St. John's University
St. John's Scholar

Theses and Dissertations

2023

ON THE IMPORTANCE OF TRUE PEER NORMS IN THE ASSESSMENT OF ENGLISH LEARNERS: A VALIDATION STUDY OF THE ORTIZ PICTURE VOCABULARY ACQUISITION TEST

Jane Yan Ting Wong

Follow this and additional works at: https://scholar.stjohns.edu/theses_dissertations

 Part of the Psychology Commons

ON THE IMPORTANCE OF TRUE PEER NORMS IN THE ASSESSMENT OF ENGLISH LEARNERS: A VALIDATION STUDY OF THE ORTIZ PICTURE VOCABULARY ACQUISITION TEST

A dissertation submitted in partial fulfillment
of the requirements for the degree of

DOCTOR OF PSYCHOLOGY

to the faculty of the

DEPARTMENT OF PSYCHOLGY

of

ST. JOHN'S COLLEGE OF LIBERAL ARTS AND SCIENCES
at
ST. JOHN'S UNIVERSITY
New York
by

Jane Yan Ting Wong

Date Submitted: 4/16/2023 Date Approved: 4/18/2023

Jane Yan Ting Wong

Samuel O. Ortiz, Ph.D.

© Copyright by Jane Yan Ting Wong 2023

All Rights Reserved

ABSTRACT

ON THE IMPORTANCE OF TRUE PEER NORMS IN THE ASSESSMENT OF ENGLISH LEARNERS: A VALIDATION STUDY OF THE ORTIZ PICTURE VOCABULARY ACQUISITION TEST

Jane Yan Ting Wong

Traditional efforts in aiding English learners (ELs) to achieve better test performance such as modifications to the testing process or the use of native language or nonverbal tests are problematic and disregard the unique language developmental experiences of ELs (Ortiz & Wong, 2022). The Ortiz Picture Vocabulary Acquisition Test (Ortiz PVAT; Ortiz, 2018), an assessment of English receptive vocabulary, incorporates ELs' proportion of lifetime exposure to English (LEE) in test norms to allow for true peer comparison, thus ensuring test fairness in measurement and score interpretation. The current study aimed to add to the existing validity evidence for the test and to provide support for the use of true peer norms which account for LEE when assessing ELs' English receptive vocabulary development in an archival dataset comprising a sample of ELs from the New York City metropolitan area. Results indicate that performance on the Ortiz PVAT was not affected by gender nor home language spoken, suggesting that the Ortiz PVAT measures receptive vocabulary in English in a fair manner, irrespective of one's gender or heritage language. LEE significantly correlated with receptive vocabulary, such that higher LEE was associated with better performance on the Ortiz PVAT. Furthermore, significant differences in standard scores based on the English Speaker (ES) norms with a medium effect size was found between

the groups of EL with low or medium levels of LEE and the ES sample but not when their test performance was compared to the EL normative sample when LEE is accounted for. Lastly, an additional 18% of the variance in Ortiz PVAT raw scores was accounted for by LEE above and beyond age, and LEE was found to exert more influence on the variance in raw scores compared to age. Results from the current study provide further support for the existing validity evidence for the Ortiz PVAT and contribute to the knowledge base regarding test fairness for ELs, specifically regarding true peer comparison with LEE accounted for in the valid and defensible evaluation of ELs' language abilities.

ACKNOWLEDGEMENTS

First and foremost, I would like to extend my sincere gratitude to my mentor, Samuel O. Ortiz, Ph.D., for his invaluable tutelage, guidance, and support since 2011 when we collaborated on the development of the Ortiz Picture Vocabulary Acquisition Test (Ortiz PVAT). The astute Dr. Ortiz knew within the first year we met that I would find success in the field of school psychology because of my unique skillset and experiences, while it took me more than six years (almost the entire duration of the development of the test) to finally realize what had been missing in my former career—the amalgamation of my passion in helping others, my knowledge in psychometrics, and my personal experience as an English learner navigating between two very distinct languages and cultures since my teenage years. I am beyond grateful for the ability to wake up every morning doing what I love. This dissertation, a validation study of the test I developed with Dr. Ortiz, is dedicated to him and all English learners.

Secondly, I would like to thank my Dissertation Committee, Dawn P. Flanagan, Ph.D. and Marlene Sotelo-Dynega, Psy.D. for their brilliant comments and guidance on my dissertation. I attribute my solid foundation in the area of assessment and CHC theory to these two inspiring professors. I am also thankful to all the professors at St. John's University as well as all the supervisors in my practicum/externships/internship who have helped shape me as a competent researcher/practitioner. Special thanks to my former mentor, Richard Lalonde, Ph.D. at York University, Canada, who taught me how to do research and use SPSS more than a decade ago.

I am deeply grateful to my parents and my uncle, Dr. Nelson Wong, for their constant support, encouragement, and keen interest in my academic achievement. Finally,

I would like to thank all my friends, including my entire Ortiz PVAT development team at MHS, for their moral support throughout my doctoral journey.

As part of "Jane's graduate school survival kit," a former MHS colleague shared this poem by Ralph Waldo Emerson on *What is Success* with me, and it has been my north star in life ever since:

> To laugh often and much;
> To win the respect of intelligent people and the affection of children;
> To earn the appreciation of honest critics and endure the betrayal of false friends;
> To appreciate beauty;
> To find the best in others;
> To leave the world a bit better, whether by a healthy child, a garden patch or a redeemed social condition;
> To know even one life has breathed easier because you have lived;
> This is to have succeeded.

With this research, my clinical work, and the amazing connections I have built over the years, I am inclined to say I have succeeded.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

Statement of the Problem

According to the National Center for Education Statistics (NCES), the percentage of students in public schools in the United States (U.S.) who were identified as English learners[1] (ELs) increased from 9.2% (4.5 million students) from fall 2010 to 10.4% (5.1 million students) in fall 2019, with 12 states having 10% or more ELs in their total public school population (2022). These states include Alaska (12.0%), California (18.6%), Colorado (11.0%), Delaware (11.1%), Illinois (12.3%), Maryland (10.6%), Massachusetts (10.6%), Nevada (14.5%), New Mexico (16.5%), Rhode Island (12.2%), Texas (19.6%), and Washington (11.7%; NCES, 2022). In particular, Spanish was found to be the home language of 3.9 million EL public school students in fall 2019. This language group represents 75.7% of all EL students and 7.9% of all public school students, followed by other languages such as Arabic (spoken by 131,600 students), Chinese (100,100 students), and Vietnamese (75,558 students). Furthermore, 792,000 ELs were identified as students with disabilities in fall 2019. This number represents 15.5% of the total EL student enrollment, compared to the 14.4% of students with disabilities within the total public school enrollment in 2019–2020 (NCES, 2022).

As the population of ELs in K–12 continue to grow, there are increasing demands for school psychologists to conduct psychoeducational assessments to determine eligibility for special education services as well as appropriate school placements for ELs (Ortiz et al., 2018). There has long been an issue of disproportionality or

[1] Adapted from Ortiz and Wong (2022), the term *English learner* (or its attendant acronym, EL) will be used in the present paper to represent individuals who are neither monolingual, nor native English speakers, regardless of the number of languages one has been exposed to, the age at which the learning of English began, or the number of years of formal instruction in English.

overrepresentation of minority and EL students in special education, and researchers have attributed educators' attitudes and expectations as well as the practice of testing as contributing factors (Ford, 2012; Cormier et al., 2014; Cormier et al., 2022). Yet few guidelines or consensus exist in terms of standards for evaluating ELs whose language development trajectory differ greatly from that of native English speakers (ESs), let alone the diverse experiences that exist among ELs from different cultural or linguistic backgrounds which adds to the complexity of test fairness (Ortiz, 2014; Ortiz, 2018; Ortiz et al., 2018; Cormier et al., 2022; Ortiz & Wong, 2020a, Ortiz & Wong, 2022). A quote by Valdes and Figueroa (1994) illustrates the dire state within the field of school psychology:

> The unique American tragedy of bilinguals has been that over the last century, both test makers and testers have generally ignored the psychological robustness of bilingualism. The result has been a waste of human potential. Bilingual persons have needlessly been misled and misdiagnosed, especially children. (p. 87)

The current study aims to explore the concept of test fairness as it relates to the assessment of ELs' language abilities, with a focus on the importance of using true peer comparison to account for ELs' differential linguistic developmental experiences when evaluating test performance.

Chapter I

Literature Review

Threats to Test Fairness in the Assessment of English Learners

The assessment of cognitive abilities, academic achievement, and other neuropsychological domains is heavily predicated upon the use of standardized psychometric instruments. Standardized tests of cognitive and language functioning are validated for use with native English-speaking normative samples with controls for cognitive maturation (i.e., age) along with other demographic variables that have been theorized to covary with the construct that a particular test is intended to measure (e.g., gender, parental education level [PEL] as a proxy for socio-economic status [SES], and geographical region). Using age to control for differences in cognitive development for ESs is valid because every monolingual native ES begins learning English at the same time (from birth), and any significant deviation from their same age peers in levels of performance is likely attributable to individual differences in the ability being measured, provided that the test was given in the same manner and/or conditions as the normative sample (i.e., following standardization).

However, for ELs, although they have also been learning their native language (L1) since birth, the acquisition of the English language, their second language (L2), can begin at any point after birth. As such, their language development in English cannot be assumed to follow the same pattern as a typical ES. Age alone cannot account for the variable of experience and exposure to English which can impact ELs' test performance when ELs are tested in L2 (Rhodes et al., 2005; Ortiz, 2014; Ortiz et al., 2018; Ortiz & Wong, 2022). In other words, the ability that is being measured in a test given in English

is confounded with an examinee's exposure to English due to linguistic demands in the tests in terms of the ability to comprehend task instructions and/or to provide appropriate verbal responses. When an EL does not perform at the expected level compared to same age peers in the normative sample, there is no way to ascertain whether the below average performance is due to an inherently lower ability in the construct being measured or the different amounts of exposure to or opportunities in learning English.

It is important to note also that language development is not the only way in which an EL's test performance can be affected. Language and culture are inextricably linked as culture informs the use of language and vice versa (Vygotsky, 1986). The same concept can be communicated differently due to regional and cultural variations even when the same language is used. For instance, napkin refers to a piece of garment used at a table to wipe the lips or protect the clothes in the U.S., but it can also mean diaper in British English (Merriam-Webster, n.d.). Thus, performance on standardized tests developed in the North American context is also affected by one's level of acculturation or knowledge of the culture from which the tests are based (Rhodes et al., 2005). Researchers such as Ortiz and Flanagan (1998), Flanagan and Ortiz (2001), Rhodes et al. (2005) and Ortiz et al. (2017) refer to the concepts of linguistic demand and cultural loading as the two dimensions that can impact ELs' test performance. Because the variable of exposure to language and/or culture differs between ELs and ESs, and between individual ELs with differing language and cultural experiences, the use of age alone to control for experiential differences and cognitive maturation is no longer sufficient to maintain normative comparability for ELs. In essence, when the same approach for assessing ESs is applied to the assessment of ELs, fairness is compromised.

Even when an ES and an EL are of the same age, they cannot be validly presumed to share comparable experiences, exposure, or development in English and the North American culture from which many cognitive tests have been developed. Even for minority students who have been exposed to both L1 and L2 simultaneously from birth and thus become functionally fluent in both languages, their language development is still qualitatively different from monolingual ES, rendering the comparison to monolingual ES normative samples invalid. Hence, ELs will always be at a disadvantage and at greater risk for misidentification of disability when they are assessed with the use of English-based (L2) tests with norms that are based on native ESs with the assumption that every examinee has had similar experience and exposure to the language the test is administered in and the culture from which the test is based as the normative sample.

Despite issues outlined above, school psychologists are still tasked with conducting evaluations for ELs as there is not an option to do nothing. A number of alternative approaches to testing have been proposed in an effort address the issue of test fairness and threats to test score validity in evaluating ELs. However, each of the approaches frequently used in the field is associated with flaws that do not fully address the problem they are intended to solve. Modifications to the testing procedure to compensate for test fairness in the assessment of ELs include the use of an interpreter or translator and testing the limits. Others have relied on the use of nonverbal or native language tests or testing in the EL's dominant language. A brief discussion of the pros and cons for each method is as follows.

