Bilingualism: Language and Cognition 21 (2), 2018, 328–339 © Cambridge University Press 2017 doi:10.1017/S1366728917000074

Difficulties using standardized tests to identify the receptive expressive gap in bilingual children's vocabularies*

TODD A. GIBSON
Louisiana State University
LINDA JARMULOWICZ
University of Memphis
D. KIMBROUGH OLLER
University of Memphis

(Received: April 28, 2016; final revision received: February 1, 2017; accepted: February 13, 2017; first published online 13 March 2017)

Receptive standardized vocabulary scores have been found to be much higher than expressive standardized vocabulary scores in children with Spanish as L1, learning L2 (English) in school (Gibson et al., 2012). Here we present evidence suggesting the receptive-expressive gap may be harder to evaluate than previously thought. We compared the performance of 116 six-year-old Spanish–English bilingual children in the US to 30 monolingual Spanish-speaking peers in Mexico across two Spanish-language standardized picture naming tests and one standardized picture pointing test. The performance of 134 monolingual English-speaking peers was compared using similar English-language tests. Results revealed the presence and magnitude of a receptive-expressive gap was largely dependent on the tests used. These discrepant results likely exist because widely-used standardized tests do not offer comparable normed scores. We review possible test norming practices that may have contributed to these results and suggest guidelines to determine a meaningful receptive-expressive gap for bilingual children.

Keywords: bilingualism, second-language learning, vocabulary learning, school-age children, standardized testing

Introduction

A number of researchers have cautioned against the use of monolingual norms for bilingual test takers (Armon-Lotem, de Jong & Meir, 2015; Hoff, Rumiche, Burridge, Ribot & Welsh, 2014; Kohnert, 2013). Language exposure is a key variable in this consideration. Unlike their monolingual peers, bilingual children live in two language contexts. A bilingual child who is exposed to each of her two languages 50% of the time has half as much language exposure to each of those languages as a monolingual peer has to a single language. Therefore, it is likely that bilingual children will have language-specific vocabularies that are smaller than those of their monolingual peers (Patterson & Pearson, 2012). Additionally, some words will be used primarily in contexts where the first language (L1) is spoken and other contexts where the second language (L2) is spoken. Vocabulary, therefore, likely will be distributed across the two languages (e.g., *thimble* might be a word used at home in L1 and *protractor* might be a word used at school in L2; Oller, Pearson & Cobo-Lewis, 2007). The English vocabulary items known by a group of bilingual children, therefore, might differ from items known by a group of monolingual English-speaking peers. Therefore, vocabulary tests normed on monolingual children, even if children are matched demographically and regionally, likely are not appropriate for bilingual test takers.

The admonition that monolingual norms should not be used to assess bilingual children has been folded into best practices positions by governing bodies for speech-language therapy in the US (American Speech, Language, Hearing Association, n.d.), UK (Royal College of Speech and Language Therapists, 2007), Australia (Speech Pathology Australia, n.d.), and others. Indeed ASHA has posited that standardized tests normed on monolingual speakers can be used for descriptive purposes only. Despite these cautions, studies demonstrate that speech-language therapists in English-speaking countries regularly do not implement these best practices (Caesar & Kohler, 2007; Stow & Dodd, 2003; Williams & McLeod, 2012). This often is due to a lack of appropriate bilingual measures (Caesar & Kohler, 2007). To overcome this lapse, several groups, including researchers and test makers, have attempted to develop methods to tease apart language differences from language disorders in bilingual children. For example, some published omnibus language tests for Spanish speakers such as the *Preschool Language Scales - Fifth Edition, Spanish* (Zimmerman, Steiner & Pond, 2012) and the *Clinical Evaluation of Language Fundamentals - Fourth Edition, Spanish* (Wiig, Semel & Secord, 2006) have attempted to improve their bilingual validity by including in their norming samples more US speakers from representative geographic areas

* This research was supported by grants from the National Institutes of Health, National Institute of Child Health & Human Development (R01 HD046947 to D. Kimbrough Oller, Principal Investigator, and R01 HD44923 to Claude Goldenberg, Principal Investigator)

Address for correspondence: Todd A. Gibson, Louisiana State University, 84 Hatcher Hall, Baton Rouge, LA 70803, USA
toddandrewgibson@lsu.edu

and levels of socioeconomic diversity. Additionally, the *Bilingual English-Spanish Assessment* (BESA; Peña, Gutiérrez-Clellen, Iglesias, Goldstein & Bedore, 2014) has norms based exclusively on Spanish–English bilingual children in the US and includes both English and Spanish subtests.

Because bilingual children with typical development (TD) present with many linguistic behaviors that mimic those of monolingual children with language impairment (Oller et al., 2007), it can be difficult to disentangle language differences from disorders. A large multinational European Union project recently proposed a series of approaches to accomplish this differentiation. First, they focused on language processing rather than language knowledge (Armon-Lotem et al., 2015), which has been suggested by Kohnert (2010) and others. For example, instead of measuring children's repertoire of vocabulary items, the clinician focuses on children's ability to process phonological information by repeating nonwords (i.e., repeating strings of speech sounds that could be words but are not; Chiat, 2015). Instead of measuring children's ability to generate novel complex sentences, the clinician measures children's ability to repeat complex sentences (Marinis & Armon-Lotem, 2015). And instead of focusing on the language-specific lexical items or grammatical constructions in narratives, the clinician focuses on the structure of the narrative itself (e.g., whether the narrative includes settings, events, characters, etc.; Gagarina, Klop, Tsimpli & Walters, 2016). Second, in addition to a focus on language processing over language knowledge, these researchers stressed the importance of using bilingual measures that are constructed to be comparable across languages. Co-constructing items across languages so that they have similar levels of testing difficulty ensures that differences in performance across languages are the result of real linguistic ability rather than an artifact of the tests.

Another approach to measuring language processing rather than language knowledge and thus disentangling bilingualism from impairment is known as dynamic assessment (Miller, Gillam & Peña, 2001). This approach typically implements a test, teach, retest design. For example, in the first step of dynamic assessment, a bilingual child might be administered a vocabulary test. In the second step, the child might be taught the same vocabulary items she was tested on. In the third step, the child might be re-tested to determine not only if the vocabulary items were learned but if the teaching was effortful (e.g., the clinician provided few or continuous examples) and if the child was responsive (e.g., the child required little support or constant support). This approach is not limited to vocabulary (Camilleri & Law, 2007) but can be applied to grammar (Olswang & Bain, 1996), narrative (Gillam, Peña & Miller, 1999), and phonology (Glaspey & Stoel-Gammon, 2007) as well.

Although most current methods of assessment of bilingual children encourage a focus on processing over knowledge, some researchers suggest that adapting knowledge-based tests can reveal useful linguistic information (Gross, Buac & Kaushanskaya, 2014). For example, it likely is less meaningful that a child possesses a language-specific word for a concept than that she has the concept at all. Pearson, Fernández & Oller (1993) proposed conceptual scoring as a way to measure concepts independent of the languages in which those concepts are encoded as a way to minimize test bias for bilingual children. For example, imagine a child with both Spanish and English words for the concept chair (*silla* in Spanish) but only the Spanish word for the concept pillow (*almohada* in Spanish). If tested only in English, it would appear that the child does not have a concept of pillow. However, if both English and Spanish words are accepted as correct, then the child clearly has the pillow concept. Indeed, this bilingual child would have more lexical items in her repertoire, three in this case (*chair*, *silla*, *almohada*), than her monolingual peer, who would have only two (*chair* and *pillow*).

