HHS Public Access

Author manuscript

Biling (Camb Engl). Author manuscript; available in PMC 2018 May 02.

Published in final edited form as:

Biling (Camb Engl). 2018 March ; 21(2): 328–339. doi:10.1017/S1366728917000074.

Difficulties Using Standardized Tests to Identify the Receptive Expressive Gap in Bilingual Children's Vocabularies*

Todd A. Gibson†,
Louisiana State Unviersity

D. Kimbrough Oller, and
University of Memphis

Linda Jarmulowicz
University of Memphis

Abstract

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 because widely-used standardized tests may not offer comparable normed scores. Furthermore monolingual Spanish-speaking children tested in Mexico and monolingual English-speaking children in the US showed other, yet different statistically significant discrepancies between receptive and expressive scores. Results suggest comparisons across widely used standardized tests in attempts to assess a receptive-expressive gap are precarious.

Keywords

bilingualism; second-language learning; language attrition; vocabulary learning; school-age children

A significant body of research has documented that Spanish-English bilingual children's standardized receptive vocabulary scores are higher than their expressive vocabulary scores, even in their first language, Spanish (see Gibson, Pena, & Bedore, 2014, Table 1 for review). Such a discrepancy would not be predicted in theory because standardized tests are designed based on normative research to yield outcomes that should be comparable. It is widely recognized that picture pointing tasks (a common approach to testing receptive vocabulary) are easier than picture naming tasks (a common approach to testing expressive vocabulary). This difference should, in theory, be levelled after conversion to standard scores. Typically developing children's scores should be similar across any two such tests (e.g., children

*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.

scoring a 90 on a standardized receptive vocabulary test should score somewhere near 90 on a standardized expressive vocabulary test). We have termed the circumstance in which receptive standardized scores are significantly higher than expressive standardized scores the *receptive-expressive gap* (Gibson, Oller, Jarmulowicz, & Ethington, 2012).

The receptive-expressive gap has both theoretical and practical importance. In previous work (Gibson et al., 2012), we interpreted the gap as 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 to 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. Such a theoretical model would have clear practical implications because a receptive-expressive gap in monolinguals is associated with language disorder. Clinicians working with bilingual speakers would have to determine whether an observed receptive-expressive gap indicated language disorder or merely an epiphenomenon associated with second language learning.

In Gibson et al. (2012), we reported receptive vocabulary standardized test scores from the *Test de Vocabulario en Imagenes Peabody* (TVIP; Dunn, Padilla, Lugo, & Dunn, 1986) which we compared to the *Vocabulario Oral* (oral vocabulary) subtest of the Wo*odcock Language Proficiency Battery – Revised, Spanish Form* (WLPB-RS; Woodcock & Munoz-Sandoval, 1995). The TVIP is a picture-pointing task, and the *Vocabulario Oral* subtest of the WLPB-RS 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-Munoz Language Survey – Revised Spanish* form (WMLS-RS; Woodcock, Munoz-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 et al. (2002) 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.

The purpose of the current study was to report our discrepant findings using these similar picture naming tasks and to explore the possible contributions to 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

Bilingual participants included 116 Hispanic kindergarten (K) Spanish-English bilingual children in the US and 30 Spanish-speaking monolingual Mexican first graders. Additionally, a monolingual English-speaking comparison group of 134 kindergarteners were drawn from the same classrooms as the bilingual children (a subset of these children's results were reported in Gibson et al., 2012).

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. In our previous study (Gibson et al., 2012), we reported the results from 124 typically developing Hispanic K children for whom we had all demographic and testing data from our multi-year cross-sectional study up to that point. In the current study, we pooled the original 124 Hispanic K children with all other bilingual Hispanic K children for whom we had data in the US (109 additional children). 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 a pool of 229 children with an average age of 72.93 months, $SD = 4.71$. Because we wanted the Mexican first graders and the Hispanic K children from the USA 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 pool of 116 Hispanic 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 K, which occurred in their second semester. We chose to analyze posttest rather than pretest scores in order to match better the ages of the Mexican group. The Mexican first graders (11 girls and 19 boys) had an average age of 77.67, $SD = 3.66$, 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 appeared to be from relatively low socioeconomic (SES) backgrounds. The average number of years of formal education for the mothers of the bilingual K 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 (Woodcock & Munoz-Sandoval, 1995) and the Woodcock-Munoz Language Survey – Spanish form (Woodcock, Munoz-Sandoval, Ruef, & Alvarado, 2005). Both tests are picture-naming tasks in which the child

names a single 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.

In order to create Spanish language norms for the WLPB-RS, the test's examiner's manual (Woodcock & Munoz-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. The US sample of Spanish speakers (pre-K through adult) included 1,325 individuals (the number of K and first grade Spanish-speaking children in the US sample was not reported separately). From the US sample, the majority was born in the US (331), with the rest having immigrated to the US. Of the US Spanish-speaking sample, 75% reported using Spanish 100% of the time at home, which the examiner's manual reported was an indication of 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 is 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 norming sample. An additional statistical procedure was applied to allow test givers to use the same norming tables for both the English and Spanish forms of the test. 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 and rank ordered using Rasch analysis. A statistical

procedure was applied to rescale the difficulty of the Spanish item bank to the English item bank. The examiner's manual reports high validity and reliability.

Receptive vocabulary was measured by the *Test de Vocabulario en Imágenes Peabody* (TVIP; Dunn, Padilla, Lugo, & Dunn, 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 administering Spanish items translated from the English PPVT-R to Spanish-speaking groups in Mexico City and Puerto Rico. Data for SES was reported for the Puerto Rico group only. Scores from the two groups were combined and Rasch analysis was used to calibrate the items for difficulty.

