Clinical Forum

Conceptual Versus Monolingual Scoring: When Does It Make a Difference?

Lisa M. Bedore
Elizabeth D. Peña
Melissa García
Celina Cortez
The University of Texas at Austin

The unique challenge for assessing bilingual vocabulary development is that the speaker’s vocabulary may be distributed across the languages spoken. Young simultaneous and sequential learners use vocabulary from both of their languages. Children under the age of 2 years growing up in a bilingual environment—simultaneous bilinguals—often use a single set of words (from both languages) to communicate with adult bilinguals in their environment. For example, in their case study of bilingual language acquisition, Deuchar and Quay (2000) reported fluctuating levels of translation equivalents based on vocabulary diaries—the same word from each of their languages such as pan (bread) and bread. The percentages of translation equivalents, or overlapped vocabulary, in the child’s lexicon ranged from a high of 67% at the onset of word use to a low of 10% at the acquisition of the first 50 words. Beyond that point, the numbers of translation equivalents slowly increased again to 44% at age 1;10 (years;months). Reports for simultaneous bilingual toddlers based on parent report using equivalent instruments (i.e., The MacArthur Communicative Development Inventories [CDI; Fenson et al., 1993] and the
MacArthur Inventario de Desarrollo de Habilidades Comunicativas [IDHC; Jackson-Maldonado et al., 2003]) indicated that approximately 30% of vocabulary is overlapped and approximately 70% is unique (e.g., Pearson, Fernández, & Oller, 1995). When young children use translation equivalents, it is thought that some of these words may be treated by the child as having two separate meanings, for example, pan referring to rolls and sweet breads and bread for making sandwiches or toast.

Code switching is sometimes interpreted to indicate a lack of semantic knowledge. However, in young simultaneous bilinguals, it seems to be dictated by several factors related to language dominance. First, children tend to use more vocabulary items in their dominant language (Genesee, Nicoladis, & Paradis, 1995). In addition, it is more likely that children will use the language of the conversational context (Deuchar & Quay, 1999). Sometimes, code switches can be attributed to gaps in the lexicon in the nondonominant language (Fantini, 1985). It is important to keep in mind that code switching does not inhibit semantic development. Gawlitzeck-Maiwald and Tracy (1996) argued that semantic knowledge in both of a bilingual’s languages (as reflected in vocabulary use) boosts productivity across the lexical and syntactic systems. Thus, the need for vocabulary in one language may nudge the child to seek out comparable vocabulary in the other language.

Evaluations of early sequential bilingual children (i.e., children who begin to acquire their second language [L2] at age 3 or 4 years) via single-word receptive vocabulary tests such as the Peabody Picture Vocabulary Test—Revised (PPVT–R; Dunn & Dunn, 1981) and Test de Vocabulario en Imágenes Peabody (TVIP; Dunn, Padilla, Lugo, & Dunn, 1986), show that these children demonstrate lower than average skills in both of their languages as compared to monolingual speakers (Umbel, Pearson, Fernández, & Oller, 1992). Their overlapping vocabulary was 65%; note that this task is a receptive, picture identification task, and the English and Spanish versions share a high degree of vocabulary. Recently, Peña, Bedore, and Zlatic-Guinta (2002) used conceptual scoring to evaluate performance on a category generation task that was administered in Spanish and English to a group of young sequential bilingual learners ranging in age from 4 to 7 years. In this production task, children generated more singlets (or language-unique items) than overlapping vocabulary. Younger children (mean age = 5;1) produced 29.5% overlapped vocabulary and older children (mean age = 6;5) produced 33.8% overlapped items. This pattern of overlapped vocabulary holds even to adulthood. When adult bilinguals generated lists of words (by categories such as foods or clothing) in each of their languages, approximately 45%–60% of responses were overlapped and the rest were language specific (Roberts & Le Dorze, 1997). These studies illustrate two aspects of distributed vocabulary knowledge that need to be considered in assessment. The first is that the total knowledge of bilinguals in a single language may not be directly comparable to that of monolinguals. The second is that some but not all of the vocabulary overlaps across languages.

Work by Kohnert, Bates, and Hernández (1999) illustrated that access to vocabulary in each of a sequential bilingual’s languages shifts over time. For example, 5- and 6-year-olds demonstrated higher levels of accuracy and faster response times on a picture naming task in their first language (L1), Spanish, than in their L2, English. By 8–10 years of age, children were much closer in their levels of accuracy across languages, and response times across languages were comparable. However, by age 14 to 16 years, children were still accurate in the picture naming tasks but they responded faster in their L2, English.