Existing Solutions to Increase Test Fairness

Use of Interpreter or Translator

This approach includes the use of a translator or interpreter for test administration. While the use of a translator/interpreter likely leads to better understanding of task instructions and allows for responses given in an EL's native language, it clearly violates of standardization protocol of a test, rendering the test scores derived from such administration invalid and uninterpretable (Ortiz & Wong, 2020a). Unless the tests are developed to be given by a translator/interpreter with standardized task instructions and scoring rubrics in the specific language (and dialect) spoken by both the interpreter and EL, along with adequate control for differences in language and cultural development among individuals in the norm sample, the use of a translator/interpreter will inevitably continue to represent a threat to test score validity. Another issue concerning the use of interpreter/translator is that many of whom are not well trained or well versed in the administration of psychoeducational evaluation. There is no way for the examiner to verify if the items were given in the way that were intended by the test developer, and if the items responses translated to English match the exact wording or intended meaning by the examinee. Not to mention the complexities involved in variations in dialects and regional usage of many native languages. Even when the translator/interpreter is highly trained and experienced, there is no manner in which one can determine to what degree the use of interpreter/translator has hindered or helped the examinee and to what extent. The psychometric properties of a test administered with the use of a translator/interpreter is unknown. It is impossible to establish the mean, standard deviation, reliability, and validity of any test that is administered in a way that differs from standardized procedure.

Thus, it is erroneous to assume that scores derived from such test procedure to be equal to scores derived from the administration procedures or scoring protocols originally recommended by the test publisher. At the same time, the use of translator/interpreter can provide useful qualitative information regarding an EL's ability levels as well as instructional needs. The approach only becomes problematic when the test scores are erroneously treated as valid and are used for diagnostic or eligibility decision-making (Flanagan et al., 2013; Ortiz & Wong, 2020a).

Testing the Limits

In recognition of the inherent disadvantage for EL in traditional testing procedures, a variety of methods have been used to help examinees to perform to the best of their ability. Such approaches have generally been referred to as "testing the limits" (Flanagan et al., 2013). These test methods include alterations and modifications of test items, permitting responses provided in the examinee's native language, repeating task instructions (even when it is not allowed as per test instructions), tutorials of task concepts given prior to actual administration, extension and/or complete removal of time constraints for timed tasks, etc. Despite the good intention of "leveling the playing field" for ELs, such practices nonetheless violate the standardization, rendering the test scores invalid and uninterpretable. Even when modifications are permitted by the test publisher, there exists no norms for valid comparisons. However, similar to the use of translator/interpreter, qualitative information about the examinee's performance derived from testing the limits can be useful to inform instructional needs.

Nonverbal or Language-Reduced Testing

Another approach to address fairness issues when testing ELs is to avoid the use of language altogether by engaging in "nonverbal" methods of testing. Although the impact of language differences on ELs' test performance may be reduced when the required use of receptive and expressive language in English is lower in measures of nonverbal abilities, it is virtually impossible to actually administer any test without some type of communication occurring between the examiner and the examinee (Ortiz & Wong, 2022). Even when illustrations or nonverbal means of communication are used for task instructions, the meaning of such visuals or gestures must be conveyed in some manner to the examinee, and it is difficult to do so in the complete absence of verbal communication. Moreover, when gestures are used in communicating task instructions, once the examinee learns and understands the meaning of the gestures, the gestures essentially become a form of sign language that is used for the purpose of testing. Thus, some form of language is used regardless of the extent to which verbal communication is involved during the testing process. It is therefore more accurate to refer to such "nonverbal" tasks as *language-reduced*. Moreover, the reduction of language does not always result in the elimination of acculturative content. For instance, the difficulty level of a nonverbal test of visual memory involving the presentation of objects commonly found within a particular culture may not be the same for individuals from a different culture who may have markedly different exposure or experience with the object presented (Ortiz & Wong, 2022). Furthermore, research has indicated that the strict verbal-nonverbal conceptualization of test performance for ELs is far too simplistic because the extent of cultural and linguistic influences on test performance requires

careful consideration of the unique characteristics of individual subtest and the construct it purports to measure (Ortiz et al., 2018). Lastly, given the vast majority of school-based referrals are related to reading and/or writing difficulties, avoiding the assessment of language-based abilities such as crystallized intelligence (Gc), auditory/phonological processing (Ga), or long-term storage and retrieval (Glr) would not facilitate the determination of the nature and extent of the learning problems to inform appropriate interventions (Ortiz & Wong, 2022).

Use of Native Language Tests

In an effort to mitigate possible bias in testing ELs in their L2, the use of native language (L1) tests has been recommended as one possible solution. In fact, the Individuals with Disability Education Act (IDEA; 2004) and its attendant regulations (i.e., Code of Federal Regulations [CFR]) stipulates that assessments and other evaluation materials used to assess non-native English speakers are to be

> provided and administered in the child's native language or other mode of communication and in the form most likely to yield accurate information on what the child knows and can do academically, developmentally, and functionally, unless it is clearly not feasible to so provide or administer. (§§300.304 Evaluation Procedures; U.S. Department of Education, 2017)

Although administered in an individual's native language, L1 tests remain problematic given that ELs are neither monolingual L1 speakers as they have begun learning L2, nor do they all possess the same learning experiences in L1. For instance, many standardized Spanish tests gather normative data from countries such as Mexico, Peru, or Puerto Rico with monolingual speakers of Spanish who have received formal education in that

language only. This makes them inappropriate for testing bilingual individuals in America as the use of such tests effectively ignore ELs' bilingual status and overlook the variability in experience, exposure, and learning in both their native language and English (Rhodes et al., 2005). Great variability also exists in terms of the amount of exposure ELs receive in their native language in the home or community upon their age of arrival in the host country. The perception of fairness in the use of native language test is thus illusory and presumes incorrectly that two ELs of the same age, gender, grade, socio-economic status (SES), geographic location, and race/ethnicity must have similar levels of exposure to and education in their native language (Flanagan et al., 2013; Ortiz, 2014).

Evaluation in the Dominant Language

Another workaround to allow ELs to perform to the best of their ability is to evaluate them in their dominant language. However, testing in an EL's dominant language suffers from the same limitations already noted for evaluation in L1 or L2 because it presumes equivalency between an EL's age and developmental language proficiency with the normative sample (Flanagan et al., 2013). It is not uncommon for an EL to be dominant in their native language before starting school and then become dominant in English (L2) after having received formal instruction for only a few years, resulting in L1 attrition (Umbel et al., 1992). Despite not being age-appropriate in the development of L1 or L2, ELs are nevertheless evaluated as if they were. Their performances are compared to norms that do not represent their developmental experiences in exposure as well as opportunity for learning L1 or L2.

For example, Ortiz and Wong (2020b) reported some preliminary data ($N = 14$) regarding the use of the Woodcock-Munoz Language Survey, Third Edition (WMLS-III;

Woodcock et al., 2017) in the determination of language dominance in a sample of EL students ranging from kindergarten to the 5th grade (average age = 7 years) from a large suburban school district in the southern part of the U.S. The group obtained a mean standard score (SS) of 54 for their general English language ability score, which was more than three standard deviations below the standardization mean when compared to same age monolingual, native English-speaking peers on the WMLS-III. The highest SS obtained in the group was 69, indicative of very poor English proficiency. When the group of students were compared to monolingual, native Spanish-speaking peers on the WMLS-III Spanish, their mean general Spanish SS was even lower, mean SS = 54. An examination of individual scores showed that only two out of the 14 students obtained SS that were within normal limits. The rest of the students scored poorly and well below normal limits, with the highest score being SS = 72. Based on a comparison to each student's English language scores, of the 12 students who obtained below average scores on WMLS-III Spanish, only three could be considered "dominant" in Spanish. Although they scored slightly higher in Spanish than in English, the "dominance" was determined on the basis of scores such as SS = 73, 59, and 57 in Spanish vs. SS = 40, 40, and 43 in English, respectively. One would argue that despite the comparatively higher scores obtained on the Spanish subtests, such Spanish language scores would hardly qualify one to be considered age-appropriate in their Spanish language development. Along with the other two students who obtained average scores on WMLS-III, if the five students were evaluated further in their "dominant language" of Spanish, these students would likely be penalized for their lack of learning opportunity (i.e., formal education) in Spanish compared to their native Spanish-speaking same age peers and thus be misidentified as

having some type of speech-language impairment or disability depending on the nature and pattern of the additional testing in the Spanish language.

Similarly, based on the English language scores, of the 14 students, nine were found to be English "dominant," based on score comparisons such as SS = 64 in English vis-à-vis SS = 40 in Spanish. If further testing were to be conducted in English, then their poor performance would likely be attributed to speech-language impairment or some type of disability. Irrespective of the language in which the 14 students were to be evaluated further in, the reliance on test scores on the WMLS-III in determining language dominance and then testing in their so-called "dominant language" could potentially lead to 12 of the 14 students as being misidentified with some type of disability. This 86% identification rate is alarming, because none of the students' bilingual language developmental experiences were taken into account when their language proficiency was assessed in either language. Their performances on the language tasks in both languages were assessed by unfair comparisons with that of the normative samples who possess learning experiences of a native Spanish or English speaker with formal education in either language. As the above example illustrates, evaluation in the dominant language is not an equitable solution in the assessment of ELs' cognitive or language abilities.

The Influence of Exposure to English in the Assessment of ELs

It is clear that all of the above methods commonly lead to misinterpretation of low scores obtained by ELs on both native and English language tests. As Fisher and Frey (2012) pointed out,

> It is unlikely that a second-grade English learner at the early intermediate phase of language development is going to have the same achievement profile as the native

> English-speaking classmate sitting next to her. The norms established to measure fluency, for instance, are not able to account for the language development differences between the two girls. A second analysis of the student's progress compared to linguistically similar students is warranted. (p. 40)

It is apparent that for fair evaluation to occur, ELs must be distinguished from ESs who have been learning English from birth. As well, rather than being compared to monolingual speakers of their native language, they should be compared to fellow bilingual ELs with similar experiences in language development who are their true peers. In addition, because the acquisition of English can begin at any given point in an EL's development, the amount of exposure or opportunity in learning English must be controlled for, in addition to the variable of age which controls for cognitive development (Ortiz, 2014; Ortiz, 2018; Ortiz & Wong, 2022).

The importance of accounting for ELs' different levels of exposure to and experience in learning English when evaluating their English speaking ability is not new or considered important only by researchers in the field of psychology. Using data from the 2017 American Community Survey (ACS) conducted by the U.S. Census Bereau, demographers Dietrich and Bauman (2019) examined the relationship between levels of exposure (as categorically indexed by native-born versus foreign-born; age of entry at a young age versus an older age; and living in the U.S. for a few years versus many years) and English-learning ability (a dichotomous variable recoded from the ACS: those who speak only English or speak a language other than English but English "very well" versus those who speak English less than "very well"). Logistical regression analyses revealed that foreign-born children were less likely to speak English very well compared to native-

born chidlren, and that the younger the children were entereing into the U.S., the greater the likelihood of them speaking English very well. Furthermore, the longer that a child had spent living in the U.S., the more likely that they would speak English very well. On the other hand, children who entered the U.S. at an older age also tended to make greater gains in relation to their English-speaking ability (i.e., learning at a faster rate) and eventually caught up with their counterparts who had entered the U.S. at a younger age. This popluation-based research highlights the importance of exposure in learning English as a foreign language.