Another possible approach to assess bilingual children is to include monolingual norm-referenced measures in the language of evaluation but to adjust cut-off points using a sliding scale based on language exposure (Thordardottir, 2015). This approach is only appropriate for children learning two languages since birth. Thordardottir (2011) found that 5-year old children who were essentially balanced in their exposure to two languages across a lifetime scored between .5 and 1 standard deviation below the mean on a standardized vocabulary test, but still within normal limits. Performance diminished as experience in the targeted language declined. Though imperfect, until more data is obtained, clinicians might use a sliding scale approach to inform their diagnoses of language impairment.

The receptive-expressive gap

Bilingual children with TD have a profile of linguistic behaviors that often mimics monolingual children with language impairment (Paradis, 2010). One such behavior is the receptive-expressive gap in which bilingual children perform better on receptive than expressive standardized tests (Oller et al., 2007). Such a gap, if real, would not be predicted, especially for a language which a child has been speaking since birth. For example, Leonard (2009) highlighted that many norm-referenced, standardized language tests are constructed to provide separate receptive and expressive standard scores. Such test construction, he argued, implies and formalizes a distinction between these language modalities that may not be real, since children with expressive language

https://doi.org/10.1017/S1366728917000074 Published online by Cambridge University Press

difficulties usually also experience receptive language difficulties. Nonetheless, tests like the *Clinical Evaluation of Language Fundamentals, Fourth Edition - Spanish* (Wiig et al., 2006), the *Preschool Language Scales - Fifth Edition, Spanish* (Zimmerman et al., 2012), and the *Test of Early Language Development - Third Edition, Spanish* (Ramos, Ramos, Hresko, Reid & Hammill, 2006), provide separate standard scores for receptive and expressive language. These tests also provide ways to determine whether discrepancies between receptive and expressive language are meaningful. Cases in which expressive language scores are meaningfully lower than receptive language scores frequently are interpreted as expressive language disorders in accordance with guidelines similar to those from the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (American Psychiatric Association, 2000, p. 58), which explicitly requires the use of standardized scores for the diagnosis. In contrast, therefore, a common expectation for children with typical language development is that receptive and expressive standard scores should be similar, not only at the group level but at the individual level. This expectation has been codified by the World Health Organization's (2005) ICD-10 system, which provides distinct codes for expressive and receptive language disorders.

A discrepancy between receptive and expressive language performance does not necessarily indicate a problem, however. A review of the examiners' manuals of prominent English language tests reveals that a small proportion of the individuals in norming samples themselves perform better in receptive than expressive language (see Gibson, Peña & Bedore, 2014a). A discrepancy of 7 to 14 standard score points (on tests normed with a mean of 100 and standard deviation of 15) occurred in about 12 percent of the participants within several norming samples. For example, comparing the receptive and expressive standard scores of the *Preschool Language Scales - Fifth Edition* (PLS-5; Zimmerman, Steiner & Pond, 2012) resulted in receptive standard scores 7 points higher than expressive standard scores within the monolingual English-speaking norming sample. A similar pattern can be found within other standardized tests that include receptive and expressive scores (*Oral and Written Language Scales*, Carrow-Woolfolk, 1995; *Clinical Evaluation of Language Fundamentals - Fourth Edition*, Semel, Wiig & Secord, 2003).

Because this gap occurs infrequently, Gibson, Peña and Bedore (2014b) asserted that at the group level, a statistically significant difference between receptive and expressive standard scores should be treated as a receptive-expressive gap. Such a gap can be observed embedded in the reported standard scores in a number of studies (Barnett, Yarosz, Thomas & Jung, 2007; Miccio, Tabors, Páez & Hammer, 2005; Oller & Eilers, 2002; Uchikoshi & Maniates, 2010; Windsor & Kohnert, 2004). For example, Gross et al. (2014) administered to 6 year olds the *Test de Vocabulario en Imágenes Peabody* (TVIP, Dunn, Padilla, Lugo & Dunn, 1986) and the *Vocabulario Sobre Dibujos* subtest of the *Bateria III Woodcock-Muñoz Pruebas de Aprovechamiento* (WMPA-III; Muñoz-Sandoval, Woodcock, McGrew & Mather, 2005) to measure Spanish receptive and expressive vocabulary, respectively. The TVIP is a picture pointing task and the WMPA-III is a picture naming task. Results revealed a receptive-expressive gap of 21.7 points for simultaneous bilinguals and 19.1 for sequential bilinguals. These outcomes were remarkably similar to the 21 point gap we identified in an earlier study (Gibson et al., 2012) for a group of 5 year old children that included both simultaneous and sequential bilinguals.

Because the Spanish–English bilingual children in the studies reported above had TD, we interpreted in previous work (Gibson, Oller, Jarmulowicz & Ethington, 2012) that the receptive-expressive gap in Spanish was an indicator of either first language (L1) inhibition or the differential activation of the second language (L2). We reasoned that in the context of the US, Spanish-speaking children inhibited their L1 (Spanish) to allocate cognitive resources for learning L2 (English). Alternatively, because these children's focus was on L2 learning, they may have activated L2 to a greater degree than L1, thus mimicking L1 inhibition. We speculated that this inhibition/differential activation had minimal impact on the easier receptive vocabulary task but had a significant impact on the more difficult expressive vocabulary task, resulting in a receptive-expressive gap. However, the tests themselves might have contributed to the gap.

A possible psychometric contribution to the receptive-expressive gap in Spanish

For each of the above investigations that identified a receptive-expressive gap in Spanish, researchers used the TVIP to measure receptive vocabulary but a variety of picture-naming tests to measure expressive vocabulary. In Gibson et al. (2012), we administered the TVIP and compared it to the *Vocabulario Oral* (oral vocabulary) subtest of the Wo*odcock Language Proficiency Battery – Revised, Spanish Form* (WLPB-RS; Woodcock & Muñoz-Sandoval, 1995), which is a traditional picture-naming task. As part of a larger study, however, we subsequently administered the picture-naming task (the *Vocabulario Sobre Dibujos* subtest) from the *Woodcock-Muñoz Language Survey – Revised Spanish* form (WMLS-RS; Woodcock, Muñoz-Sandoval, Ruef & Alvarado, 2005) and were able to compare the same children's results on both the WLPB-RS and WMLS-RS. Results of the WMLS-RS for the children in Gibson

https://doi.org/10.1017/S1366728917000074 Published online by Cambridge University Press

et al. (2012) had not been analyzed at the time of our previous publication on the receptive-expressive gap. Subsequent analysis revealed that the receptive-expressive gap we had identified using the WLPB-RS was greatly diminished when the receptive outcomes were compared with results from the WMLS-RS. But to interpret the apparent discrepancy, we deemed it important to conduct the same tests with a group of true Spanish monolinguals, for whom we did not have the relevant data.

Knowledge of discrepancies across tests is valuable because, although it is generally accepted that monolingual norms should not be used to assess bilingual children, it is not clear exactly how bilingual children deviate from such norms. The purpose of the current study was to report our discrepant findings using these similar picture naming tasks and to explore the possible reasons for these differences. To rule out the possibility that the tests were inappropriately normed for monolingual Spanish speakers, we administered the same tests, the TVIP, WLPB-RS, and the WMLS-RS to monolingual first graders in Guadalajara, Mexico. In what follows, we report these efforts and offer some speculations for the results.