Ordóñez, Carlo, Snow, and McLaughlin (2002) explored the extent to which vocabulary knowledge might transfer across languages. Bilingual fourth and fifth graders (who varied in levels of English and Spanish proficiency as measured by single-word receptive vocabulary tests and reading comprehension measures) defined high-frequency concrete nouns in each of their languages. Children’s responses were coded to reflect their paradigmatic and syntagmatic knowledge of the word as well as the communicative adequacy of the response. The vocabulary test results were used as measures of breadth and the definitional task was used as a measure of depth. Prediction of depth of vocabulary knowledge was possible only when performance on the breadth measures in both languages was taken into account. Metalinguistic transfer (i.e., knowledge of how to define words) did not occur unless children had acquired specific vocabulary in the L2, English. As a group, these studies emphasize the importance of taking into account that bilingual children’s vocabulary development is particularly likely to be influenced by the frequency of exposure to specific words and context when they learn one language at home and the other at school.

Fantini (1985) observed in his case study that his son was especially likely to use a word from another language when he had no ready lexical equivalent in the language in which he was speaking. Similarly, sequential bilinguals who are in the process of acquiring vocabulary may have lexical gaps where they know a word in only one of their languages. An example is a child who might know the word toronja (grapefruit) in Spanish but not know what to call the fruit in English. Alternatively, a child may learn a word in his or her L2 but not know it in his or her L1. Thus, a child studying ladybugs at school might ask what one is called in the other language. When there are such gaps, bilingual children could code switch as reported by Fantini or produce errors resulting from competition from more familiar lexical items, as is observed in monolingual children (Gershkoff-Stowe & Smith, 1997). This notion of interference or competition can be applied to bilingual children who are starting to use vocabulary from both languages (Peña & Stubbe Kester, 2004). Interference and competition between lexical items in both languages could cause an increase in the number of errors produced by bilingual children in tasks in which they are expected to generate specific vocabulary items. Looking at the number of errors produced by the bilingual children in each of their languages can also be beneficial for establishing assessment and intervention guidelines.
TESTING BILINGUAL VOCABULARY DEVELOPMENT

The recommendation that speech-language pathologists (SLPs) need to evaluate skills in both of a bilingual’s languages seems consistent with the patterns of findings for simultaneous and sequential bilingual children discussed above (Chamberlain & Medeiros-Landurand, 1991; Langdon, 1989). Single-language measures ignore the fact that bilingual children may choose to use different words depending on the setting, interlocutor, and context (Iglesias, 2001) as well as their cultural experiences (Peña, 2001). These difficulties in assessment are further compounded when tests are used with children for whom they were not designed. Fully evaluating both of a bilingual child’s languages may help to provide a better estimate of his or her semantic knowledge, thereby accounting for similarities and differences between languages (Pearson, Fernández, & Oller, 1993). However, this is a time-consuming solution that will not necessarily yield more accurate assessments if clinicians do not have the normative information needed to interpret such results.

Tests that address both languages potentially yield scores that reflect a comparable (combined) vocabulary size for bilingual learners when compared to that of monolinguals (Alvarado, 2000; Clarizio, 1982; Figueroa, 1989). Integrated language assessments of this type are supported by Grosjean’s (1989) holistic view of bilingualism in which bilinguals are recognized as having unique linguistic configurations that reflect a distinctive system of dual-language acquisition that is different from monolinguals. Scoring conventions may also help capture a bilingual’s integrated system.

Pearson et al. (1993) described conceptual scoring as scoring the meaning of a response regardless of the language in which it is produced. Thus, when describing a ball, if a child said, “It’s red and blue y tiene una raya y una estrella” (It’s red and blue and has a stripe and a star), she would achieve a monolingual score of 2 in English or Spanish but a conceptual score of 4 because she expressed unique concepts in each language. This illustrates how the use of a conceptual score acknowledges bilingual children’s knowledge in their two languages for an overall estimate of bilingual semantic development. Gaps in the child’s lexicon that are compensated for by the use of code switching would not be penalized with this scoring procedure. True errors would, however, still be observed. For example, if the child said that a red ball was yellow or said that it was amarillo, the error would be evident and could be noted.