The above demographic study concerns general English speaking ability in the U.S. population. Further support for accounting for ELs' different levels of exposure to and experience in learning English when evaluating their abilities is illusrated by Sotelo-Dynega et al.'s (2013) study. The researchers examined the relationship between English proficiency level (EPL) as measured on the New York State English as a Second Language Achievement Test (NYSESLAT; New York State Education Department [NYED] & Harcourt, 2006) and performance on seven subtests of the Woodcock Johnson Test of Cognitive Abilities, Third edition (WJ III; Woodcock et al., 2001). Participants' performance on the NYSESLAT resulted in four groups at varying levels of EPL—Beginner, Intermediate, Advanced, and Proficient. The EPL can thus be thought of as an objective measure of exposure level to English. Results from the study showed that performance on the WJ III was most attenuated by the participants' EPL on subtests with the most linguistic demand and cultural loadings, such as Verbal Comprehension and Concept Formation which measure the knowledge of word meanings and the ability to identify the rule for a novel concept in relation to colored circles and triangles,

respectively, compared to tasks such as Visual-Auditory Learning, Sound Blending, Number-Reverse, Visual Matching, and Spatial Relation (in order of descending linguistic demand and cultural loading). Although the task of Concept Formation taps fluid reasoning which is usually considered a less verbally demanding cognitive ability domain, the task on the WJ III is also considered a "learning" test because it requires the examinee to comprehend task instructions as well as verbal corrective feedback that are given to the examinee in order to perform well. Furthermore, within-group differences were also found as a function of linguistic demand and cultural loading of the WJ III task, such that individuals scoring at the Beginning and Intermediate EPLs on the NYSESLAT performed increasingly worse on tasks that require more developmental language proficiency and cultural knowledge, with mean SS as low as 60.67 on Verbal Comprehension and 78.33 on Concept Formation for the Beginning group, and 68.67 and 81.86 on those tasks respectively for the Intermediate group, compared to mean SS of 82.45 and 86.24 on those tasks for the Advanced group, and 93.50 and 96.36 for the Proficient group, respectively. As evident in the statistically significant subtest scores between the Beginner and Intermediate groups versus the Advanced and Proficient groups, it is crucial to consider ELs' EPL when assessing their cognitive abilities, particularly on tests that place the most demand on language and cultural knowledge. As linguistic demand and cultural loading increase across the subtests on the WJ III, the performance in all the groups became more and more affected by their English proficiency and exposure. In fact, 31% of the variance across the seven subtests were accounted for by the NYSESLAT proficiency level. In other words, exposure to English predicted one-third of the variance in subtest performance on the WJ III. This shows the

importance of accounting for exposure to and experience in both language and culture when assessing ELs, especially in any tests that assess those abilities.

The above study replicated an earlier study by Dynda (2008) where ELs' performance on the Woodcock-Munoz Language Survey-Revised (WMLS-R; Woodcock et al., 2005) was used to assess the ELs' developmental proficiency in (or exposure to) the English language which resulted in the categories of Low, Intermediate, and High. Dynda (2008) compared the performance of the three groups of ELs across four subtests from the Wechsler Abbreviated Scales of Intelligence (WASI; Wechsler, 1999), namely, Matrix Reasoning, Block Design, Similarities, and Vocabulary, two subtests on the WMLS-R, namely, Letter-Word Identification and Picture Vocabulary, and one subtest on the WMLS-III, Dictation. Similar to Sotelo-Dynega and colleagues' (2013) study, the researcher found the same attenuation effect on cognitive test performance when the subtests were arranged along a continuum of linguistic demand, with all three groups with varying EPL performing close to the standardization mean of 100 on tasks that were considered more "language-free" (e.g., Matrix Reasoning and Block Design on the WASI) and worst on tasks that were considered more "language-based" (e.g., Vocabulary on WASI and Picture Vocabulary on WMLS-R). The decline was the most significant for the Low proficiency group, with mean SS dropping to 40 on Picture Vocabulary, compared to mean SS = 70 for the Intermediate proficiency group, and mean SS = 90 for the High proficiency group.

The linear pattern of decline in EL test performance based on linguistic demand and cultural knowledge on cognitive tasks was also observed in a study by Cormier and colleagues (2014). In their study, the extent to which linguistic demand and cultural

loading influence test performance on 20 WJ III subtests was examined. The researchers used a subset of the normative sample from the WJ III (ages 7-10, 11-14, and 15-18) for their analysis which contained both native-English speakers as well as some ELs identified a priori by demographic information that were obtained during the norming process. In this study, linguistic demand was defined as the examinee's level of ability in receptive and expressive language in the language of administration of the test, as determined by the individual's performance on four oral language measures (Understanding Directions, Oral Comprehension, Story Recall, and Picture Vocabulary) on the co-normed Woodcock-Johnson Tests of Achievement, Third Edition (WJ III ACH; Woodcock et al., 2001), while cultural loading was defined as quantifiable characteristics that contribute to the examinee's association with a particular cultural group including a) foreign born status, b) race, c) language spoken at home, and d) first language. Results indicated that variance in test performance across all three age groups was explained by individual differences in expressive and receptive language abilities to a very high degree on the test with the highest demand for linguistic development, Verbal Comprehension (79% to 86% across the three age ranges) and the test with the highest demand for acculturative knowledge, General Information (71% to 86% across the three age ranges), followed by tests with moderately high language demand due to the complex verbal instructions involved, Concept Formation (67% to 71% across the three age ranges), and ranged downward to a very low degree on subtests within the domains of visual-spatial processing (Gv) and Processing Speed (Gs) such as Picture Recall (7% to 11% across age groups) and Planning (2% to 10% across the three age groups), with all other tests falling somewhere between the two *very high* and *very low* anchors. Interestingly, the finding

that language abilities exerted an equally important influence on the performance of the monolingual, native-English speakers compared to the EL sample came as a surprise to the researchers, so much so that a recommendation was included to caution evaluators to consider the effect of language on test performance when assessing monolingual ESs with suspected speech-language difficulties. Using data obtained from the Woodcock-Johnson IV (WJ IV; Schrank et al., 2014) normative sample, Cormier and colleagues (2022) provided further support for the important role language ability plays in cognitive test performance. Using a mixed-effects modeling approach, they found that language abilities (both expressive and receptive, but particularly receptive) appear to exert a significant influence on cognitive test performance, whereas test characteristics (i.e., test directions) did not influence performance, after accounting for language abilities. Based on the study's results, the researchers cautioned that when assessing ELs, limited English proficiency can lead to linguistically biased test results. That, in turn, can lead to a misinterpretation of the examinee's true cognitive abilities.

Taken together, the above studies illustrate that in order to validly measure ELs' abilities, two factors must be considered. First, it is necessary to consider that in comparison to typically developing monolingual ESs whose language development can be presumed to commensurate with age, ELs' performance on tasks can be affected by the age- or grade-level expected English language development as well as acculturative knowledge that are inherently built into language-based tests. Second, developmental differences in language among ELs of the same age must be accounted for. This is because ELs are not one monolithic group, and age alone does not account for development in either an EL's L1 or L2. Thus, true peer comparison for ELs involves

comparing ELs with other same age ELs with the same amount of exposure or proficiency in English (Ortiz, 2018; Ortiz et al., 2018; Ortiz & Wong, 2020b; Ortiz & Wong, 2022). Recent advances in test fairness have focused on improving true peer group comparisons via the creation of independent norms for both ESs and ELs, whereby the amount of English exposure is controlled for in the EL normative sample. An example of a new assessment that exemplifies the true peer group comparison approach is the Ortiz Picture Vocabulary Acquisition Test (Ortiz PVAT; Ortiz, 2018).

The Use of True Peer Norms in the Assessment of ELs

The Ortiz PVAT (Ortiz, 2018) represents a new way of assessing ELs that considers differences in language development among ELs via the use of dual norms—a set of age norms for ESs for evaluating monolingual, native ESs and a different set of norms for evaluating ELs with the proportion of lifetime exposure to English controlled for alongside the variable of age. Because it has two sets of norms, it provides valid comparison groups for assessing both ESs and ELs. For ELs, language development information gleaned from clinical interviews with the primary caregiver including an examinee's chronological age, age at first English exposure, active learning experiences as well as formal education in English are used to calculate a percentage value in the test program to represent the amount of life-time English exposure possessed by the examinee. The examinee is then compared to other same-aged ELs with exactly the same percentage of life-time exposure to English, which can range from 1% to 99%. In other words, true peer comparison for ELs is achieved on this test because each EL examinee is compared to their same age EL peers in the normative sample with precisely the same proportion of life-time exposure to English. The Ortiz PVAT provides SSs with a mean

of 100, standard deviation of 15, and classifications of *Extremely Low* (SS < 70), *Very Low* (SS = 70 to 79, *Low* (SS = 80 to 89), *Average* (SS = 90 to 109), *High* (SS = 110 to 119), *Very High* (SS = 120 to 129), and *Extremely High* (SS ≥ 130).

The Ortiz PVAT focuses on the assessment of English receptive vocabulary skills, which makes it easier to evaluate ELs who are at the very beginning stages of their second language acquisition (e.g., preproduction or silent period). Moreover, testing L2 vocabulary rather than that in L1 allows for English to be a common metric for evaluation for all students and permits all clinicians (monolingual or bilingual) to begin the process of evaluating an EL's language functioning in English (Ortiz, 2018; Ortiz & Wong, 2020a; Ortiz & Wong, 2020b; Ortiz & Wong, 2022).

Test Fairness Achieved via True Peer Group Comparison on the Ortiz PVAT

The Ortiz PVAT technical manual (Ortiz, 2018) provided strong evidence for the need for separate norms when assessing ELs' receptive vocabulary. When scored against the ES normative sample, individuals in the EL normative sample scored consistently below 100 (the ES normative sample mean), especially when exposure to English was low (i.e., 0–10% of the lifespan). The mean SS based on ES norms for the highest exposure group (i.e., 51–100% of the lifespan) most closely mirrors the mean SS for the monolingual ES group, as expected due to their substantial amount of exposure to English. However, when the EL groups' performance was scored with the EL norms, mean SS across all three levels of exposure were classified in the *Average* range (SS = 90–109). The above results demonstrated that ELs in the Ortiz PVAT normative sample achieved developmentally typical scores when compared against a valid reference group of multilingual peers (i.e., their true peer) based on proportion of lifetime exposure during

the scoring process. This finding supports the validity of dual norms and the necessity of creating two reference samples to ensure fairness when evaluating language.

Additional support can be found in García (2022)'s unpublished dissertation in which she evaluated the English vocabulary knowledge of 24 Spanish-speaking ELs between the ages of 5 and 11 years (M_{age} = 8.25) who resided in the Northeastern part of the U.S. with English language exposure ranging from 25% to 100% across their lifespan. She found that the Ortiz PVAT which accounted for the ELs amount of lifetime English exposure offered a more accurate measure of English vocabulary knowledge. Specifically, scores obtained on the Ortiz PVAT for the participants fell within the *Average* range. She also found statistically significant differences between scores obtained using the Ortiz PVAT and those obtained using the non-exposure-based norms on the WMLS III Test 1: Analogies, Test 3: Picture Vocabulary, and Basic English Oral Language Cluster with large effect sizes.

Furthermore, in the previously mentioned preliminary study on the use of WMLS-III to determine language dominance reported by Ortiz and Wong (2020b), the school district had also begun to use the Ortiz PVAT as a measure to identify students who might be in need for a referral for assessment. When the students' test performances were compared to their same age EL peers with the same amount of exposure to English on the Ortiz PVAT, their mean SS was 92, which places the group mean in the *Average* range. One student obtained a score in the *Very Low* range (SS = 71), two students received SS of 84, and the rest of the students scored 87 or higher. This pattern of results indicates that most of the students in this group were actually acquiring English vocabulary at a rate that was commensurate with their same age EL peers with same amount of English

exposure. In other words, other than the student who scored in the *Very Low* range and perhaps the two students who obtained a borderline score of SS = 84, this group of students generally showed vocabulary performance typical of other ELs with similar language development experience and therefore did not require further assessment for diagnostic purposes. Although they might require support in their second language development to help with their learning (e.g., bilingual instructions, modifications in the way English instructions are delivered, increased explicit instruction of English vocabulary and/or speech sounds; Ortiz, 2018), there was insufficient evidence to suspect that these students had an inherent speech/language impairment or disability that preclude them from learning English at a pace that is expected of a typical EL.

Compared to the 86% identification rate when the language dominance method was used based on WMLS-III English/Spanish test performance, the use of the Ortiz PVAT with EL norms that control for exposure to English resulted in only one out of 14 students (7%) whose score would be definitively considered atypical compared to their EL peers, with two other students who obtained scores that were fairly close to normal limits (SS = 84 for both students). Considering the possibility of measurement error with borderline scores, the pattern of results obtained with the Ortiz PVAT suggested that for a majority of the students in the group (11 students, or up to 13 students), any difficulty in the classroom were most likely unrelated to a speech-language impairment or an inherent disability. Their rate of acquisition of the English language was commensurate with what would be expected of their same age EL peers with the same amount of exposure to English, showing typically observed ability to learn English compared to their true peers. Thus, using the Ortiz PVAT, at most, three students (21%), and perhaps just one (7%), of

the 14 students who were referred for evaluation might have a disability related to language and warranted further evaluation as opposed to the 86% of students in the group using the dominant language approach based on their WMLS-III English or Spanish scores. The difference is phenomenal, and it illustrates the importance of true peer norms in ensuring test fairness for ELs to prevent the misidentification of disabilities in ELs.