Method

Participants

Participants included 116 Spanish–English bilingual kindergarten children from the US, 30 Spanish first-grade monolinguals from Mexico, and 134 monolingual English-speaking kindergarteners from the US. A subset of the US children's results were reported in Gibson et al., (2012). No formal Spanish language educational support was provided by the school system to the Spanish-speaking children in the US.

Children in the US were treated as bilingual if caregivers reported that Spanish was spoken in the home, whether exclusively or along with other languages. This provided a sample with a wide range of language abilities, including some children who were functionally monolingual in one of their languages. Because we were comparing scores across the TVIP, WMLS-RS, and the WLPB-RS, children were included in the analysis only if they had scores from all three tests. This resulted in an original US Spanish–English bilingual participant pool of 229 children with an average age of 72.93 months, $SD = 4.71$. Because we wanted the Mexican first graders and the Spanish–English bilingual kindergarten children from the US to have similar ages for comparison, we excluded those bilingual children in the US who were 72 months of age or younger. This resulted in a sample of 116 bilingual K children in the US (53 girls and 63 boys) with an average age of 76 months, $SD = 3.77$ at the time of posttest in kindergarten, which occurred in their second semester. The decision to use posttest rather than pretest scores for the US children was validated by the results of a repeated measures ANOVA that found statistically significant changes in standard scores across the school year for the TVIP, $F(1, 114) = 9.15$, $p = .003$, $\eta^2 = .07$, and WLPB-RS, $F(1, 114) = 4.45$, $p = .04$, $\eta^2 = .04$, but not for the WMLS-R, $F(1, 92) = 1.95$, $p = .17$, $\eta^2 = .02$. The Mexican first graders (11 girls and 19 boys) had an average age of 77.67 months, $SD = 3.66$, in the first semester of their first grade school year. There was no statistically significant difference in age between the two groups, $F(1, 144) = 2.14$, $p = .15$.

Both groups of children came from similarly low socioeconomic status (SES) backgrounds based on available information. The average number of years of formal education for the mothers of the bilingual kindergarten group in the US was 8 years, $SD = 2.89$. (There was no data for mother's education for 3 of the 116 children.) Although we did not have data on mother's education for the children from Mexico, census data from Mexico showed that our target school was located in a municipality described as a high *grado de marginación*, or area where opportunities for development were minimal or not present (Consejo Nacional de Población, 2012). The presumed relatively low SES agreed with observations of the native Mexican collaborator who helped recruit the sample.

Measures

Expressive vocabulary

Expressive vocabulary was measured using both the *Woodcock Language Proficiency Battery – Revised Spanish* form (WLPB-RS; Woodcock & Muñoz-Sandoval, 1995) and the *Woodcock-Muñoz Language Survey – Spanish* form (WMLS–RS; Woodcock, Muñoz-Sandoval, Ruef & Alvarado, 2005). Both tests are picture-naming tasks in which the child names a single color picture presented by the tester. Raw scores for both of the tests are the number of accurately named pictures. Both tests produce standard scores with a mean of 100 and standard deviations of 15.

Both the WLPB-RS and the WMLS-RS were developed using similar procedures. The WLPB-RS is an adaptation of its English language counterpart, the *Woodcock Language Proficiency Battery – Revised* (WLPB-R; Woodcock, 1991). The norming data for the WLPB-R was acquired during the norming of the *Woodcock-Johnson Psycho-Educational Battery – Revised* (1989). The examiner's manual for the WLPB-R reports the sampling variables that the test developers attempted to control for with respect to the 3,245 English-speaking participants in the K – 12th grade sample. These included census region, community size, sex, race, Hispanic, and household income.

https://doi.org/10.1017/S1366728917000074 Published online by Cambridge University Press

In order to create Spanish language norms for the WLPB-RS, the test's examiner's manual (Woodcock & Muñoz-Sandoval, 1995) explains the following procedures. First, using Rasch analysis, the test developers created a pool of English language items drawn from the norming sample and calibrated them from easy to difficult. Those English items for which there was a reasonable Spanish counterpart were translated to Spanish (e.g., *authority* was translated to *autoridad*) and used as equating items. The equating items and other items were administered in Spanish to Spanish speakers both inside and outside of the USA. Of the 3,911 subjects who were administered one or more of the tests in Spanish, 1,325 (or 33.8%) of individuals were drawn from a US sample (pre-K through adult; the number of K and first grade Spanish-speaking children in the US sample was not reported separately). From the US sample, 331 were born in the US, with the rest having immigrated to the US.

The test developers attempted to select a functionally monolingual Spanish-speaking US sample. Toward this end, they administered a language history questionnaire to informants for the Spanish-speaking US sample (presumably parents of children) that asked for the percent of time that the participant used Spanish at home, others used Spanish at home, participant used Spanish in informal situations, and participant used Spanish in the classroom. Seventy-five percent reported using Spanish 100% of the time at home, 71% reported that others used Spanish 100% of the time at home, and 60% reported Spanish use 100% of the time in informal settings. Spanish use in the classroom was distributed across four categories, with 13% using Spanish less than 50% of the time, 36% using Spanish 50% of the time, 30% using Spanish 75% of the time, and 22% using Spanish in the classroom 100% of the time. The examiner's manual reported that these results indicated an essentially monolingual Spanish-speaking US sample. In addition to the US sample, over 2000 Spanish-speaking participants were tested in other Spanish-speaking countries. There was no report of the SES of the Spanish-speaking participants.

Using Rasch analysis, the Spanish items were ranked from easiest to most difficult based on the performance of the Spanish-speaking calibration sample. Outliers were dropped from the equating items. Items were then selected for inclusion in the Spanish test version. The test targeted a difficulty gradient, but, as noted in the examiner's manual, the spread between Spanish and English difficulty may have been different. Therefore, in order to guarantee the targeted difficulty gradient, items were selected based on Spanish performance BEFORE equating Spanish and English difficulty scales. The Spanish difficulty scale was rescaled to the English difficulty scale using an equation that was not published but was reported to be based on linear regression. The examiner's manual reports high reliability and validity for the WLPB-RS.

A similar procedure for the WMLS-RS was reported in its examiner's manual (Alvarado, Ruef & Schrank, 2005). The norming data was drawn from the data used to develop the *Woodcock-Johnson III* (Woodcock, McGrew & Mather, 2001). For the English-speaking K – 12th grade sample (4,783 participants), the test developers attempted to match norming variables to the US population. These variables included census region, community size, sex, race, Hispanic, type of school (public, private, or home), father's education, and mother's education.

Using Rasch analysis, English items were ranked from easiest to most difficult, and equating items were translated to Spanish. Equating items and other Spanish items were administered to Spanish-speaking participants inside and outside of the US. Of the 1,157 Spanish-speaking individuals tested in the Spanish calibration, 85 (or 7.3%) were drawn from the US. The rest were drawn from Spanish-speaking countries. Although the names of the testing sites are provided in the examiner's manual, no SES information for the Spanish speakers is provided. Spanish items were rank ordered using Rasch analysis. Equating item outliers were dropped. Means and standard deviations for the equating items were calculated. In the next step, these means and standard deviations were included in the transformation equation (reported in the examiner's manual) to rescale the Spanish item difficulties to the English item difficulties. Items were then selected for inclusion in the Spanish form. This means that, unlike the WLPB-RS, item selection took place AFTER equating Spanish and English difficulty scales. The examiner's manual reports high validity and reliability.

