PDF File]. Retrieved from https://www.asha.org/uploadedFiles/Demographic-Profile-Bilingual-Spanish-Service-Members.pdf

American Speech-Language-Hearing Association. (2018b). *Bilingual service delivery: Key issues.* Retrieved from https://www.asha.org/PRPSpecificTopic.aspx?folderid=8589935225&section=Key_Issues#Bilingual_Service_Providers

Arroyo-Romano, J. E. (2016). Bilingual education candidates' challenges meeting the Spanish language/bilingual certification exam and the impact on teacher shortages in the state of Texas, USA. *Journal of Latinos and Education, 15*(4), 275-286.

Aukerman, M. (2007). A culpable CALP: Rethinking the conversational/academic language proficiency distinction in early literacy instruction. *The Reading Teacher, 60*(7), 626-635.

Austin Independent School District [Austin ISD]. (2018). Austin ISD 2018-19 hiring salary schedule [PDF File]. Retrieved from https://www.austinisd.org/sites/default/files/dept/hr/docs/2018-19%20Teacher-Librarian%20Salary%20Schedule.pdf

Bialystok, E., & Craik, F. (2010). Cognitive and linguistic processing in the bilingual mind. *Current Directions in Psychological Science, 19*(1), 19-23.

Bujang, M.A., & Adnan, T. H. (2016). Requirements for minimum sample size for sensitivity and specificity analysis. *Journal of Clinical and Diagnostic Research, 10*(10), 1-6.

Cornish, N. (2011). What it takes to call yourself a bilingual practitioner. *The ASHA Leader, 16*, 16-18.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge, UK: Cambridge University Press & Council of Europe.

Cummins, J. (2000). Language, power, and pedagogy: Bilingual children in the crossfire. Clevedon [England]; Tanawand [N.Y.]: Multilingual Matters.

De Houwer, A. (2017). Bilingual language acquisition. In Fletcher, P. & MacWhinney, B. (Eds.), The Handbook of Child Language (221-250). Oxford, UK: Blackwell Publishing Ltd.

Delgado, P., Guerrero, G., Goggin, J.P., & Ellis, B. B. (1999). Self-assessment of linguistic skills by bilingual Hispanics. *Hispanic Journal of Behavioral Sciences, 21*, 31-46.

Dollaghan, C. A. (2007). The handbook for evidence-based practice in communication disorders. Baltimore, MD: Paul H. Brookes Publishing Co.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test* (3rd ed.). Circle Pines, MN: AGS.

Dunn, L. M., Padilla, E. R., Lugo, D. E., & Dunn, L. M. (1986). *Test de Vocabulario en Imágenes Peabody* (Adaptación Hispanoamericana) [Peabody Picture Vocabulary Test Hispanic American version]. Circle Pines, MN: AGS.

Furnham, A. (2001). Self-estimates of intelligence: Culture and gender in difference in self and other estimates of both general (g) and multiple intelligences. *Personality and Individual Differences, 31*, 1381-1405.

Gaffney, C. (2018). Understanding the causes of inaccurate self-assessments: Extraversion's role. *Proceedings of the 42nd annual Boston University Conference on Language Development*.

Gasquoine, P. H., Croyle, K. L., Cavazos-Gonzalez, C., & Sandoval, O. (2007). Language of administration and neuropsychological test performance in neurologically intact Hispanic American bilingual adults. *Archives of Clinical Neuropsychology, 22*(8), 991-1001.

Gollan, T., Wissberger, G., Runnqvist, E., Montoya, R., & Cera, C. (2011). Self-ratings of spoken language dominance: A multilingual naming test (MINT) and preliminary norms for young and aging Spanish-English bilinguals. *Bilingualism: Language and Cognition, 15*(3), 594-615.

Gray, T. (2017). Bilingual aphasia: An intervention roadmap and the dynamic interplay between lexical access and language control. *Perspectives of the ASHA Special Interest Groups, 2*, 15-22.

Gray, T. & Kiran, S. (2012). Bilingual aphasia: What is the role of proficiency and impairment? Paper presented at the American Speech-Language-Hearing Association 2012 Annual Convention, Atlanta, GA.

Hajian-Tilaki, K. (2014). Sample size estimation in diagnostic test studies of biomedical informatics. *Journal of Biomedical Informatics, 48*, 193-204.

Holliday, S. L., Navarrete, M. G., Hermosillo-Romo, D., Valdez, C. R., Saklad, A. R., Escalante, A., & Brey, R. L. (2003). Validating a computerized neuropsychological test battery for mixed ethnic lupus patients. *Lupus, 12*(9), 697-703.

Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research.* Amsterdam: John Benjamins.

Hyltenstam, K. & Abrahamsson, N. (2000). Who can become native-like in a second language? All, some, or none? *Studia Linguistica, 54*(2), 150-166.

Kastenbaum, J. G., Bedore, L. M., Peña, E. D., Sheng, L., Mavis, I., Sebastian-Vaytadden, R., Rangamani, G., Vallila-Rohter, S., & Swathi, K., (2018). The influence of proficiency and language combination on bilingual lexical access. *Bilingualism: Language and Cognition,* 1-31.