Test Fairness and Generalizability of the Ortiz PVAT Across Demographic Groups

According to the 2014 *Standards for Educational and Psychological Testing* published by the American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), "fairness is a fundamental validity issue and requires attention throughout all states of test development and use" (p. 49). Although it is of critical importance in the process of test development, the term *fairness* is difficult to define in technical or empirical terms. Rather, it remains a fundamentally conceptual issue (Ortiz, 2018). Historically, fairness was conceived as largely an issue of measurement bias. However, advances in psychometrics have contributed to increasing evidence of validity for many large-scale standardized measures developed for diverse populations (Cormier et al., 2022), with much attention paid to item development and norming. In their analysis of the theoretical, empirical, and practical issues in testing ELs, Ortiz and Wong (2022) postulated that most tests are not biased on the basis of technical deficiencies in test development. Rather, bias in the assessment of ELs stems from differences in language and acculturative knowledge acquisition that are confounded with the constructs the tests purport to measure because adequate language proficiency is required to perform well on the tests. And it is such confounds that can lead to interpretive or diagnostic errors when

normative samples that do not control for differences in both age and amount of exposure to English for ELs are used to make test score comparisons because they assume that ELs' language development is commensurate with same age monolingual English-speaking peers. On the other hand, when test performances of ELs who are by no means a monolithic group in terms of their language developmental or acculturative experiences are compared to that of speakers of their heritage language, fairness is again compromised. When diagnostic decisions are made based on results generated from such methods, one cannot be certain whether any deficits in test performance is due to an inherent disorder in language or psychological processes or a difference in language developmental/acculturative experiences. The Ortiz PVAT was developed precisely to eliminate such confounds in test score interpretation by accounting for the proportion of lifetime exposrue to English, in addition to age.

Nevertheless, to ensure that the test is fair for use for diverse populations, an investigation of the assessment's specificity to the measure of vocabulary ability (what the test purports to assess but not other individual characteristics) must be provided as sources of evidence for the fairness of the Ortiz PVAT. The test's technical manual provides a few analyses to examine the generalizability of the scores by examining the effects of demographic group membership via the comparison of mean SSs. A series of ANCOVAs was conducted to compare scores across demographic variables (i.e., gender, PEL, and racial/ethnic groups for the ES normative sample; gender, PEL, and home language spoken for the EL normative sample; Ortiz, 2018). Because there are two sets of norms, SSs based on the respective norms were examined separately for the ES and EL normative samples, and they were entered as the dependent variables. The target

demographic variable to be assessed for fairness was entered as the independent variable in each analysis, along with relevant demographic characteristics statistically controlled. The test developer also used a conservative coefficient alpha level of $p < .01$ to control for Type I errors that could arise from multiple comparisons. In addition to significance levels, measures of effect size (Cohen's *d* ratios and partial η^2) were also included in their analyses.

Demographic Effect of Gender

When gender differences were investigated in the ES normative sample, the demographic variables of region, PEL, and race/ethnicity were included as covariates in the ANCOVAs to control for their possible effects. The test developer found that on both Ortiz PVAT A and B forms, there was no evidence of a statistically significant difference between male and female examinees, and the effect sizes were negligible in the ES sample (Cohen's $d = -0.04$ for both Form A and Form B; based on Cohen's [1988] guidelines in the interpretation of effect sizes where Cohen's $d = .20$ indicates a small effect, Cohen's $d = .50$ indicates a medium effect, and Cohen's $d = .80$ indicates a large effect). The results indicated that the measurement of receptive vocabulary comprehension with the Ortiz PVAT using the ES norms was unaffected by gender. For the EL normative sample, with region, PEL, and language spoken at home (in place of race/ethnicity) included as covariates to control for their possible effects, the main effect of gender was also found to be nonsignificant, with negligible effect sizes (Cohen's $d = -0.01$ and 0.00 for Form A and Form B, respectively). The nonsignificant results provided support that receptive vocabulary comprehension ability is measured in a similar manner across genders for both Ortiz PVAT normative samples.

Demographic Effect of Parental Education Level

Because higher PEL is associated with increased vocabulary size in children (Hart & Risley, 2003), the effect of PEL was investigated in both the Ortiz PVAT ES and EL normative samples with the possible confounding effects of other demographic variables such as gender, geographic region, race/ethnicity (for ESs) and home language (for ELs) controlled for. Mean SSs were compared among the four PEL groups: a) less than high school diploma, b) high school graduate, c) some college or associate degree, and d) bachelor's degree or graduate/professional degree). An overall main effect was observed as theorized for both ES and EL samples on both forms, indicating statistically significant differences between the four levels of PEL. However, the size of this effect was small for all mean comparisons (Partial $\eta^2 = .010$ for Form A and .011 for Form B for the ES sample and Partial $\eta^2 = .018$ for both Form A and Form B for the EL sample, based on Cohen's [1988] proportion of variance effect size cut-off points where small = .01; medium = .06; large = .14). Pairwise comparisons indicated that small differences were observed between the lowest PEL (less than high school diploma) and the two highest levels (some college or associate degree and bachelor's degree or graduate/professional degree) on the ES sample such that higher scores were observed for examinees with parents who had higher levels of education relative to examinees with parents who did not have a high school diploma. On the other hand, medium-sized differences were observed between individuals with parents who completed some college or an associate degree and the two lower PEL groups (i.e., less than high school diploma and high school graduate) on the EL sample. The pattern of these above effects was in line with existing

research showing that higher levels of parental education positively contribute to greater vocabulary ability in their children (Ortiz, 2018).

Although group mean score differences by PEL were observed, the test developer postulated that such findings alone did not necessarily indicate a potential for bias. Thus, a differential test functioning (DTF) which measures whether the overall test score was truly measuring the same construct for different groups by comparing test response functions for each PEL group was also conducted. Results indicated that the curves for the two PEL group pairs that were indicated as statistically different in the pairwise comparisons described above appeared to be parallel and close together, suggesting similar test functioning between groups. The test developer concluded that such findings demonstrated the lack of variance in the measurement of vocabulary ability across PEL groups, and the high degree of congruence in DTF curves provided support that the significant effect of group mean score differences by PEL did not indicate test bias. Rather, PEL, as theorized based on research vocabulary development, is likely to influence an examinee's score on the Ortiz PVAT to some extent, such that higher PEL is associated with higher scores (Ortiz, 2018). Given the similarity observed between the DTF curves for the PEL groups, the test developer further concluded that the Ortiz PVAT accurately captures the same construct of receptive vocabulary acquisition for individuals with both lower and higher PELs.

Demographic Effect of Race/Ethnicity for the ES Normative Sample

Differences across the four racial/ethnic groups as defined by major U.S. Census categorizations: Black, Hispanic, White, and Other (Asian, Native, Multiracial, and Other were combined due to small sample size for each) in the ES normative sample on the

Ortiz PVAT were examined for meaningful differences in mean SSs, while controlling for the effects of gender, geographic region, and PEL. No significant effects were found for Form A or Form B, and effect sizes were negligible (Partial $\eta^2 = .005$ for both forms). The test developer thus concluded the lack of bias in the measure of receptive vocabulary across racial/ethnic groups on both Ortiz PVAT forms. This finding is noteworthy, considering historically, research showed that non-White individuals tend to perform, on average, 10 standard scores below Whites on cognitive measures (particularly on Gc subtests with a high degree cultural loading), reflecting systematic bias against individuals who do not share similar acculturative experiences with the mainstream culture (e.g., test items involving experiences that are more common among Whites such as *camping*; Dunn & Dunn, 2007). In recent years, advances in psychometrics such as the use of item response theory in test item development and the expansion of floor (very easy) items have led to improvements in the reduction of system bias against ethnic and racial minorities (Dunn & Dunn, 2007). The fact that no statistically significant differences were found on the basis of race/ethnicity in test performance for the Ortiz PVAT ES normative sample demonstrates that items did not vary signficantly in difficulty depending on one's racial/ethnic background and/or acculturative experiences. It also points to the importance of adhearing to a strict inclusion criteria for the monolingual native ES sample (as described in the *Measur*e section under *Methods*).

Demographic Effect of Home Language for the EL Normative Sample

For the EL sample, home language replaced race/ethnicity as the variable for stratification. Four major language groups were targeted, in proportions that were representative of U.S. general population of individuals who spoke a language other than

English in accordance with U.S. Census figures (United States Census Bureau, 2015), namely, Spanish & Spanish Creole, Indo-European languages (note that this category includes all Indo-European languages other than Spanish), Asian & Pacific Islander languages, and Other languages. Differences in mean SSs between these language groups were analyzed with possible effects of other demographic characteristics such as gender, PEL, and geographic region controlled for. In line with hypothesized expectations from the test developer, no statistically significant differences were observed between the language groups and effect sizes were negligible (Partial $\eta^2 = .004$ for both forms). Thus, according to the test developer, no language group were found to outperform any other group as Ortiz PVAT was designed to assess receptive vocabulary ability in English only (Ortiz, 2018). DTF analyses were not conducted by language group due to the unequal sample sizes as a result of the unequal proportions in the U.S. population.

Chapter II

Purpose and Hypotheses

Based on the results from the generalizability studies above, the test developer concluded that no evidence of test bias was found on the basis of gender, PEL, race/ethnicity (for the ES sample), or home language spoken (for the EL sample). Moreover, the developmentally typical mean SS among EL groups with varying levels of lifetime exposure to English when their test performances were compared against their true peer group supported the validity of the Ortiz PVAT dual norms. Taken together, there appears to be strong evidence that the Ortiz PVAT meets the fairness requirements outlined in the 2014 *Standards*. Nevertheless, limited evidence is available regarding the test's generalizability beyond the normative sample since its publication in 2018. Research is needed to replicate the above generalization studies to validate the use of the test in diverse populations to further elucidate the test's clinical utility. Moreover, it is important to establish the utility of the proportion of lifetime exposure to English variable in the prediction of the variability in receptive vocabulary test performance in a community sample to establish the need for ELs norms that account for different language developmental experiences.

Purpose

The purpose of the current study was three-fold. First, the present study sought to replicate the generalizability studies conducted with the Ortiz PVAT normative sample to show a lack of systematic bias on the basis of available demographic variables such as gender and home language in a local community sample. Second, the current study aimed to provide support for the use of dual norms when assessing ELs' English receptive

vocabulary development to account for the influence of exposure to English. Specifically, the present study sought to replicate the Ortiz PVAT validity study to show that overall developmentally typical or *Average* performance by ELs can only be achieved when their test performances are compared to their true peer group with equivalent amounts of lifetime exposure to English, but not when their test performances are compared to monolingual native English speakers. Lastly, in order to extend the research by Dynda (2008) and Sotelo-Dynega et al. (2013), the current study aimed to examine the extent in which proportion of lifetime exposure to English can predict performance on the Ortiz PVAT.

Hypotheses

Based on existing theories and previous research discussed in the literature review section, the following hypotheses were generated:

1) It was hypothesized that no significant main effect of gender would be found when the effects of other relevant demographic variables are controlled for.
2) It was hypothesized that no significant main effect of home language would be found when the effects of other relevant demographic variables are controlled for.
3) It was hypothesized that higher amounts of lifetime exposure to English would be associated with higher test performances on the Ortiz PVAT.
4) When ELs were classified into low, medium, and high lifetime exposure groups, a significant effect of exposure was predicted for mean raw scores and SS based on ES norms, but not SS based on EL norms because the proportion of lifetime exposure to English is factored into the computation of SS_EL.

5) It was hypothesized that lifetime exposure to English would predict a significant proportion of the variance in raw scores, demonstrating again the need for the EL norms to appropriately control for ELs' language developmental experience.