Receptive vocabulary was measured by the *Test de Vocabulario en Imágenes Peabody* (TVIP; Dunn et al., 1986), which provides a standard score with a mean of 100 and standard deviation of 15. The TVIP is a Spanish adaptation of the *Peabody Picture Vocabulary Test – Revised* (PPVT-R; Dunn & Dunn, 1981). The TVIP was developed by procedures not dissimilar to those for the expressive vocabulary tests. First, a pool of 350 English items from the PPVT-R were translated to Spanish by two translators. This list was then reviewed by experts in Mexico and Puerto Rico, and items that were judged not to be universally used by Spanish speakers were replaced by more universal items. To standardize the test for test takers ages 2 years to 17 years 11 months, the remaining 167 items were administered to 1,219 children in the Mexico City metropolitan area and 1,488 children in Puerto Rico. (The age 6 standardization sample, however, included only 243 children; 96 from Mexico and 147 from Puerto Rico.) Field testing for standardization took place from September 1981, until February 1983. Because children from high socioeconomic backgrounds were overrepresented in the sample, especially at younger ages, the test makers weighted the scores to fit the distribution of SES reported in the U.S. Census statistics. Scores from the

https://doi.org/10.1017/S1366728917000074 Published online by Cambridge University Press

Table 1. *Standard score Mean (Standard Deviation) for receptive and expressive vocabulary tests for US bilinguals, US monolinguals, and Mexican monolinguals.*

	WLPB-REnglish	WLPB-R Spanish	WMLS-R Spanish	TVIP	PPVT
US Bilinguals (N = 116)	62.56^{c} (20.57)	61.96^{a},^{b} (21.47)	87.03^{a} (12.42)	89.34^{b}(15.06)	71.09^{c} (15.64)
Mex Monolinguals(N = 30)	NA	96.83 (16.48)	95.77^{d} (10.26)	104^{d} (9.68)	NA
US Monolinguals(N = 134)	100.36^{e} (14.75)	NA	NA	NA	92.38^{e} (12.63)

Notes. Superscripts indicate a statistically significant difference.

Mexico City and Puerto Rico groups then were combined and Rasch analysis was used to calibrate the items for difficulty. This process led to 42 items being eliminated, resulting in a total of 125 items in final version of the TVIP. The examiner's manual reports high validity and reliability.

Procedures

As part of a larger study of bilingualism, Spanish–English bilingual children in the US were administered a battery of Spanish and English language tests at the beginning and the end of the kindergarten school year In the current analysis, we report Spanish and English vocabulary results. Children were tested in their schools, and attempts were made to secure administration in quiet areas. Testers were fluent in the language of testing. For each score reported here, the same tester administered both the WLBP-RS and WMLS-RS. Order of testing was based on convenience which resulted in 35 occasions in which WMLS-RS was administered between one and 14 days before the WLPB-RS, seven days in which they were administered on the same day, and 74 occasions in which the WLPB-RS was administered one to 28 days before the WMLS-RS. Order of administration was not documented on the 7 days in which both tests were administered. A control group of monolingual English-speaking kindergarteners drawn from the same classrooms as the bilingual participants were administered the WLPB-R (English) and the PPVT-III (English).

After identifying the discrepancy in the outcomes of the WLPB-RS and WMLS-RS for the Spanish–English bilingual children, we sought to test monolingual Spanish-speaking children in Mexico to determine if the same pattern of results occurred. Testing in Mexico was performed by a native Mexican, Spanish-speaking tester who had also participated in testing children in the US bilingual group. All children from the Mexico group attended the same school and were in their first semester of first grade. They were tested at school, and attempts were made to minimize distractions during testing. Children from Mexico were administered only three tests in the following order: WLPB-RS, TVIP, followed by WMLS-RS. For each child, all testing was undertaken in the same session.

Results

Descriptive statistics are reported in Table 1. We performed a series of paired sample *t*-tests with a Bonferroni correction for multiple comparisons to contrast test scores within each group. We additionally calculated Cohen's *d* effect sizes for dependent groups. For the US bilingual children, there was a statistically significant difference of more than 25 standard score points between the WLPB-RS and WMLS-RS outcomes, $t(115) = 17.06$, $p < .001$, $d = 1.86$. The effect size can be characterized as very large. Further, the WLPB-RS scores differed from the TVIP scores by more than 27 points, $t(115) = 15.04$, $p < .001$, $d = 1.46$, again a very large effect size, suggesting a very large receptive-expressive gap. No statistically significant difference, however, occurred (only two points, a very small effect size) between the TVIP and WML-RS scores, $t(115) = 1.91$, $p = .06$, $d = .18$, suggesting either no receptive-expressive gap, or a very small one for the overall group of bilingual children. The pattern of better performance on the WMLR-RS than on the WLPB-RS in the bilingual children applied strongly regardless of order of administration. In English testing, US bilingual children presented with a 9 point receptive-expressive gap, with PPVT scores better than WLPB-R scores, $t(115) = 6.19$, $p < .001$, $d = .60$, at a moderate effect size.

The large difference between the outcomes for monolingual and bilingual children on the expressive vocabulary tests is also illustrated in Figure 1. These differences illustrate incontrovertibly that something is amiss with one or both of these tests, since they yield dramatically discrepant results for different groups. The differences confound the interpretation of any possible receptive-expressive vocabulary gap that could be based on results of these tests.

Because children from the US were treated as bilingual if Spanish was spoken in the home, there was a wide range of fluency levels within each language. In order to explore a possible role for fluency on the receptive-expressive gap,

Table 2. *Standard score Mean (Standard Deviation) for US bilinguals by language knowledge group.*

	WLPB-REnglish	WLPB-R Spanish	WMLS-R Spanish	TVIP	PPVT
Spanish Dominant (N = 21)	51.90^{d} (15.54)	67.95$^{a, b}$ (17.85)	91.38$^{a, c}$ (8.85)	100.57$^{b, c}$ (5.83)	60.76^{d} (8.85)
Balanced Bilinguals(N = 19)	74.47 (17.64)	48.42$^{e, f}$ (16.36)	77.21^{e} (16.04)	78.37^{f} (11.91)	82.84 (9.31)

Notes. Superscripts indicate a statistically significant difference.

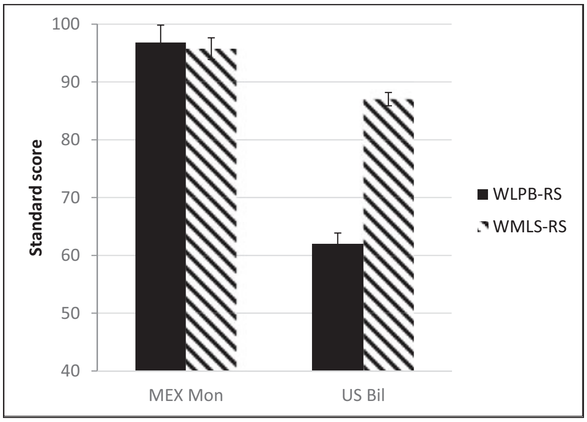

Figure 1. Spanish expressive vocabulary tests: Mexican and US children. Bars indicate Standard Error.