Kiran, S. & Iakupova, R. (2011). Understanding the relationship between language proficiency, language impairment and rehabilitation: Evidence from a case study. *Clinical Linguistics & Phonetics, 25*(6-7), 565-583.

Kiran, S., Peña, E., Bedore, L., & Sheng, L. (2010). Evaluating the relationship between category generation and language use and proficiency. Paper presented at the Donostia Workshop on Neurobilingualism, San Sebastian, Spain.

Luk, G. & Bialystok, E. (2013). Bilingualism is not a categorical variable: Interaction between language proficiency and usage. *Journal of Cognitive Psychology, 25*(5), 605-621.

Luo, L., Luk, G., & Bialystok, E. (2010). Effect of language proficiency and executive control on verbal fluency performance in bilinguals. *Cognition, 114*, 29-41.

Mahon, E. A. (2006). High-stakes testing and English language learners: Questions of validity. *Bilingual Research Journal, 30*(2), 479-497.

Marian, V., Blumenfeld, H., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research, 50*, 940-967.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276-282.

Morgan-Short, K., Steinhauer, K., Sanz, C., & Ullman, M. T. (2012). Explicit and implicit second language training differentially affect the achievement of native-like brain activation patterns. *Journal of Cognitive Neuroscience, 24*(4), 933-947.

Muñoz-Sandoval, A. F., Cummins, J., Alvarado, C. G. & Ruef, M. L. (2005). *Bilingual Verbal Abilities Test Normative Update* (BVAT-NU). Retrieved from https://www.hmhco.com/shop/k12/Bilingual-Verbal-Ability-Tests-BVAT-Normative-Update/id/920967

Muñoz-Sandoval, A. F., Woodcock, R. F., McGrew, K. S., & Mather, N. (2005). *Pruebas de aprovechamiento* [Tests of achievement] (Batería III, Woodcock-Muñoz). Itasca, IL; Riverside.

Perani, D., Abutalebi, J., Paulesu, E., Brambati, S., Scifo, P., Cappa, S. F., & Fazio, F. (2003). The role of age of acquisition and language usage in early, high-proficient bilinguals: An fMRI study during verbal fluency. *Human Brain Mapping, 19*(3), 170-182.

Pray, L. (2005). How well do commonly used language instruments measure English oral-language proficiency? *Bilingual Research Journal, 29*(2), 387-409.

Ryan, C. (2013). Language use in the United States: 2011 [PDF File]. Retrieved from https://www2.census.gov/library/publications/2013/acs/acs-22/acs-22.pdf

Sackett, D. L., Haynes, R. B., Guyatt, G. H., & Tugwell, P. (1991). *Clinical epidemiology: A basic science for clinical medicine.* Boston: Little, Brown.

Sackett, D. L., Straus, S. E. Richardson, W. S., Rosenberg, W., & Haynes, R. B. (2000). *Evidence-based medicine: How to practice and teach EBM.* Edinburgh, Scotland: Churchill Livingstone.

Schwarz, A. L. Resendiz, M., Gonzales, M. D., Gragera, A., Tipps, J., & Perez, C. (in preparation). Translation practices of bilingual speech-language pathologists with storybooks.

Sorace, A. (2010). Using magnitude estimation in developmental linguistic research. In Blom, E. & Unsworth, S. Editor (Eds.), *Experimental Methods in Language Acquisition Research*, (57-72). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Texas Educator Certification Examination Program. (2019). Bilingual target language proficiency test (BTLPT) (190). Retrieved from: https://www.tx.nesinc.com/TestView.aspx?f=HTML_FRAG/TX190_TestPage.html

Texas Statewide Leadership for Autism Training [TSLAT]. (2015). Academic achievement assessment: Bilingual verbal ability tests (BVAT-NU) [PDF File]. Retrieved from http://txautism.net/assets/uploads/docs/BVAT-NU-TG.pdf

Tremblay, A. (2011). Proficiency assessment standards in second language acquisition research. *Studies in Second Language Acquisition, 33*, 339-372.

Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K., & Crowther, D. (2016). Flawed self-assessment: Investigating self- and other-perception of second language speech. *Bilingualism: Language and Cognition, 19*(1), 122-140.

Valdés, G., & Figueroa, R. (1994). Bilingualism and testing: A special case of bias. Norwood, NJ: Ablex Publishing.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. *Handbook of Statistics, 26*, 81-124.

Weismer, G. & Laures, J. S. (2002). Direct magnitude estimates of speech intelligibility in dysarthria. *Journal of Speech, Language, and Hearing Research, 45*, 421-433.

White, L. & Genesee, F. (1996). How native is near-native? The issue of ultimate attainment in adult second language acquisition. *Second Language Research, 12*(3), 233-265.

Woodcock, R.W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson Tests of Cognitive Abilities* (3rd ed.). Itasca, IL: Riverside.

Woodcock, R.W., Alvarado, C. G., & Ruef, M. L. (2017). Woodcock-Muñoz Language Survey III. Itasca, IL: Houghton Mifflin Harcourt.