Chapter III

Methods

Participants

A sample of 226 participants between the ages of 2 years 7 months and 18 years 11 months (M_{age} = 8 years 1 month, SD_{age} = 3 years 7 months) was drawn from archival test data obtained from an aggregate of prior smaller studies and clinical applications conducted in the New York City metropolitan area. Anonymity and confidentiality of participants were protected as only CaseID were used as identifier in the archival data. Demographic information available from the archival data included age, gender, grade, first home language, second home language, and proportion of lifetime exposure to English (in percentage). No other identifying information such as participant name, residence, or name of attending school was available in the dataset, nor demographic information involving race/ethnicity and PEL. Within the sample, 25 were identified as monolingual native ES and 201 were identified as EL. One hundred and thirty-four individuals were identified as males and 91 were identified as females (one participant's gender was omitted). Within the ES sample, there were 11 males and 13 females. The EL sample consists of 123 males and 78 females. Table 1 presents the minimum and maximum values as well as the means and standard deviations for the demographic variables of age, grade, and proportion of lifetime exposure to English. Table 2 presents a frequency breakdown of grade by sample.

Table 3 presents the frequency breakdown of first home language in the EL sample by major language groups as defined by the U.S. Census Bureau (2021) for indivduals in the U.S. who speak a langauge other than English. As shown in Table 3,

Spanish and Other European Languages were the first home language identified by a vast majority of the EL sample (46.8% and 36.8%). Six individuals (3%) in the EL sample identified English as their first home language, with three of them having Spanish indicated as their second home language and one of them with Russian indicated. Of note, home language (the primary language spoken at home) is not the equivalent of heritage language (the language that is spoken by individuals who share the same cultural heritage as the EL). For the remaining two ELs with English indicated as their home language without a second home language indicated, they should not be mistaken as monolingual native ES, as ELs can have varying levels of exposure and proficiency levels in both their heritage language and second language (e.g., nonbalanced, balanced, or mixed bilinguals; see Rhodes [2005] and Ortega [2009]). One participant had 71% of lifetime exposure to English while the other had 100% lifetime time exposure (i.e., having begun learning English from birth). Overall, within the EL sample, a second home language were indicated for 55 participants, with English being the most frequent second language identified (44 individuals or 20.4% of the EL sample), followed by Spanish which is the second language spoken by three individuals (1.3% of the EL sample), German and Q'equuchi', each spoken by two participants (.9% of the EL sample, respectively), and Mam and Russian, each spoken by one participant (.4% of the EL sample, respectively).

Measure

The Ortiz PVAT (Ortiz, 2018) is an individually administered test of English receptive vocabulary developed for children and youth between the ages of 2 years 6 months and 22 years 11 months that can be administered via Microsoft Windows

computers or an Apple iPad. The test includes two forms (Forms A and B). Examinees are presented with four pictures depicting real life objects, actions, or scenes on the computer screen along with the audio recording of the English target word. The examinee then selects the picture that best corresponds to the word they just heard. Each correctly answered item constitutes one raw score. The total raw score obtained by the examinee is converted to a SS. The test includes both ES norms as well as EL norms that control for proportion of lifetime exposure to English. Specifically, performance of an examinee identified by the evaluator as a monolingual native English speaker is compared to that of the ES normative sample based on their age at testing. The following criteria describe an ES: a) the language the examinee first learned to speak is only English, b) the language used in the home prior to entering school was only English, and c) the language used for instruction at school, if and when attended, was English only, or English within a dual-language/dual-immersion program). On the other hand, performance of an examinee identified as an EL is compared to that of the EL normative sample based on both their age at testing as well as their proportion of lifetime exposure to English in percentage calculated based on their current age and age at which they first began learning English actively (i.e., the earliest opportunity for the examinee to learn English in a formal setting such as when starting school, in an informal manner such as conversations at home with parents or siblings, or through other consistent exposure such as participating in community activities, preschool, use of technology, interactions with English-speaking extended family). For an EL examinee, the test report obtained from the test developer provides the raw score and standard score based on the EL norms (SS_EL). SS based on the ES norms (SS_ES) is also calculated in the background within the test program to

inform individualized classroom instructional level needs or intervention services that may be required for academic growth, although the actual score is not displayed in the score reports to prevent misuse of the score for diagnostic purposes.

In relation to the test's psychometric properties, it demonstrates good evidence for alternate form reliability (r = .991–.996 for the ES and EL normative samples), internal consistency (marginal reliability coefficient = .98 for both Form A and Form B across the two normative samples and .99 for both forms for the clinical sample), test-retest reliability (corrected r = .72–.81, $p < .001$ for both normative samples), content validity (via subject-matter expert review), internal structure (via support for a single-factor model), clinical utility (via a validity study involving the following clinical groups: Language Disorder, Intellectual Disability, Language Disorder with Expressive Impairment, and Attention-Deficit/Hyperactivity Disorder), and convergent validity (via significant, positive correlations with other established tests of vocabulary or verbal Intelligence, such as the Peabody Picture Vocabulary Test, Fourth Edition [PPVT-4; Dunn & Dunn, 2007] and the Verbal Comprehension Index [VCI] of either the Wechsler Intelligence Scale for Children, Fourth Edition [WISC-IV; Wechsler, 2003] or the Wechsler Intelligence Scale for Children, Fifth Edition [WISC-V; Wechsler, 2014]).

Procedure

A proposal for the present research was submitted to the St. John's University Institutional Review Board (IRB) for a review, and it was determined that the current study (IRB-FY2023-212) qualifies for an exemption due to the secondary use of anonymous research data which is presumed to pose no more than minimal risk to human subjects. The archival dataset was obtained, and raw scores for the EL sample were

converted to SS_ES based on scoring algorithms obtained from the test developer as those scores were not readily available via the test reports generated. SPSS Version 18.0 was used to perform various analyses.

Chapter IV

Results

Relationships among Demographic Variables and Performance on the Ortiz PVAT

As mentioned, three types of scores are available on the Ortiz PVAT: raw score, SS based on the ES norms (SS_ES), and SS based on the EL norms (SS_EL; for ELs only). The following demographic variables were available from the dataset: age, gender, grade, major language group based on first home language, and proportion of lifetime exposure to English. Pearson product-moment correlations (2 tailed) were calculated to explore the relationships among various demographic variables and the three types of scores. Some support for the first hypothesis regarding generalizability of the Ortiz PVAT across gender and home language was found based on correlation analysis. As shown in Table 4, no statistically significant relationship was found between gender and any of the test performance variables (r ranged from .02 to .06, n.s.). Similarly, major language group (i.e., first home language) was not significantly associated with any of the test performance variables (r ranged from -.13 to .11, n.s.). Age in years and months was significantly related to grade ($r = .98$, $p < .01$) as one might expect. Age is also significantly and positively related to raw score ($r = .70$, $p < .01$), which is indicative of vocabulary development and/or general cognitive development that is associated with simple maturation. A significant and negative relationship was found between age and SS_EL ($r = -.23$, $p < .01$) but not SS_ES ($r = -.09$, n.s.). Because age is factored into the calculation of standard scores in both of the age-based norms, the significant negative correlations are considered spurious. In this particular EL sample, the older the EL participants were, their obtained SS also appeared to be lower. The same pattern in terms

of SS_EL was observed for the variable of grade (r = -.18, p < .01) due to the strong relationship between age and grade. With regard to the proportion of lifetime exposure to English (LEE) variable, it was positively associated with age and grade (r = -.46 and -.48, respectively, both at p < .01) as one would expect in relation to the passage of time. LEE is significantly and positively related to all three test performance variables, with the strongest relationship shown with raw scores, followed by SS_ES, then by SS_EL (r = .64, .37, and .19, respectively, all at p < .01). The above finding provides strong support for the third hypothesis where higher LEE was predicted to be associated with higher test performances on the Ortiz PVAT. Lastly, a significant relationship was found between LEE and major language group (r = .32, p < .01). It should be noted that the number of participants in various major language groups were highly uneven in the present study with Spanish and Indo-European languages spoken by the vast majority of the participants in the EL sample (46.8% and 36.8%, respectively). Speakers of Asian and Pacific Island Languages and all other languages represent only 4.5% and 9.0% of the EL sample, respectively. Thus, the positive correlation is likely an artifact of highly uneven and nonrepresentative samples.

To further explore the relationship between the home language and LEE, a one-way ANOVA was conducted to compare the mean LEE among the home language groups. Levene's test for equality of variance was found to be violated, $F(4,215) = 11.90$, p < .001. Thus, a Welch test which is considered a more robust test against a violation of the homogeneity assumption (van den Berg, 2023) was performed, and a significant difference was found F_{Welch} (4,40.96) = 137.62, p < .001). Note that this analysis was done on the entire sample (N = 220) for both ESs (whose first language was assumed to

be English, n = 25) and ELs (whose first language was indicated, n = 195) to match the correlation analysis above. Pairwise comparisons indicated that the mean LEE for the English group (M = 97.0%, SD = 9.7%) was, as one would expect, significantly higher than all of the other language groups ($p < .001$ for all pairwise comparisons). The mean LEE of the Spanish group (M = 51.8%, SD = 19.9%) was significantly higher than that of the Other Indo-European Languages group (M = 36.8%, SD = 28.7%) and the All Other Languages group (M = 20.5%, SD = 16.6%, $p < .001$ for both). On the other hand, mean LEEs do not differ significantly between the Spanish group and the Asian and Pacific Island Languages group (M = 43.3%, SD = 24.2%, p = 1.00). For the Indo-European Languages group, other than the aforementioned significant differences between their mean LEE and that for the English as well as the Spanish groups, a trend toward higher exposure compared to the All Other Languages group (p = .06) was found. No significant difference was found between the Other Indo-European Languages and Asian and Pacific Island Languages groups (p = 1.00).

Demographic Effects on Ortiz PVAT Standard Scores

Gender

To explore the demographic effect of gender on SS, two ANCOVAs were performed with gender entered as the independent variable and home language group entered as a covariate to control for its possible effects (information regarding participants' PEL, race/ethnicity, and geographic region were not available in the current dataset). In relation to SS_ES, as hypothesized, the main effect of gender was not significant, $F(1,215) = .14$, $p = .71$, and effect sizes were negligible (Partial $\eta^2 = .001$). Both the main effect of home language group and the interaction between gender and

home language were non-significant with negligible effect sizes, $F(1,215) = 3.03$, $p = .08$, Partial $\eta^2 = .014$ and $F(1,215) = .651$, $p = .42$, Partial $\eta^2 = .003$, respectively. Also, in line with the first hypothesis, the same pattern of results was found for SS_EL. The main effect of gender was nonsignificant, $F(1,191) = .55$, $p = .547$, with negligible effect sizes (Partial $\eta^2 = .002$). No main and moderating effects of home language were found, $F(1,191) = .002$, $p = .962$, Partial $\eta^2 = .000$ and $F(1,191) = .837$, $p = .361$, Partial $\eta^2 = .004$, respectively. Taken together, the above findings provide strong support for the first hypothesis, demonstrating that the measurement of receptive vocabulary comprehension with the Ortiz PVAT is unaffected by gender.

Home Language

To explore the demographic effect of home language on SS for the EL sample (SS_EL), a one-way ANOVA was performed. Due to the variability in sample sizes among the home language groups, to allow for more meaningful comparisons, the Asian and Pacific Island Languages group ($n = 9$) and the All Other Languages group ($n = 18$) were combined to form a larger *All Other* group. The 6 EL cases with English identified as their first home language were added to the Other Indo-European Languages group ($n = 68$) to form a larger *Indo-European Languages* group. The new home language variable which now included the *Spanish* group ($n = 94$; M = 101.04, SD = 9.27), *Indo-European Languages* group ($n = 74$; M = 97.26, SD = 15.82), and the *All Other* group ($n = 27$; M = 97.44, SD = 9.15) was entered as the independent variable. Because no information on PEL and geographic region were available in the current dataset, no covariate was used in the current analysis. Levene's test for equality of variance was found to be violated, $F(1,192) = 14.68$, $p < .001$. A Welch test which is considered to be a more robust test

against a violation of the homogeneity assumption (van den Berg, 2023) was thus performed. As predicted by the second hypothesis, the main effect of home language was not statistically significant, F_{Welch} (2, 73.82) = 1.73, p = .184. This result was in line with the generalizability study on the Ortiz PVAT normative EL sample regarding home language. No language group was found to outperform any other language group in the current study. Furthermore, despite the significant differences found in mean LEE between the language groups in the EL sample, the lack of mean differences in home language groups is significant because it shows that the EL norms appropriately controlled for differences in LEE between speakers of different home languages. This finding provides further support to the claim that the Ortiz PVAT was designed to assess receptive vocabulary ability in English only, irrespective of the home language spoken by an EL.