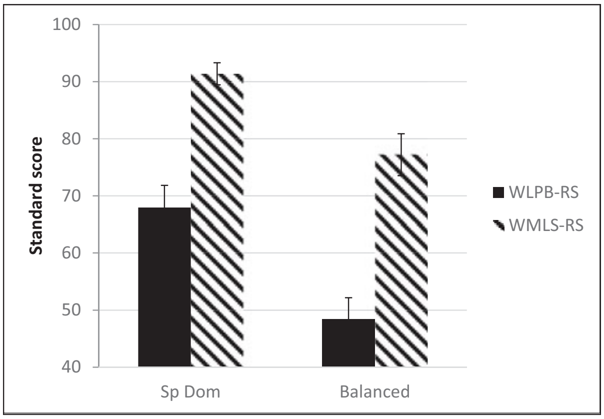

Figure 2. Spanish expressive vocabulary tests: US children by language dominance. Bars indicate Standard Error.

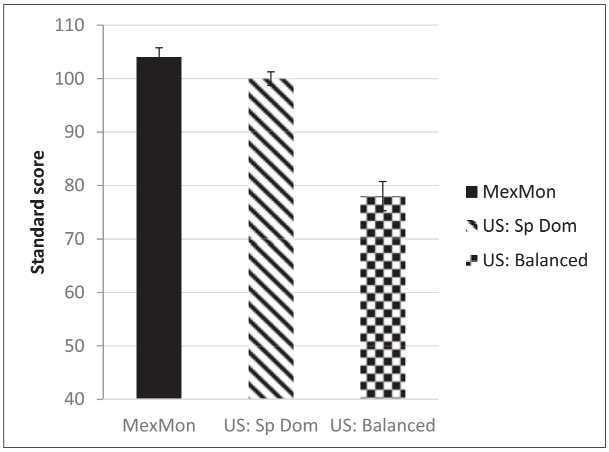

Figure 3. Spanish receptive vocabulary test: US and Mexican children. Bars indicate Standard Error.

we split the bilingual children into two groups based on receptive vocabulary performance in the two languages. If their TVIP score exceeded the median (92 standard score points) and their PPVT score fell below the median (70 standard score points), they were categorized as Spanish dominant. We used this median-split procedure intending to create an English dominant group (English better than Spanish), but, in fact, the resultant group appeared to have more or less balanced receptive vocabulary across their two languages. Therefore, we referred to them as balanced bilinguals. Twenty-one children were included in the Spanish-dominant group and 19 in the balanced bilingual group. Descriptive statistics for each group are reported in Table 2. The differences in Spanish receptive vocabulary between the two groups in comparison with the Spanish monolinguals from Mexico is illustrated in Figure 2.

For the Spanish dominant group, there was a statistically significant difference between WLPB-RS and WMLS-RS outcomes, $t(21) = 7.37$, $p < .001$, $d = 1.92$. The effect size can be characterized as very large. Further, the WLPB-RS scores differed from the TVIP scores by more than 32 points, $t(21) = 8.76$, $p < .001$, $d = 2.76$, again a very large effect size, suggesting a very large receptive-expressive gap. Unlike the overall group of bilingual children, however, the Spanish dominant group showed a statistically significant difference between WMLS-RS scores and TVIP scores, $t(21) = 5.08$, $p < .001$, $d = 1.164$, resulting in a 9 point receptive-expressive gap. These children presented with a 14 point receptive-expressive gap in English testing, $t(21) = 3.20$, $p = .005$, $d = .79$.

For the balanced bilingual group, there was a statistically significant difference between WLPB-RS and WMLS-RS outcomes, $t(18) = 9.56$, $p < .001$, $d = 2.19$. The effect size can be characterized as very large. WLPB-RS scores differed from the TVIP scores by 29.95 points, $t(18) = 6.89$, $p < .001$, $d = 1.61$, again suggesting a very large receptive-expressive gap.

https://doi.org/10.1017/S1366728917000074 Published online by Cambridge University Press

Similar to the overall group of bilinguals, there was not a statistically significant difference between WMLS-RS scores and TVIP scores, $t(18) = .30$, $p = .77$, $d = .07$. Additionally, with the application of the Bonferroni correction, there was no statistically significant receptive-expressive gap in English, $t(18) = 2.17$, $p = .04$, $d = .54$.

Figure 2 illustrates that the large difference between bilingual children's scores on the two expressive language tests applied both to the Spanish dominant and the balanced bilingual groups. Thus, the confound in interpretation of any possible receptive-expressive gap when using these tests as a basis for determining expressive vocabulary appears to affect children across a wide range of Spanish–English balance.

For the Mexican monolingual Spanish-speaking children (in dramatic contrast with the bilinguals), there was NO statistically significant difference between the WLPB-RS and WMLS-RS, $t(29) = .43$, $p = .67$, $d = .09$ (as illustrated in Figure 1). However, there WAS a statistically significant difference between the TVIP and the WMLS-RS, $t(29) = 3.87$, $p = .001$, $d = .71$ and for the TVIP and WLPB-RS, $t(29) = 2.55$, $p = .016$, $d = .50$. Differences of 7–8 points with moderate effect sizes suggested a receptive-expressive gap that one would not normally expect.

For the US monolingual English-speaking children, there was also a statistically significant difference between WLPB-R, $M = 100.36$, $SD = 14.75$, and PPVT-III, $M = 92.38$, $SD = 12.63$, $t(133) = -8.40$, $p < .001$, $d = -.74$, but in the opposite direction, with expressive skills outranking receptive skills, and again at moderate effect size.

Finally the monolinguals had significantly higher scores than bilinguals on every possible comparison (Mono Eng > Bil Eng on both WLPB-R and PPVT-III, and Mono Sp > Bil Sp on WLPB-RS, WMLS-RS, and TVIP).

Summary and Discussion

Results from the current study demonstrate markedly different outcomes depending on the tests used and the child language backgrounds. Significant discrepancies between receptive and expressive tests often are interpreted, whether justified or not, as real differences in abilities (Leonard, 2009). Indeed, monolingual individuals performing substantially lower in expressive compared to receptive language skills often are diagnosed with expressive language disorders. For the Spanish–English bilingual kindergarteners in this study, a substantial discrepancy existed between receptive and expressive skills when comparing the TVIP to the WLPB-RS but this discrepancy greatly diminished when comparing the TVIP to the WMLS-RS. This was true whether children were dominant in Spanish or more or less balanced across Spanish and English vocabulary knowledge. Furthermore, the results on the WLPB-RS and WMLS-RS were significantly different for the Spanish–English bilinguals in the US but not for the Spanish-speaking monolinguals in Mexico.

For the monolingual Spanish-speaking Mexican children, there was no statistically significant difference between the standard scores of the WLPB-RS and WMLS-RS. However, both expressive vocabulary tests resulted in a statistically significant receptive-expressive gap when compared to the TVIP. Because all of the tests were standardized to a mean of 100 and standard deviation of 15, such differences suggest either anomalies in one or more of the tests or anomalies in the current sample. We note that the current sample provides repeated measures comparisons, with each child as a control for himself/herself. The tests we used, on the other hand, were all normed on different samples, so it would seem a possible source of the discrepancies is in the differences among the norming samples or the norming procedures for the various tests.