Marchman and Martínez-Sussman (2002) found positive correlations between children’s conceptual vocabulary score (based on Pearson et al., 1993) for combined administrations of the English CDI and the Spanish IDHC and language production in language samples. Single-language scores correlated with use of the corresponding language but did not correlate with scores in the other language. These findings suggest that the composite score is potentially useful as a predictor of a bilingual child’s word use (Marchman & Martinez-Sussman, 2002). Thus, for the purposes of quantifying development, it is important to use a scoring system that accounts for the distribution of total vocabulary knowledge across two languages. However, it should be noted that in work to date with conceptual scoring (e.g., Marchman & Martínez-Sussman, 2002; Pearson et al., 1993), researchers have used English and Spanish tests that were designed primarily for monolingual populations, and they were used with very young children for whom parent report was possible (see Kan & Kohnert [in press] and Kohnert, Hernández, & Bates [1998] for exceptions). This pattern of findings needs to be verified for older children.

To summarize, bilingual children have vocabulary knowledge distributed in two languages. Some of this vocabulary is language specific and some is shared. This presents a unique challenge when developing tests for bilingual children. There are two potentially related solutions to this problem: One is to develop tests that accurately reflect bilingual children’s vocabulary knowledge; another is to develop ways to holistically analyze bilingual children’s distributed vocabulary knowledge.

The current study aims to explore the performance of bilingual children on a set of items that were specifically designed to reflect bilingual children’s language experiences and then to explore the scoring conditions under which typically developing (TD) children will be correctly classified. Figure 1 provides a graphic representation of the problem and the predictions. The first two bars illustrate the assumption that monolingual and bilingual children are expected to have comparable amounts of vocabulary overall, but that bilingual children’s vocabulary may be distributed across two languages. The lower bars illustrate three ways of quantifying bilingual vocabulary: total score, single-language scoring, and conceptual scoring. A total score counts all vocabulary items in each language. When there is greater overlap between items on the tests of the L1 and L2, there is greater potential for a higher total score. However, along with the higher score comes the possibility of a greater variability in scores that will lower the cutoff for identification. Single-language scoring results in a lower score because of L2-unique vocabulary. If, for example, a child had 30% overlap (discussed earlier) and relatively balanced skills in each language, the child would score in the low average or just below the average range as compared to a monolingual norm. Even if a child was dominant in one of the languages and had overlapping vocabulary, he or she would have a somewhat lower than expected score based on monolingual norms (Umbel et al., 1992). This scoring method thus introduces the possibility of overidentification of language impairment (LI). Also, the amount of bilingual exposure would also be likely to increase the variability. Finally, a conceptual score potentially yields a range of scores that is comparable to that of monolingual children and may be useful over a range of bilingual language exposure.

STUDY 1

Purpose

Studies of vocabulary development as well as results of studies testing both languages call into question the
assumption that bilingual children will have highly similar vocabulary in both of their languages. In Study 1, we explored the nature of bilingual children’s responses to open-ended, expressive questions that were designed to reflect the language experience of children growing up in bilingual communities (regardless of whether they were bilingual). We evaluated the nature of the children’s responses to determine if they produced equivalent responses to equivalent questions in Spanish and English. If children based responses on language-specific experiences, then they may have responded in distinct ways when asked the equivalent question in each language. We also looked for evidence of gaps in the children’s lexicon (as evidenced by code switching and errors). Children’s responses were quantified using total, single-language, and conceptual scores.

Method

Participants. Fifty-five TD Latino American children between the ages of 4:0 and 7:11 participated in Study 1. These children were selected from a larger pool of 84 children who participated in the first phase of a larger test development project. The subset of 55 children was selected to meet the following criterion: They exhibited no LI (based on parent or teacher report), they were in the targeted age range, and they had completed all of the tasks under analysis. These children were further divided into four language groups based on either the percentage of Spanish and English output per parent and teacher report or grammaticality in the target language on a narrative story-telling task. The primarily English-speaking (PE) group (N = 11) used English 80% or more during daily language use, with less than 20% Spanish use. The bilingual English (BE) group (N = 7) used English 50%–80% of the time on a daily basis and Spanish 20%–50% of the time. The bilingual Spanish (BS) group (N = 13) used Spanish 50%–80% of the time on a daily basis and English 20%–50% of the time. The primarily Spanish-speaking (PS) group (N = 24) used Spanish 80% or more of the time on a daily basis and English less than 20% of the time. In the cases (N = 9) that parent interview information was not provided, children were placed in a language group based on grammaticality in the target language on a narrative story-telling task because this measure correlates highly with percentage of input and output in a given language (Gutierrez-Clellen & Kreiter, 2003; Stubbe Kester & Peña, 2002). Children who produced 75% or more grammatical utterances at age 4, 81% or more grammatical utterances at age 5, and 86% or more grammatical utterances at ages 6 and 7 were considered fluent in that language. These cutoff scores were extrapolated from Gutierrez-Clellen and Kreiter’s work with slightly older children. Cutoff points were set for each age between the average percentage of grammaticality for children for whom there were no concerns about language development and the percentages for children for whom there was parent and/or teacher concern.