Proportion of Lifetime Exposure to English

In order to explore the effects of LEE, three groups of ELs were created by dividing the EL sample into a *Low Exposure* group (0-10%, n = 31), a *Moderate Exposure* group (11-50%, n = 91), and *High Exposure* group (51-100%, n = 79). The cut-points were adapted from the validity studies on the EL normative sample in the Ortiz PVAT technical manual published by the test publisher (Ortiz, 2018). Figure 1 displays the mean standard scores for each group compared to the ES group (n = 25). A visual examination of the graph indicates that when scored against the ES normative sample (using the ES norms), all three EL groups across different levels of LEE scored consistently below the SS of 100 (the ES normative sample mean) as well as the mean SS of 96.3 obtained by the ES sample in the current study. This was particularly true when

LEE was low (i.e., 0–10% of the lifespan). The highest exposure group (i.e., 51–100% of the lifespan) most closely mirrors the mean standard score for the monolingual ES group. This finding was expected due to the High Exposure group's substantial amount of lifetime exposure to English.

Performance based on English Speaker Norms. To examine mean differences in standard scores based on the ES norms (SS_ES) among the various EL groups with different levels of LEE as well as the ES group, a one-way ANOVA was performed. A significant effect of group membership was found, $F(3, 222) = 6.04$, $p = .001$, with group membership (i.e., the level of LEE for EL vs. native English speakers) accounting for 8% of the variance in the SS based on ES (where a Partial $\eta^2 = .06$ indicates a medium effect based on Cohen's [1988] guidelines). Pairwise comparisons indicated that other than the High Exposure group (M = 93.97, SD = 1.36) whose performance did not differ significantly from the mean SS_ES of the ES group (M = 96.32, SD = 2.41, $p = .343$), both the Low Exposure group (M = 86.07, SD = 2.16) and the Moderate Exposure group (M = 88.044, SD = 1.26) scored significantly lower than the ES group ($p = .002$ and $p = .004$, respectively) when their test performance was scored using ES norms which did not account for their LEE. Furthermore, the Moderate Exposure group also scored significantly lower than the High Exposure group ($p = .005$). On the other hand, no statistically significant difference in mean SS_ES was found between the Low Exposure group and the Moderate Exposure group ($p = .344$).

Performance based on English Learner Norms with Lifetime Exposure to English Controlled For. Another one-way ANOVA was conducted to compare the mean standard scores between the four groups similar to the analysis above, but this time, test

performance was indexed by SS_ES for the ES group while SS_EL was used to measure test performance for all three EL groups to control for ELs' LEE as intended by the Ortiz PVAT. No statistically significant effect was found for group membership, $F(3, 222) = .998$, $p = .394$, Partial $\eta^2 = .013$, when test performance for all three EL groups were compared to the Ortiz PVAT EL normative sample with LEE controlled for. Taken together, as predicted by the fourth hypothesis, results from the two ANOVAs above demonstrate that regardless of an EL's level of lifetime exposure to English, on average, ELs obtained a standard score that is equivalent to the mean standard score of 96.32 obtained by the ES group in the current study only when their test performance was compared against a valid reference group of multilingual peers with lifetime exposure factored into the scoring process. When their test performance was compared to the Ortiz PVAT ES normative sample, groups with the lowest and moderate levels of LEE scored, on average, statistically lower than the ES group or the group with the highest LEE. This finding supports the validity of dual norms and the necessity of using two reference samples when assessing ELs' receptive vocabulary development to ensure fairness in score interpretation, particularly for those with the low to moderate LEE.

Predicting Performance on the Ortiz PVAT using Proportion of Lifetime Exposure to English

To examine the unique contribution of LEE in test performance for ELs, a hierarchical regression was performed on raw scores to determine whether LEE improved the prediction of test performance beyond that provided by age. Raw scores were used in the regression analysis to tease apart the unique contribution of the two demographic variables to the total variance in test performance since neither age nor LEE were

factored into the calculation of raw scores. The demographic variable of grade was excluded from the regression analysis due to concerns for multicollinearity stemming from its exceptionally high correlation with age (r = .98). Gender (r = .06) and major language group (r = -.05) were excluded from the regression analysis because of their lack of statistically significant relationship with raw scores (see Table 4).

Table 5 displays the standardized regression coefficients (β), R^2, and change R^2 (ΔR^2). In Step 1, age was entered into the equation, R = .66, F (1, 198) = 155.49, p < .001. Forty four percent of the variance in Ortiz PVAT raw scores was accounted for by the predictor of age alone in Step 1. In Step 2, LEE was added into the equation, R = .79, F (1, 197) = 160.78, p < .001. Sixty two percent of the total variance in Ortiz PVAT raw scores in the EL sample was accounted for by the equation in Step 2. Specifically, when LEE was added into the model, an additional 18% of the variance was accounted for.

A comparison of the standardized regression coefficients revealed that in Step 1, a .67 increase in raw scores was predicted for every SD increase in age (p < .001). In Step 2, when LEE was added into the equation, a .52 increase in raw scores was predicted with every SD increase in LEE, assuming age was held constant (p < .001), and only a .37 increase in raw score is now predicted for each standardized unit increase in age while the effect of LEE was held constant (p < .001). A comparison of the above standardized regression coefficients indicates that LEE exerted a larger effect on the total variance in Ortiz PVAT raw scores in the current sample than age. Results again provide strong support for the fifth hypothesis, LEE significantly predicted performance on the Ortiz PVAT.

Chapter V

Discussion

With the exception of the analyses involving the demographic variables of PEL and race/ethnicity which were not available in the archival dataset, the current study replicated all of the validity studies performed on the Ortiz PVAT normative sample. As well, support for all five hypotheses were found.

Test Fairness in Relation to Generalizability Across Demographic Groups

Specifically, similar to the generalizability studies performed on the normative sample where the effects of gender and home language spoken for the EL sample were found to be nonsignificant, the current study also failed to find any statistically significant differences in SS_ES or SS_EL on the basis of gender, suggesting test fairness for both males and females. In relation to home language spoken by the EL sample, the current study also failed to find any statistically significant differences in SS_EL between participants who spoke Spanish, Indo-European languages (including English), or all other home languages, suggesting that the Ortiz PVAT is a fair measure of English receptive vocabulary regardless of one's native home language. This lack of mean differences in SS_EL between language groups was even more impressive given the significant differences in LEE found between the different language groups in the current sample, suggesting that the influence of LEE was adequately controlled for among different language groups on the EL norms. Because information regarding PEL and race/ethnicity were not available in the current archival dataset, whether the Ortiz PVAT is generalizable across individuals from different racial/ethnic groups or with different levels of parental education await to be explored in future studies. Nevertheless, results

from the current study provide strong evidence for the test's ability to generalize across demographic groups on the basis of gender and home language.

As stated in the 2014 edition of the *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014), fairness is "a fundamental issue in protecting test takers and test users in all aspects of testing" (p. 49). Specifically, fairness is a principal in assessment that must apply to all test takers and not simply to designated subgroups of the population of test takers, such that all test takers regardless of gender, home language spoken, PEL, or race/ethnicity have a right to protection from the harmful influences of bias and unfair influences made from tests. Although the Ortiz PVAT demonstrated good reliability as shown in the test's technical manual, strong reliability alone is necessary but not sufficient to provide evidence for test validity and fairness (Ortiz, 2018). The findings regarding generalizability of the Ortiz PVAT are critical in relation to the concept of test fairness because they suggest that there is no advantage or disadvantage when it comes to learning English vocabulary on the basis of an individual's gender or first/heritage language.

Rationale for the Lack of Score Differences Found between Home Language Groups

Perhaps the finding with regard to the lack of difference found between home language groups was surprizing. After all, some languages bear higher orthographical and/or phonological similarities with English than others (e.g., compared to Chinese or Arabic, Spanish or French perceptibly bear more similarities with English). However, according to Cummins (1981, 1984), English, like any other language, is not learned in a random manner, but rather, in a predictable and common developmental sequence. ELs or learners of any language first learn very basic conversational part of the language, such

as those that are commonly used in everyday social interactions (e.g., yes, no, hot, sit, stop). This type of language is referred to as Basic Interpersonal Communication Skills (BICS; Cummins 1984). Later on, language learners begin to develop the part of the language that facilitates learning and higher order/abstract information processing, known as Cognitive Academic Language Proficiency or CALP (Cummins, 1984). According to Cummins, the developmental sequence from BICS to CALP is universal in learning all languages. However, Cummins (1984) demonstrated in his research on second language acquisition that having developed a high level of CALP in one's heritage language contributes to more successful learning of a second language (i.e., a linguistic transference process). In essence, there appears to be a conferral of advantage in being competent in one language prior to the learning of a second language, but the advantage would only be seen in relation to a faster rate of acquisition, not in terms of the sequence of acquisition. In other words, the developmental sequence of learning English cannot be altered because the L2 is not learned in an aberrant fashion but in highly predictable sequence (i.e., BICS first then CALP; Ortiz & Wong, 2022). On the other hand, the rate at which English is acquired can be influenced by developmental factors, such as formal education in both L1 and L2 as well as proficiency in one's heritage language which may impact the transfer of L1 CALP to the development of L2 CALP. Similarities between one's heritage language and English may facilitate the development of BICS and CALP in English, but research has shown that, on average, it takes 5 years of exposure to a language for CALP to emerge, following the development of BICS. And precisely because variability in language developmental differences in L2 (in this case, English), or

LEE, has been accounted for in the Ortiz PVAT EL norms, test fairness is established for all takers of the test, irrespective of their home language.

Test Fairness in Relation to Proportion of Lifetime Exposure to English

Further support for the use of true peer norms when assessing ELs' receptive vocabulary development to account for the influence of exposure to English was found in the analyses involving the LEE variable. Firstly, as expected, higher LEE was associated with higher test performance on the Ortiz PVAT. This finding is in agreement with what has been observed in multiple research studies (e.g., Dynda, 2008; Sotelo-Dynega et al., 2013; Dietrich & Bauman, 2019; Cormier et al., 2014; Cormier et al., 2022). The stronger associations between LEE and raw scores and SS_ES compared to SS_EL were likely due to the fact that LEE has already been accounted for in the computation of SS_EL but not the other two types of scores. In addition, when ELs were classified into Low, Medium, and High Exposure groups, as predicted, a significant effect of exposure was found on both the mean raw scores and mean SS based on ES norms, but not SS based on EL norms, again, because LEE has been factored into the computation of SS_EL. Although no significant difference in mean performance was found between the EL group with the High Exposure group and the ES group, the mean SS_ES of 93.70 of the High Exposure group is still about 2/3 SD below the normative mean of SS_ES = 100. The lack of significant difference found between the two groups was likely due to observed mean SS_ES of 96.3 for the ES group in this particular community sample. Note that this community sample may include individuals who have been referred for evaluation due to academic difficulties. It is possible that some of the individuals in the ES sample may have an inherent language disorder or disability which might have, in

turn, led to a lower performance on the Ortiz PVAT. Because no information regarding clinical status was available in the archival dataset, it was impossible to determine whether the current ES sample is systematically different from the typically developing Ortiz PVAT ES normative sample. Nevertheless, the statistically significant differences in mean SS_ES for the Low and Medium Exposure groups compared to that for the ES group as well as the overall pattern of mean differences obtained via SS_ES vis-à-vis SS_EL for the EL groups with varying levels of LEE compared to the ES sample as illustrated in Figure 1 appeared to replicate findings from the Ortiz PVAT normative sample, suggesting that the proportion of lifetime exposure to English is an essential variable to be included in the test norms to fairly assess ELs' English receptive vocabulary. Lastly, findings from the hierarchical regression analyses revealed that an additional 18% of the variance in the performance on Ortiz PVAT was accounted for when LEE was added into the model as a predictor alongside age. Furthermore, compared to age, LEE was found to exert a larger influence on Ortiz PVAT raw scores (.52 SD increase vs. .37 SD increase in raw scores with each additional SD increase in LEE or age, respectively, while holding the other variable constant). This finding from the hierarchical regression analysis combined with the mean comparisons involving the different LEE groups and the ES group further illustrate the importance of accounting for LEE when assessing ELs' language related abilities.