English vocabulary test scores were also discrepant. For the US bilingual children, receptive vocabulary performance was nine standard score points greater than expressive vocabulary performance. For the US monolingual English-speaking kindergarteners, the reverse occurred. Their expressive vocabulary performance was eight standard score points greater than receptive vocabulary performance. That is, monolingual English-speaking children in this study performed better at naming pictures than pointing to pictures. It appears again that the tests must have had differences in norming samples or in norming procedures.

It is not clear why discrepancies across these tests occurred. However, it is logical to assume that, when comparing any two tests, each point of divergence in the sequence of test development affords an opportunity for discrepant results. In what follows, we provide some speculations for such discrepancies by focusing on the WLPB-RS and WMLS-RS.

Reflecting on each stage in the Spanish equating of the WLPB-RS and WMLS-RS leads us to identify at least five steps where procedures/samples may have diverged between the two tests in such a way as to produce incomparable outcomes. First, the equating samples included different sets of human beings, and this could have been a source of potential error. Neither of the examiner's manuals report demographic information regarding the individuals participating in these samples. Furthermore, the tests were administered in different settings, including different countries, which sometimes did not overlap between tests. When the same country was included in both testing samples, the distribution of participants was widely discrepant. For example, the WLPB-RS included almost five times the proportion of participants from the US (34%) as the

https://doi.org/10.1017/S1366728917000074 Published online by Cambridge University Press

WMLS-RS (7%). Additionally, the local contexts in which the testing took place may have varied widely between tests. It is not clear to what extent if any environmental factors were controlled. Contexts may have varied based on home vs. school settings, private testing rooms vs. public, air conditioned environments vs. non-air conditioned environments, etc. Each of these variables potentially could introduce error. In the current analysis, however, we used a repeated measures design where all participants served as their own controls, and the tests were administered by the same testers in the same schools. Therefore, our study had inherent controls against procedural and sample differences. And there was no reason to believe that the children in our sample were unusual.

Second, not only could a potential source of error have been differences in samples but there may have been differences in the data gathering procedures. For example, although both tests relied on experts to develop the list of to-be-tested Spanish items, it is not clear how these lists may have differed across tests. Spanish equating items were drawn from cognate words, for example *authority-autoridad*. English words that serve as Spanish cognates likely are Latinate words, which typically are relatively low in frequency in English. This might result in unusual and potentially different sets of equating items across the tests. Basic procedures for administering these words also were not reported. For example, it is not clear what stopping rules were applied for either test during data gathering. All children could have been given a prescribed number of words before stopping, the stopping rule could have been based on children's linguistic behavior, or stopping could even have been predicated on children's non-linguistic behavior. Additionally, it is not clear how the final list of words for each test was selected; presumably there were many more items administered than were included in the final published test in both cases, but the original numbers could have differed importantly. Differences across tests on any data gathering procedure may have introduced error that could have produced discrepancies in test results.

Third, it is not clear how or if each of the test's results was verified. A possibility is that estimates were made based on the original larger number of items, many of which were not included in the final list. Ideally, testers would have re-administered the final list of chosen items to children in the previous settings. Whether this occurred, however, is not revealed in the manuals, so differences in verification procedures across tests may have caused discrepancies in results.

Fourth, the mathematical procedures for norming appear to have differed between tests. For example, although the tests used similar statistical procedures, the sequence of events differed. The final item list for the WLPB-RS was selected before equating the English and Spanish difficulty scales, but the final item list for the WMLS-RS was selected after equating the difficulty scales. Differences in statistical procedures introduce the possibility of error that could contribute to discrepancies such as were found in the current analysis. Indeed, we suspect this may have been a significant contributor to the discrepancies found in the current analysis. Of the two tests, the norming sample for the WLPB-RS was likely the more similar to the sample of children in the current analysis since 34% of that norming sample came from the US as opposed to 7% for the WMLS-RS. We would have predicted therefore that between the WLPB-RS and WMLS-RS the former would have resulted in higher scores than the latter, but the opposite was the case. So it remains possible that differing norming samples could have played an important role in the unexpected discrepancies found in the present study, but these outcomes make it difficult to see how.

Finally, more than a decade passed between the original English data collection (which formed a foundation for item selection in the Spanish versions of the tests) and the final publication of the Spanish versions of these tests. There may have been significant dialectal shifts in either or both languages across those years that could have contributed to differences across tests. Furthermore, such a significant time lag could have resulted in procedural drift. For example, later sets of test administrators may have been more efficient than earlier administrators because test-givers learned from earlier mistakes. The variable of time, therefore, may have contributed to the discrepancies that were identified in the current analysis.

The critiques of the norming processes leveled at the expressive vocabulary tests also can apply to the receptive vocabulary test. In all of the studies cited in this paper in which a Spanish receptive-expressive gap was identified in Spanish–English bilingual speakers, the researchers used the TVIP. Therefore, there exists the possibility that this test contributes to the receptive-expressive gap, perhaps by inflating receptive vocabularies in Spanish. The standardization process for the TVIP took place in the early 1980s and lexical items were selected for their universality. Perhaps those items have increased in frequency over the intervening years, providing more exposure to each item and thus increasing the likelihood of children's success.

The outcome for the monolingual English-speaking children in this study, too, might be attributed to psychometric considerations. These children's expressive standard scores were greater than their receptive standard scores. In Gibson et al. (2012), we proposed two possibilities underlying this outcome. First, the norming sample perhaps differed from the tested children culturally and/or dialectally. The study took place in the Southeastern United States, which has a rich tradition

https://doi.org/10.1017/S1366728917000074 Published online by Cambridge University Press

of dialectal variation. Second, the school environment in which the children took the test may have encouraged more talking than where the norming sampling took place. Hadley, Wilcox and Rice (1994) found that teachers' expectations for speaking at school differed across classrooms and age groups. These sorts of expectations regarding talking may have played a role in children's naming performance.

Possible theoretical explanations of the results

Of course, our original proposition explaining the receptive-expressive gap might still hold (Gibson et al., 2012). In the context of L2 learning, Spanish-speaking children might be so focused on learning L2 that L1 becomes either deactivated or inhibited. This deactivation/inhibition might impact the more difficult expressive task to a greater degree than the receptive task, resulting in a receptive-expressive gap. Some words might be more immune to deactivation/inhibition because they are used frequently, while infrequently used words might be more susceptible to deactivation/inhibition. Lexical items across tests might differ in their degree of immunity to deactivation/inhibition, which might help explain the discrepancy between WLPB-RS and WMLS-R scores. However, other psycholinguistic alternatives exist.

Yan and Nicoladis (2009) found that French–English bilingual children in Canada had more difficulty naming pictures than did their monolingual English-speaking peers despite having similar receptive vocabulary sizes. They appealed to Gollan and Acenas (2004) and proposed that the locus of naming difficulties for bilinguals was at the level of lexical access. The mechanism underlying this difficulty is explained by the WEAKER LINKS HYPOTHESIS, which asserts that the links between semantic and phonological representations for bilinguals are weaker than those for monolinguals. The result is a Tip of the Tongue state, in which the speaker knows the word, can even describe the word, but cannot access the phonological form of the word (Brown & McNeill, 1966). Because these semantic-phonological links are sensitive to language experience, some words (e.g., items from the WLPB-RS) might be more susceptible to lexical access difficulties than other words (e.g., items from the WMLS-R). The weaker links hypothesis, therefore, might help explain both the discrepancy between WLPB-RS and WMLS-R as well as the receptive-expressive gap.