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 K school year In the current study, 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. 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-R (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 *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 between the WLPB-RS and WMLS-RS, $t(115) = 17.06$, $p < .001$, $d = 1.86$, and the WLPB-RS and the TVIP, $t(115) = 15.04$, $p < .001$, $d = 1.46$, but no statistically significant difference between the TVIP and WML-RS, $t(115) = 1.91$, $p = .06$, $d = .18$. In English testing, US bilingual children presented with a 9 point receptive-expressive gap, with PPVT scores better than WLPB-R scores, $t(115) = 6.19$,

$p < .001$, $d = .60$. For the Mexican monolingual Spanish-speaking children, there was no statistically significant difference between the WLPB-RS and WMLS-RS, $t(29) = .43$, $p = .67$, $d = .09$. 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, $t(29) = 2.55$, $p = .016$, $d = .50$. For the US monolingual English-speaking children, there was a statistically significant difference between WLPB-R, $M = 100.36$, $SD = 14.75$, and PPVT-III, $M = 92.38$, $SD = 12.63$, $t(133) = -8.40$, $p < .001$, $d = -.74$, in the opposite direction, with expressive skills outranking receptive skills.

Summary and Discussion

Results from the current study demonstrate markedly different outcomes depending on the tests used and the child language-backgrounds. The standardized assessments used here had a mean of 100 and a standard deviation of 15; therefore, a group of children with a standard score of, for example, 89 on one of the tests should score somewhere near 89 on the other tests. Significant discrepancies between tests should indicate real differences in abilities. For the Spanish-English bilingual kindergarteners in this study, however, comparisons with the TVIP yield a receptive-expressive gap for the WLPB-RS but not for the WMLS-RS.

Indeed, 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. This circumstance raises the question: Which test is appropriate for the Spanish-English bilingual speakers in the US? They cannot both be right.

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 the test or anomalies in the current sample's performance.

English vocabulary test scores also yielded discrepancies. 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.

The review of the literature indicates converging evidence across a number of different studies that indicate a receptive-expressive gap for bilingual children in the US. But the discrepancies reported here clearly indicate that we should be circumspect about comparisons across tests. Future research will likely require evidence using a wide variety of methodologies to provide the best conclusions regarding the receptive-expressive gap.

References

Alvarado, C., Ruef, M., Schrank, F. Woodcock-Munoz language survey-revised. Itasca, IL: Riverside Publishing; 2005.

Dunn, L., Dunn, L. Peabody Picture Vocabulary Test (PPVT). Circle Pines, MN: AGS Publishing; 1981.

Dunn, L., Padilla, E., Lugo, D., Dunn, L. Test de Vocabulario en Imágenes Peabody (TVIP). Circle Pines, MN: AGS; 1986.

Gibson TA, Oller DK, Jarmulowicz L, Ethington CA. The receptive–expressive gap in the vocabulary of young second-language learners: Robustness and possible mechanisms. Bilingualism: Language and Cognition. 2012; 15:102–116.

Gibson TA, Peña ED, Bedore LM. The relation between language experience and receptive-expressive semantic gaps in bilingual children. International Journal of Bilingual Education and Bilingualism. 2014; 17:90–110. [PubMed: 29670456]

Oller, D., Jarmulowicz, L., Gibson, T., Hoff, E. First language vocabulary loss in early bilinguals during language immersion: A possible role for suppression. In: Caunt-Milton, H.Kulatilake, S., Woo, I., editors. Proceedings of the 31st Annual Boston University Conference on Language Development. Somerville, MA: Cascadilla Press; 2007. p. 474-484.

Schrank, F., McGrew, K., Ruef, M., Alvarado, C., Muñoz-Sandoval, A., Woodcock, R. Overview and technical supplement (Batería III Woodcock-Muñoz. Itasca, IL: Riverside Publishing; 2005. Assessment Service Bulletin No. 1

Secretaría de Gobernación, Consejo Nacional de Población. Ïndice de Marginación por Localidad 2010. 2012. Retrieved from http://www.conapo.gob.mx/work/models/CONAPO/indices_margina/2010/documentoprincipal/Capitulo01.pdf

Woodcock, R. Examiner's manual. Itasca, IL: Riverside Publishing; 1991. WLPB-R: Woodcock language proficiency battery—revised.

Woodcock, RW., McGrew, K., Mather, N. Woodcock-Johnson tests of achievement. Itasca, IL: Riverside Publishing; 2001.

Woodcock, RW., Muñoz-Sandoval, AF. Woodcock language proficiency battery-revised: Spanish form (WLPB-RS). Itasca, IL: Riverside Publishing; 1995.

Highlights

- For bilingual children, the magnitude of the discrepancy between receptive and expressive standardized vocabulary scores was dependent on the tests administered
- For monolingual children, expressive standardized vocabulary scores was greater than receptive
- Comparisons across widely used standardized vocabulary tests in attempts to assess a receptive-expressive vocabulary gap are precarious

Table 1

Average standard scores.

	WLPB-R English	WLPB-R Spanish	WMLS-R Spanish	TVIP	PPVT
US Bilinguals	62.56(20.57)	61.96 (21.47)	87.03 [a] (12.42)	89.34[a] (15.06)	71.09(15.64)
Mex Monolinguals	NA	96.83 (16.48)[b]	95.77 (10.26)[b]	104 (9.68)	NA
US Monolinguals	100.36(14.75)	NA	NA	NA	92.38(12.63)

Notes. Superscripts indicate no statistically significant difference. US monolinguals were tested only in English

Biling (Camb Engl). Author manuscript; available in PMC 2018 May 02.