According to the test developer, one of the main goals of developing the Ortiz PVAT was to create a test that can solve one of the most critical issues when testing ELs—failure in existing tests to account for differences in language development and English proficiency of ELs compared to native ESs which results in attenuated test

performance and creates an impression of "disorder" as opposed to a mere "difference" (Ortiz, 2018). As indicated by the review of literature in the current paper, attempts to modify the testing process including the use of interpreter or translator, testing the limits, administration of native language or language-reduced tests, or solely evaluating ELs in their "dominant language" do not fully address the problem because the influence of language and culture still permeates the testing process, and the measurement of language and language-related abilities (e.g., vocabulary acquisition) is too crucial in understanding an individual's overall academic success to avoid altogether. Thus, the effects of the examinee's language developmental and acculturative experiences should not be ignored.

Interestingly, for an individual who does not speak English at all, their test performance on a test of English vocabulary acquisition is not necessarily unfair or invalid, because it would clearly indicate that the individual has no vocabulary *in English*. The score, however, does not indicate that the individual has no vocabulary in any language. This distinction is an important one, because it reflects a validity issue that is related to test score interpretation, which is different from a validity issue stemming from test measurement bias in a psychometric sense. As Ortiz (2018) pointed out, the threat to validity in the evaluation of ELs does not involve test construction as advances in psychometrics can ensure test fairness toward diverse groups of test takers, as the previous section on test fairness regarding generalizability across demographic groups illustrates. Rather, it involves the consequence of testing (i.e., how the test result is interpreted). As the results from the current study regarding the LEE variable demonstrate, the dual norms on the Ortiz PVAT allows the evaluator to correctly interpret

an EL's receptive vocabulary in English with the use of EL norms. It is not sufficient, however, to simply gather a sample of ELs and construct a normative sample among them that is stratified based on age and the usual demographic characteristics of gender, PEL, race/ethnicity, and geographic region. Doing so would erroneously imply that all ELs are comparable merely because they are learners of English when, in reality, some individuals have far more exposure and developmental proficiency in English compared to other ELs of the same age. Thus, the development of dual norms is not merely a recognition that ELs require their own true peer group, but it is also intended to acknowledge that ELs norms need to be inherently different and far more complicated than typical normative samples to address the concerns with regard to the consequences of testing with ELs (i.e., fair and accurate test score interpretation). The fact that an additional 18% of the variance was accounted for by adding the variable of LEE alongside age in the current study demonstrates the importance in constructing dual norms that account for ELs' differential language developmental experiences to ensure test fairness and the non-discriminatory evaluation of vocabulary acquisition in both native ES and ELs who speak diverse heritage languages with varying exposure or opportunity to learn English.

Moreover, in their study of linguistic influences on cognitive test performance using the WJ IV normative sample, Cormier et al. (2022) pointed out a surprising finding related to the influence of examinee characteristics on the variance in WJ IV test scores. In their study, the influence of examinees' English ability, as measured by their expressive and receptive language abilities in English, appeared to eliminate the contribution of test characteristics related to linguistic demands of test instructions. More

specifically, the influence of receptive language ability was found to be more influential than age on cognitive test performance. Given that English abilities is highly correlated with LEE, it is therefore not surprising that in the current study, LEE was also found to exert more influence on the variance of the raw scores on the Ortiz PVAT compared to age. And it is likely because the Ortiz PVAT measures receptive language, or specifically receptive vocabulary, in English, the strong effect of LEE above and beyond age was observed.

Limitations and Directions for Future Studies

As stated above, the current study was limited by its archival nature as no information regarding the participants' PEL or race/ethnicity was available to extend the generalizability studies beyond gender and home language. In addition, information regarding the participants' clinical status or reason for referral was also unavailable and thus made it difficult to determine why the mean SS of the ES sample was lower than the normative mean of 100. This might have been crucial information to explain why statistically significant difference was not found between the EL sample with the highest LEE when their performance was scored with the ES norms compared to the ES sample. Another limitation involves the convenience sampling from the various studies that this current archival dataset was based on. Data was drawn from archival test data obtained from an aggregate of prior smaller studies and clinical applications conducted in the New York City metropolitan area. It is possible that results from this current study may not be generalizable to other geographic regions. However, judging from the large sample size of the Ortiz normative ES and EL samples (over 1000 per sample) and the stratification on the basis of geographical region across the U.S. matching the U.S. Census figures

(Ortiz, 2018), there are reasons to believe that follow-up studies conducted with samples from other geographical locations would obtain similar results (i.e., lack of systematic bias on the basis of gender and home language). Future studies can replicate and extend the results of the current study by collecting data from participants from different geographic locations.

In addition to issues related to geographical representativeness, the current sample consists of an extremely small sample of speakers of Asian and Pacific Island Languages (n = 9) and speakers of all other languages (n = 18) compared to individuals whose home language was Spanish (n = 94) or other Indo-European languages (n = 68), rendering the comparisons between the four major language groups difficult. Moreover, the sampling of different home languages within each major language group also left much to be desired in terms of representativeness, with only Chinese/Mandarin and Uzbek representing the Asian and Pacific Island Languages group, and Arabic, Hebrew, and Wolof representing the All Other Language group. Future studies should incorporate more diverse language groups to demonstrate the generalizability of the Ortiz PVAT.

Lastly, because no other test scores were available in the current dataset, it was impossible to compare the utility of the Ortiz PVAT dual norms vis-à-vis other language tests that only provide norms based on native ES. For instance, future studies can examine the extent of convergence in scores obtained from the Ortiz PVAT and other language proficiency tests, such the WMLS-III. While both the WMLS-III Picture Vocabulary subtest and Ortiz PVAT assess English vocabulary development, they assess different modalities of lexical knowledge (expressive vocabulary on the WMLS-III vs. receptive vocabulary on the Ortiz PVAT). Furthermore, based on the preliminary school

data discussed in Ortiz and Wong (2020b), due to the lack of consideration of language development in the monolingual ES norms on the WMLS-III, a significant positive correlation would be expected but the strength of the correlation would likely be attenuated due to the aforementioned differences between the two tests. Indeed, in García's (2022) study involving a sample of 24 Spanish-speaking ELs, scores obtained on the Ortiz PVAT based on the EL norms and scores obtained on the WMLS-III on the Analogies and the Picture Vocabulary subtests and the Basic Oral Language Cluster showed statistically significant but moderate correlations (r ranged from .414 to .611, with p ranging from <.05 to <.001), with the highest correlation found between the Ortiz PVAT and the most reliable indicator of English language abilities on the WMLS-III, the Oral Language Cluster, which is comprised of performance from multiple subtests. García (2022) also found statistically significant mean differences in the test scores obtained on the Ortiz PVAT compared to the WMLS-III with large effect size despite the small sample size. The researcher urged future studies to examine the above effects using larger sample sizes.

It would also be interesting to compare the potential identification rate based on the Ortiz PVAT score vis-à-vis the dominant language method using the WMLS-III English vs. Spanish scores similar to the comparisons conducted by Ortiz and Wong (2020b). Furthermore, future validity studies involving different clinical populations (e.g., individuals with language disorder, learning disabilities, attention-related disorders, mood disorders, or intellectual disability as well as those considered neurotypical) can be conducted to compare the sensitivity and specificity between the Ortiz PVAT and the

WMLS-III in the identification of disorder or disability which can impact language functioning.

Chapter VI

Conclusion and Practical Implications

Regardless of the above limitations, the current study added strong evidence to the generalizability and validity of a relatively new measure for assessing second language acquisition for ELs. It also advanced the knowledge on test fairness for ELs, in relation to the importance of incorporating the variable of language experience in the development of norms for true peer group comparison. The use of test norms that control for differences in ELs' different language developmental background presents a step toward culturally competent practice and social justice in resolving the issue of overrepresentation of ethnic and linguistic minorities in special education.

In their literature review and neuropsychological assessment case studies of five EL children who were evaluated in a pediatric medical setting, Canas et al. (2020) strongly advised against making face value assumptions and interpretations about bilingual children's abilities when using standardized tests that compare test takers' performances to that of monolingual children because ELs are at a disadvantage that is not accounted for in their scores and a low score may reflect limited language proficiency rather than an inherent ability-related deficit. In line with recommendations from researchers mentioned in the literature review section (e.g., Flanagan & Ortiz, 2001; Rhodes et al., 2005; Cormier et al., 2022; Ortiz & Wong, 2020a, Ortiz & Wong, 2022), Canas et al. (2020) urge clinicians to conduct a careful consideration of a child's linguistic, educational, and other relevant histories (e.g., personal histories of trauma, discrimination, and/or immigration, access to resources) when making decisions regarding test selection and diagnostic impressions. In one of the case studies, an

English-dominant Spanish-speaking 3rd grade student who has been enrolled in a 50/50 dual language program since preschool was referred for a neuropsychological evaluation, and testing indicated that despite the provision of special education services including resource support in math and reading, speech/language therapy (in English only), and various accommodations, the student exhibited global deficits in receptive and expressive language as well as literacy-based academic achievement in both English and Spanish and in math. The case illustrated that some children may spend years in an academic environment that does not fully meet their specific needs, and the result can be harmful. Therefore, the authors emphasized the importance of matching the right kind of educational program, intervention services and/or accommodations to each unique case. Consider the Spanish-speaking students in the school district reported in Ortiz and Wong (2020b). Misdiagnosis/misclassification of language disorders or disabilities based on the "dominant language" approach can potentially lead to the wrong type of educational programs or services being recommended for a potentially whopping 86% of the students referred, let alone the stigma from the diagnosis or classification on the student and their families. Moreover, the time and resources spent on the unnecessary testing of EL students who were actually making progress as expected for an EL with their specific amount of exposure to English can potentially preclude or delay the access of services for students who truly warrant them. This is precisely why the Ortiz PVAT presents a pivotal step toward fair and true peer group comparison, by considering ELs' differential language background, to assist in differentiating between "disorder" and "difference."

Table 1. Demographic Characteristics of the Monolingual English Speaker and English Learner Subsamples

	English Speaker (N = 25)			English Learner (N = 201)		
Demographic Characteristic	Min	Max	M(SD)	Min	Max	M(SD)
Age	2:9	18:0	8:10 (4:7)	2:7	18:11	8:2 (3:6)
Grade	Pre-school	12th	3rd	Not Enrolled	College/ University Level	2nd
Proportion of Lifetime Exposure to English	100%	100%	100% (0%)	2%	100%	44% (26%)

Table 2. Frequency Breakdown by Grade in Each Subsample

	English Speaker		English Learner	
Grade	Frequency	Percent	Frequency	Percent
Not enrolled	0	.0	2	1.0
Preschool	6	24.0	67	33.3
K	3	12.0	9	4.5
1st	2	8.0	9	4.5
2nd	2	8.0	25	12.4
3rd	3	12.0	25	12.4
4th	1	4.0	17	8.5
5th	2	8.0	16	8.0
6th	0	.0	8	4.0
7th	0	.0	7	3.5
8th	1	4.0	9	4.5
9th	0	.0	2	1.0
10th	2	8.0	3	1.5
11th	2	8.0	1	.5
12th	1	4.0	0	.0
College/University	0	.0	1	.5
Total	25	100.0	201	100.0

Table 3. Frequency Breakdown of First Home Languages within the English Learner Sample

Language Group	First Home Language	*N*	Percent
Spanish and Spanish Creole	Spanish	94	46.8
Other Indo-European Languages	Nepali	33	16.4
	Russian	6	3.0
	French	3	1.5
	Polish	21	10.4
	Montenegrin	1	0.5
	Albanian	3	1.5
	Bulgarian	1	0.5
	Subtotal	*68*	*33.8*
Asian and Pacific Island Languages	Uzbek	2	1.0
	Chinese/Mandarin	7	3.5
	Subtotal	*9*	*4.5*
All Other Languages	Arabic	16	8.0
	Hebrew	1	0.5
	Wolof	1	0.5
	Subtotal	*18*	*9.0*
English		6	3.0
Omitted		6	3.0
Total		201	100.0