An alternative way to identify a bilingual receptive-expressive gap

Identification of a problematic receptive-expressive gap is complex and there currently is no appropriately standardized way to measure it when comparing picture pointing to picture naming tasks. This is especially true when dealing with a language for which no standardized vocabulary tests exist (Thordardottir, 2015). However, as reviewed in the literature, researchers have developed a number of ways to overcome obstacles in disentangling language differences from language disorders (Armon-Lotem et al., 2015). Here, we base a proposal for identifying a meaningful receptive-expressive gap when comparing picture pointing to picture naming tasks by reviewing the results of Gross et al. (2014).

Gross et al. (2014) reported a receptive-expressive gap of 21.7 points for simultaneous Spanish–English bilingual children in the US and 19.1 for sequential Spanish–English bilingual children in the US when participants were required to answer only in Spanish. Through the use of conceptual scoring for both receptive and expressive vocabulary tests, these gaps were reduced to 8.7 and 5.3, respectively. In English testing for these same children, however, the receptive-expressive gap for simultaneous bilinguals rose from 8.3 to 12.3 standard score points and the gap for sequential bilinguals rose from 5.8 to 6.6 standard score points. We interpret the results to mean that when comparing picture pointing tasks to picture naming tasks, conceptual scoring dramatically reduced the gap in the minority language (Spanish) and slightly increased it for the majority language (English). Either way, a gap persisted in both languages.

To our knowledge, Gross et al. (2014), are the only researchers to have reported receptive and expressive standard scores calculated both through language-specific scoring and conceptual scoring on the same set of children. Therefore, what we propose is based on little evidence and should be treated as a tentative guideline until further research is done. With that caveat, we propose a combination of accommodations mentioned in the literature review as a way to identify the receptive-expressive gap in bilingual children. These accommodations are based on comparisons between standardized picture pointing tasks and standardized picture naming tasks and are applied differently based on whether testing occurs in a minority L1 or majority L2.

For a minority language L1, we recommend that standardized scores be calculated based on conceptual rather than language-specific scoring for both receptive and expressive vocabulary. For Gross et al. (2014), a gap of roughly half a standard deviation persisted despite conceptual scoring. Therefore, we propose that a receptive-expressive gap based on conceptual scoring is meaningful and might indicate an underlying language impairment if the discrepancy is one standard deviation or greater. We posit that this conservative standard is reasonable given the lack of data currently available. As more research is done in this area, perhaps a sliding scale based on language experience can be applied *à la* Thordardottir (2015).

https://doi.org/10.1017/S1366728917000074 Published online by Cambridge University Press

On the other hand, for a majority L2, we do not recommend conceptual scoring to identify a receptive-expressive gap since, at least for Gross et al. (2014), it led to larger, not smaller, receptive-expressive gaps. The gap without conceptual scoring for the majority L2 (English) was roughly half a standard deviation. Therefore, we again recommend the conservative guideline of a one standard deviation receptive-expressive discrepancy as a way of judging a receptive-expressive gap to be meaningful and a possible indication of an underlying language impairment.

Although this recommended approach should increase the accuracy of identifying a receptive-expressive gap, it suffers limitations because the tests involved were neither co-constructed nor co-normed. Armon-Lotem et al. (2015) have developed an approach that should lead to more reliable identification of the receptive-expressive gap. They developed tests that both focused on language processing more than language knowledge and contained items that were co-constructed for several languages. The co-construction of items assures similar levels of difficulty across languages. For example, Haman, Luniewska and Pomiechowska (2015) developed Crosslinguistic Lexical Tasks (CLTs) for a wide range of languages. These tasks included co-constructed receptive and expressive items. As yet, this test has not been normed; but even without norms CLTs potentially could identify a receptive-expressive gap. Future test development should focus on the co-construction and co-norming of tasks across languages.

The review of the literature indicates converging evidence across a number of different studies regarding a receptive-expressive gap for bilingual children in the US. But the discrepancies reported here clearly indicate that we should be circumspect about these cross-test comparisons. The accommodations provided above are only small pieces to a much larger mosaic of language assessment practices that have received increased attention recently (Armon-Lotem et al., 2015; Kohnert, 2010). Although much progress has been made, significantly more research is necessary in order to confidently identify the role of the receptive-expressive gap in bilingual language impairment.

References

Alvarado, C., Ruef, M., & Schrank, F. (2005). *Woodcock-Muñoz Language Survey - Revised*. Itasca, IL: Riverside Publishing.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders IV Text Revision*. Washington, DC: American Psychiatric Association.

Armon-Lotem, S., de Jong, J., & Meir, N. (2015). *Assessing multilingual children: Disentangling bilingualism from language impairment*. Tonawanda, NY: Multilingual Matters.

Barnett, W., Yarosz, D., Thomas, J., & Jung, K. (2007). Two-way and monolingual English immersion in preschool education: An experimental comparison. *Early Childhood Research Quarterly*, *22*, 277–293.

Brown, R., & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior*, *5*, 325–337.

Caesar, L. G., & Kohler, P. D. (2007). The state of school-based bilingual assessment: Actual practice versus recommended guidelines. *Language, Speech, and Hearing Services in Schools*, *38*, 190–200.

Camilleri, B., & Law, J. (2007). Assessing children referred to speech and language therapy: Static and dynamic assessment of receptive vocabulary. *Advances in Speech-Language Pathology*, *9*, 312–322.

Carrow-Woolfolk, E. (1995). *Oral and Written Language Scales*. Bloomington, MN: Pearson Assessment.

Chiat, S. (2015). Non-word repetition. In S. Armon-Lotem, J. de Jong, & N. Meir (Eds.), *Assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 125–147). Tonawanda, NY: Multilingual Matters.

Dunn, L., & Dunn, L. (1981). *Peabody Picture Vocabulary Test*. Circle Pines, MN: American Guidance Service.

Dunn, L., Padilla, E., Lugo, D., & Dunn, L. (1986). *Test de Vocabulario en Imágenes Peabody (TVIP)*. Circle Pines, MN: American Guidance Service.

Gagarina, N., Klop, D., Tsimpli, I. M., & Walters, J. (2016). Narrative abilities in bilingual children. *Applied Psycholinguistics*, *37*, 11–17.

Gibson, T., Oller, D. K., Jarmulowicz, L., & Ethington, C. A. (2012). The receptive-expressive gap in the vocabulary of young second-language learners: Robustness and possible mechanisms. Bilingualism. *Bilingualism: Language and Cognition*, *15*, 102–116.

Gibson, T., Peña, E., & Bedore, L. (2014a). The receptive–expressive gap in bilingual children with and without primary language impairment. *American Journal of Speech-Language*, *23*, 655–667.

Gibson, T., Peña, E., & Bedore, L. (2014b). The relation between language experience and receptive-expressive semantic gaps in bilingual children. *International Journal of Bilingual Education and Bilingualism*, *17*, 90–110.

Glaspey, A., & Stoel-Gammon, C. (2007). A dynamic approach to phonological assessment. *Advances in Speech-Language Pathology*, *9*, 286–296.