Materials/task development. The study focused on the performance of children on the characteristic property items that were included in pilot testing for the semantics subtest of the BESA (Peña et al., in development). This version of the task included 24 characteristic property items (12 expressive and 12 receptive) balanced by language (12 English and 12 Spanish) (there were 172 items overall, as described in Peña et al.). For the purpose of this study, 11 of the 12 expressive characteristic property items were selected for analysis because the responses for one of the Spanish items could not be coded; the items are listed in Appendix A. These items required that the participants describe objects. Item content was taken from concepts that were familiar to preschool children as indicated by teacher–child interaction data collected in a bilingual preschool...
over a 10-week period (Ortega, 1997). English and Spanish items targeted the same concepts (e.g., object shapes, colors, sizes, function), but different questions were used to target the concept in each language. For example, in the English version of the test, children were asked to describe a school bus; in the Spanish version, children were asked to describe a truck. The selection of vocabulary and themes was based on a literature review on language development and cultural relevance in each language (e.g., Choi, McDonough, Bowerman, & Mandler, 1999; Nelson & Nelson, 1990; also see Peña, Bedore, & Rappazzo, 2003, for further discussion). For example, comparisons need to be made in the language in which bilingual children will be familiar with the vocabulary. Thus, comparing a car and a school bus may be appropriate in English, whereas comparing a cuchara (spoon) and a tenedor (fork) may be more appropriate in Spanish for children learning Spanish at home and English at school in the United States. Therefore, the items were not translations but were similar with respect to content.

An earlier study (Peña et al., 2003) indicated no significant differences on the characteristic property items (as compared to other item types) in a repeated measures analysis of variance (ANOVA). Here, an ANOVA examining performance on the characteristic property items by that same group for test language (Spanish or English) and language group (functional monolingual or bilingual) was conducted. Results indicated no significant score differences for test language, \( F(1, 81) = 1.098, p = .298, \eta^2_p = .013 \), or for language group, \( F(1, 81) = .365, p = .547, \eta^2_p = .004 \). The effect sizes were negligible, so the results were judged indicate that the items in each language were psychometrically equivalent.

Procedure. Children were individually administered the Characteristic Properties subtest of items in their school or home by trained bilingual graduate students. The PE group completed the English version of the subtest. The PS group completed the Spanish version of the subtest. The BS and BE groups completed both versions. The bilingual participants completed each subtest on different days and test order was counterbalanced. Participants’ responses were recorded in the language(s) in which they were produced on the response form designed for the subtest.

Scoring and reliability. The scores were generated for the correct responses to each item (monolingual score in English or Spanish, total response score, and conceptual score). Monolingual scores consisted of nonduplicated words produced for each item in the target language. The total response score is the number of correct items in the test language and the other language (e.g., grande and big would both be counted). The conceptual score is the number of unique correct concepts in the test language and the other language (e.g., grande and big would only be counted once). It is important to note that, unlike other studies that have employed conceptual or total scores, different questions were asked in each language in this study. Thus, the source of “other-language” responses is code-switched responses that the child generated during testing. For example, if a child responded es café (it’s brown) in response to the question, “What does the dog look like?” this would be considered a code-switched or other-language response. All scores were raw scores (the number of items generated in response to each question). In addition, error scores were generated that represent the average number of errors that the children generated to each question.

Twenty percent of the response forms were rescored for correctness in Spanish and/or English by a bilingual graduate student in speech-language pathology. Interrater reliability for scoring using the indicated scoring procedures was 95%.

Results

Table 1 lists the average number of responses generated for each of the questions for each of the groups. There are Spanish scores for the PS group as well as the BS and BE groups. English scores are listed for the PE group and the BS and BE groups. An example of a single-language