Table 4. Correlations between Ortiz PVAT Scores and Various Demographic Variables

Variable	1	2	3	4	5	6	7	8
1. Raw Score	--							
2. Standard Score (English Speaker Norms)	.60**	--						
3. Standard Score (English Learner Norms; *n* = 201)	.52**	.92**	--					
4. Gender	.06	.05	.02	--				
5. Age in Years:Months	.70**	-.09	-.23**	.06	--			
6. Grade	.72**	-.05	-.18**	.07	.98**	--		
7. Major Language Group	-.05	.11	.00	.09	-.13	-.11	--	
8. Lifetime Exposure to English	.64**	.37**	.19**	.15*	.46**	.48**	.32**	--

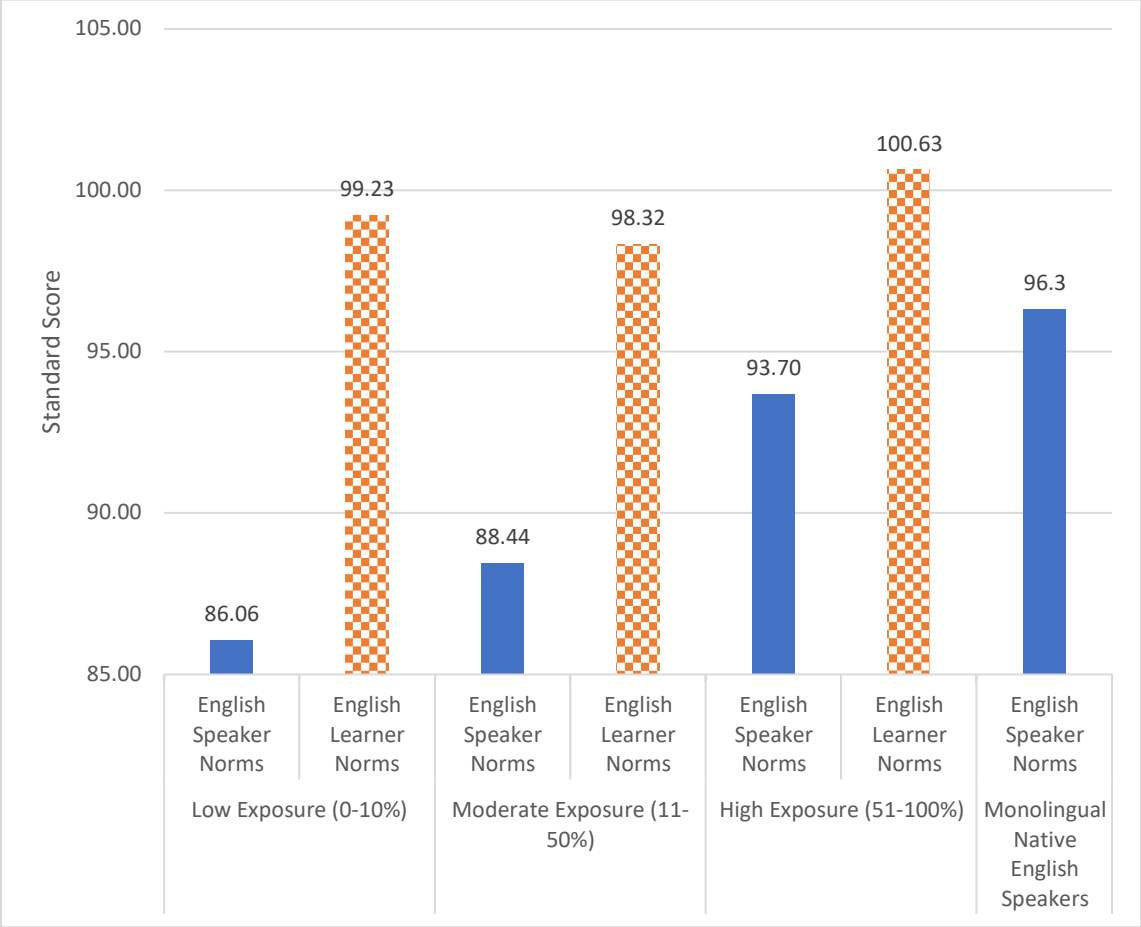
Note: N = 226 except where indicated. ** Correlation is significant at the .01 level (2-tailed). * Correlation is significant at the .05 level (2-tailed).

Table 5. Summary of Hierarchical Regression Analysis for Variables Predicting Ortiz PVAT Test Performance ($N = 200$)

Variable	β	R^2
Step 1		.66***
Age	.67***	
Step 2		.79***
Age	.37***	
Lifetime Exposure to English	.52***	
		$\Delta R^2 = .18$***

*** $p < .001$

Figure 1. Comparison of Mean Standard Scores Across Groups



References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014*). Standards for educational and psychological testing.* American Educational Research Association.

Canas, A., Edgar, V. B., & Neumann, J. (2020). Practical considerations in the neuropsychological assessment of bilingual (Spanish-English) children in the United States: Literature review and case series. *Developmental Neuropsychology, 45*(4), 211-231. https://doi.org/10.1080/87565641.2020.1746314

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Cormier, D. C., Bulut, O., McGrew, K., & Kennedy, K. (2022). Linguistic influences on cognitive tests performance: Examinee characteristics are more important than test characteristics. *Journal of Intelligence, 10*(8), 1-12. doi:https://doi.org/10.3390/jintelligence10010008

Cormier, D. C., McGrew, K., & Ysseldyke, J. (2014). The influences of linguistic demand and cultural loading on cognitive test scores. *Journal of Psychoeducational Assessment, 4*, 1-14. doi:10.1177/0734282914536012

Cummins, J. (1981). Age on arrival and immigrant second language learning in Canada: A reassessment. *Applied Linguistics, 11*, 132-149. doi:10.1093/applin/2.2.132

Cummins, J. (1984). *Bilingualism and special education: Issues in assessment and pedgogy.* Multilingual Matters Ltd.

Dietrich, S., & Bauman, K. (2019). The association between household and community characteristics and children's acculturation. U.S. Census Bureau.

Dunn, L. M., & Dunn, D. M. (2007). *Peabody picture vocabulary test, fourth edition.* Pearson, Inc.

Dynda, A. M. (2008). The relation between language proficiency and IQ test performance. Unpublished manuscript. St. John's University.

Fisher, D., & Frey, N. (2012). *The school leader's guide to English learners.* Solution Tree.

Flanagan, D. P., & Ortiz, S. O. (2001). *Essentials of cross-battery assessment.* John Wiley & Sons, Inc.

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2013). *Essentials of cross-battery assessment (3rd ed.).* John Wiley & Sons, Inc.

Ford, D. (2012). Culturally different students in special education: Looking backward to move forward. *Exceptional Children, 78*, 391-405. doi:10.1177/001440291207800401

García, A. (2022). Controlling for developmental language differences in Spanish-English learners: A comparison of the Ortiz Picture Vocabulary Acquisition Test and Woodcock Muñoz Language Survey-Third Edition. Unpublished doctoral dissertation, St. John's University.

Goldstein, B. A. (2012). *Bilingual language development and disorders in Spanish-English speakers* (2nd ed.). Paul H. Brookes Publishing Co.

Merriam-Webster. (n.d.). *Merriam-Webster Online Dictionary*. Retrieved from https://www.merriam-webster.com/dictionary/diaper

New York State Education Department & Harcourt. (2006). *New York State English as a Second Language Achievement Test: NYSESLAT Spring 2006 school administrator's manual.* NYSED & Hartcourt.

Ortega, L. (2009). *Understanding second language acquisition*. Routledge.

Ortiz, S. O. (2014). Best practices in nondiscriminatory assessment. In P. Harrison, & A. Thomas, *Best Practices in School Psychology VI* (pp. 61-74). National Association of School Psychologists.

Ortiz, S. O. (2018). *Ortiz picture vocabulary acqusition test technical manual.* Multi-Health Systems, Inc.

Ortiz, S. O., & Flanagan, D. (1998). Gf-Gc cross-battery interpretation and selective cross-battery assessment: Referral concerns and the needs of culturally and lingusitically diverse populations. In K. S. McGrew, & D. Flanagan, *The intelligence test desk reference (ITDR): Gf-Gc cross-battery assessment* (pp. 401-444). Allyn & Bacon.

Ortiz, S. O., & Wong, J. Y. T. (2020a). Psychoeducational assessment of culturally and linguistically diverse preschool children. In V. C. Alfonso, B. A. Bracken, & R. J. Nagle (Eds.). *Psychoeducational assessment of preschool children* (pp. 346-364). Routeledge.

Ortiz, S. O., & Wong, J. Y. T. (2020b, October). Fairness in tests and test score interpretation with English learners. *The Score*. Retrieved from https://www.apadivisions.org/division-5/publications/score/2020/10/assessing-english-learners

Ortiz, S. O., & Wong, J. Y. T. (2022). Addressing theoretical, empirical, and practical deficiencies when testing English learners. In K. Geisinger & J. Jonson (Eds.), *Fairness in educational and psychological testing: Examining theoretical, research, practice, and policy implications of 2014 Standards*. American Educational Research Association.

Ortiz, S. O., Piazza, N., Ochoa, S., & Dynda, A. (2018). Testing with culturally and linguistically diverse populations: New directions in fairness and validity. In D. P. Flanagan, & E. McDonough (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (pp. 684-712). The Guilford Press.

Rhodes, R., Ochoa, S. H., & Ortiz, S. O. (2005). *Assessment of culturally and linguistically diverse students: A practical guide.* The Guildford Press.

Schrank, F. A., McGrew, K., & Mather, N. (2014). *Woodcock-Johnson IV.* Riverside Publishing.

Sotelo-Dynega, M., Ortiz, S. O., Flanagan, D. P., & Chaplin, W. (2013). English language proficiency and test performance: Evaluation of bilinguals with the Woodcock-Johnson III Tests of Cognitive Abilities. *Psychology in the Schools, 50*, 781-797. doi:10.1002/pits.21706

U.S. Department of Education. (2017). *IDEA Individuals with Disabilities Education Act*. Retrieved from Sec. 300.304 Evaluation procedures: https://sites.ed.gov/idea/regs/b/d/300.304

Umbel, V. M., Pearson, B. Z., Fernandez, M. C., & Oller, D. K. (1992). Measuring bilingual children's receptive vocabularies. *Child Development, 63*(4), 1012-1020. doi:10.1111/j.1467-8624.1992.tb01678.x

United States Census Bureau. (2015). *Census Bureau reports at least 350 languages spoken in U.S. homes [Press release].* Retrieved from https://www.census.gov/newsroom/press-releases/2015/cb15-185.html

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

Vygotski, L. S. (1986). *Thought and language*. The MIT Press.

Wechsler. (2003). *Wechsler intelligence scale for children - Fourth Edition (WISC-IV).* The Psychological Corporation.

Wechsler, D. (1999). *Wechsler abbreviated scale of intelligence for children.* Pearson.

Wechsler, D. (2014). *Wechsler intelligence scale for children - fifth edition (WISC-V).* Pearson, Inc.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III tests of cognitive abilities.* Riverside.

Woodcock, R. W., McGrew, K., & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement.* Riverside Publishing.

Woodcock, R. W., Munoz-Sandoval, A., Reuf, M., & Alvarado, C. (2005). *Woodcock-Muñoz language survey-revised (WMLS-R).* Riverside.

Woodock, R. W., Alvarado, C. G., Ruef, M. L., & Schrank, F. A. (2017). *Woodcock-Muñoz language survey III.* Hougton Mifflin Harcourt.

Vita

Name	*Jane Yan Ting Wong*
Baccalaureate Degree	*Honors Bachelor of Science* *University of Toronto,* *Toronto, Canada* *Major: Psychology*
Date Graduated	*May 2003*
Other Degrees and Certificates	*Master of Arts* *York University* *University of Toronto,* *Toronto, Canada* *Social/Personality* *Psychology*
Date Graduated	*January 2009*
	Master of Science *St. John's University,* *New York* *School Psychology*
Date Graduated	*May 2021*