Gollan, T., & Acenas, L. (2004). What is a TOT? Cognate and translation effects on tip-of-the-tongue states in Spanish-English and Tagalog-English bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 246–269.

Gross, M., Buac, M., & Kaushanskaya, M. (2014). Conceptual scoring of receptive and expressive vocabulary measures in simultaneous and sequential bilingual children. *American Journal of Speech-Language Pathology*, *23*, 574–586.

Hadley, P. A., Wilcox, K. A., & Rice, M. L. (1994). Talking at school: Teacher expectations in preschool and kindergarten. *Early Childhood Research Quarterly*, *9*, 111–129.

Haman, E., Luniewska, M., & Pomiechowska, B. (2015). Designing Cross-linguistic Lexical Tasks (CLTs) for bilingual preschool children. In S. Armon-Lotem, J. de

https://doi.org/10.1017/S1366728917000074 Published online by Cambridge University Press

Jong, & N. Meir (Eds.), *Assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 125–147). Tonawanda, NY: Multilingual Matters.

Hoff, E., Rumiche, R., Burridge, A., Ribot, K. M., & Welsh, S. N. (2014). Expressive vocabulary development in children from bilingual and monolingual homes: A longitudinal study from two to four years. *Early Childhood Research Quarterly*, *29*, 433–444.

Kohnert, K. (2010). Bilingual children with primary language impairment: Issues, evidence and implications for clinical actions. *Journal of Communication Disorders*.

Kohnert, K. (2013). *Language disorders in bilingual children and adults, second edition*. San Diego, CA: Plural Publishing.

Leonard, L. (2009). Is expressive language disorder an accurate diagnostic category? *American Journal of Speech-Language Pathology*, *18*, 115–123.

Marinis, T., & Armon-Lotem, S. (2015). Sentence Repetition. In S. Armon-Lotem, J. de Jong, & N. Meir (Eds.), *Assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 95–121). Tonawanda, NY: Multilingual Matters.

Miccio, A., Tabors, P., Páez, M., & Hammer, C. (2005). Vocabulary development in Spanish-speaking Head Start children of Puerto Rican descent. In J. Cohen, K. McAlister, K. Rolstad, & J. MacSwan (Eds.), *ISB4: Proceedings of the 4th International Symposium on Bilingualism*. Somerville, MA: Cascadilla Press.

Miller, L., Gillam, R., & Peña, E. (2001). *Dynamic assessment and intervention: Improving children's narrative abilities*. Austin, TX: PRO-ED.

Muñoz-Sandoval, A., Woodcock, R., McGrew, K., & Mather, N. (2005). *The Batería III Woodcock-Muñoz: Pruebas de aprovechamiento* [Batería Woodcock-Muñoz: Achievement tests]. Itasca, IL: Riverside.

Oller, D. K., Pearson, B. Z., & Cobo-Lewis, A. B. (2007). Profile effects in early bilingual language and literacy. *Applied Psycholinguistics*, *28*, 191–230.

Oller, D., & Eilers, R. (2002). *Language and literacy in bilingual children*. Tonawanda, NY: Multilingual Matters.

Olswang, L. B., & Bain, B. A. (1996). Assessment information for predicting upcoming change in language production. *Journal of Speech and Hearing Research*, *39*, 414–423.

Paradis, J. (2010). The interface between bilingual development and specific language impairment. *Applied Psycholinguistics*, *31*, 227–252.

Pearson, B. Z., Fernández, S. C., & Oller, D. K. (1993). Lexical development in bilingual infants and toddlers: Comparison to monolingual norms. *Language Learning*, *43*, 93–120.

Peña, E., Gutiérrez, Clellen, V., Iglesias, A., Goldstein, B., & Bedore, L. (2014). *Bilingual English Spanish Assessment*. Petaluma, CA: AR-Clinical Publications.

Ramos, M., Ramos, J., Hresko, W., Reid, D., & Hammill, D. (2006). *Test of Early Language Development (TELD) Third Edition: Spanish*. Austin, TX: PRO-ED.

Royal College of Speech and Language Therapists (RCSLT), Specific Interest Group in Bilingualism (2007). *Good Practice for Speech and Language Therapists Working with Clients from Linguistic Minority Communities*. Retrieved from RCSLT website: https://www.rcslt.org/members/publications/publications2/linguistic_minorities

Semel, E., Wiig, E., & Secord, W. (2003). *Clinical Evaluation of Language Fundamentals - Fourth Edition*. San Antonio, TX: The Psychological Corporation.

Speech Pathology Australia. (n.d.). *Working in a culturally and linguistically diverse society*. Retrieved from Speech Pathology Australia website: https://www.speechpathologyaustralia.org.au/spaweb/Document_Management/Public/Position_Statements.aspx

Stow, C., & Dodd, B. (2003). Providing an equitable service to bilingual children in the UK: A review. *International Journal of Language & Communication Disorders*, *38*, 351–377.

Thordardottir, E. (2011). The relationship between bilingual exposure and vocabulary development. *International Journal of Bilingualism*, *15*, 426–445.

Thordardottir, E. (2015). Proposed diagnostic procedures for use in bilingual and cross-linguistic contexts. In S. Armon-Lotem, J. de Jong, & N. Meir (Eds.), *Assessing Multilingual Children: Disentangling Bilingualism from Language Impairment* (pp. 331–358). Tonawanda, NY.

Uchikoshi, Y., & Maniates, H. (2010). How does bilingual instruction enhance English achievement? A mixed-methods study of Cantonese-speaking and Spanish-speaking bilingual classrooms. *Bilingual Research Journal*, *33*, 364–385.

Wiig, E., Semel, E., & Secord, W. (2006). *Clinical Evaluation of Language Fundamentals - Fourth Edition, Spanish (CELF-4 Spanish)*. San Antonio, TX: The Psychological Corporation.

Williams, C. J., & McLeod, S. (2012). Speech-language pathologists' assessment and intervention practices with multilingual children. *International Journal of Speech-Language Pathology*, *14*, 292–305.

Windsor, J., & Kohnert, K. (2004). The search for common ground: Part I. Lexical performance by linguistically diverse learners. *Journal of Speech, Language, and Hearing Research*, *47*, 877–890.

Woodcock, R. (1991). *Woodcock Language Proficiency Battery - Revised*. Itasca, IL: Riverside Publishing.

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

Woodcock, R., & Muñoz-Sandoval, A. (1995). *Woodcock Language Proficiency Battery - Revised: Spanish Form (WLPB-RS)*. Itasca, IL: Riverside Publishing.

Woodcock, R., Muñoz-Sandoval, A., Ruef, M., Alvarado, C. (2005). *Woodcock Language Proficiency Battery – Revised*. Itasca, IL: Riverside.

World Health Organization. (2005). *International statistical classification of diseases and related health problems, 10th revision*. Geneva.

Yan, S., & Nicoladis, E. (2009). Finding le mot just: Differences between bilingual and monolingual children's lexical access in comprehension and production. *Bilingualism: Language and Cognition*, *12*, 323–335.

Zimmerman, I., Steiner, V., & Pond, R. (2012). *Preschool Language Scales, Fifth Edition Spanish (PLS-5 Spanish)*. San Antonio, TX: The Psychological Corporation.

https://doi.org/10.1017/S1366728917000074 Published online by Cambridge University Press