| Table 1. Average and standard deviation scores by language (primarily Spanish-speaking [PS], bilingual Spanish [BS], bilingual English [BE], and primarily English-speaking [PE]) for group in Study 1. |
|----------------|----------------|----------------|----------------|----------------|
|                | Average | SD    | Average | SD    | Average | SD    | Average | SD    |
| **Spanish**    |         |       |         |       |         |       |         |       |
| Monolingual score | 1.44   | .92   | 1.62   | 1.10  | 1.20   | .97   |         |       |
| Total score    | 1.59   | .96   | 1.77   | 1.11  | 2.26   | 1.32  |         |       |
| Conceptual score | 1.56   | .94   | 1.77   | 1.11  | 2.23   | 1.30  |         |       |
| Errors         | .47    | .40   | .37    | .31   | .29    | .32   |         |       |
| **English**  |         |       |         |       |         |       |         |       |
| Monolingual score | .69    | .54   | 1.55   | .88   | 1.71   | 1.08  |         |       |
| Total score    | .86    | .61   | 1.57   | .91   | 1.71   | 1.08  |         |       |
| Conceptual score | .86    | .61   | 1.57   | .91   | 1.71   | 1.08  |         |       |
| Errors         | .61    | .41   | .74    | .43   | .25    | .18   |         |       |

Note. Average scores represent the number of responses given per question. PS children were not tested in English; PE children were not tested in Spanish.
response included *round, baked, and has candles* in response to the question, “Tell me about a birthday cake.” A common example of a code-switched response to the same question was *it has flowers and velas* (it has flowers and candles.) The code-switched response was qualitatively similar to the English response. The children’s English-language and Spanish-language scores were evaluated separately using a repeated measures ANOVA with language group as the between-subjects factor and score type as the within-subjects factor. Effect sizes were calculated to determine the strength of the independent and dependent variables. For this study, partial eta squared ($\eta^2$) was calculated. Partial eta squared is the proportion of the total variance plus error accounted for by the effect in the sample. One advantage of this statistic is that it is not dependent on the number or magnitude of other effects. Currently, there are no guidelines for interpreting these statistics in the field. However, in previous studies (e.g., Peña et al., 2003), guidelines derived from the interpretation of correlation analysis have been adopted. Thus, effect sizes between .80 and 1 are considered very large; effect sizes between .5 and .8 are considered large; effect sizes between .25 and .5 are considered moderate; effect sizes between .1 and .25 are considered small; and effect sizes less than .1 are considered negligible.

Error production was also evaluated. The bottom row for each language in Table 1 contains the average number of errors that the children generated to each question. Responses were considered errors when characteristics that were not related to the object in question were provided. Again, these did not appear to differ qualitatively by language in which they were given. For example, when asked to describe a dog, children generated similar types of errors in English (e.g., *balloons*) and in Spanish (e.g., *libro “book”*).

**Spanish-language comparisons.** Children in the PS, BS, and BE groups all obtained comparable scores on the task in Spanish, $F(2, 40) = .410, p = .666, \eta^2 = .020$. This effect size was negligible. Within-subject comparisons (with Greenhouse Geisser correction applied to control for multiple comparisons) revealed differences between the monolingual, total, and conceptual scores, $F(1.013, 40.506) = 16.321, p = .0002, \eta^2 = .290$. This effect size was moderate. Monolingual scores were lower than conceptual scores ($p = .001$) and total scores ($p = .001$). This suggests that children produced some code-switched responses to these questions. However, on the Spanish items, total scores and conceptual scores did not differ from each other ($p = .078, \eta^2 = .00001$). The lack of difference between the total and conceptual scores suggests that the responses that the children produced in English did not duplicate their responses in Spanish. A comparison of the error rate across the groups in Spanish yielded no significant differences, $F(2, 40) = .831, p = .44, \eta^2 = .040$. This effect size was negligible.

**English-language comparisons.** There were significant differences between the scores for the children in the BS, BE, and PE groups, $F(2, 26) = .465, p = .046, \eta^2 = .210$. The effect size, however, was small. The BE group and PE group scores did not differ significantly ($p = .734$). The scores for the BS group were lower than those for the PE group ($p = .021$) but did not differ significantly from the scores for the BE group ($p = .074$). Given that the BS children are in the process of acquiring English, this difference may be attributed to their status as learners. Within-subject comparisons (with a Greenhouse Geisser correction applied to control for multiple comparisons) revealed differences between the monolingual, total, and conceptual scores, $F(1.000, 26.000) = 4.188, p = .051, \eta^2 = .139$. This was a small effect size. These findings indicate that when the task was completed in English, children were less likely to code switch than when the task was completed in Spanish. Thus, their total and conceptual scores—which incorporate Spanish knowledge—were not significantly different from their monolingual scores.

A comparison of error rate for the groups in English revealed overall differences in the error rates by language designation group, $F(2, 26) = .581, p = .019, \eta^2 = .263$. This was a moderate effect size. Post hoc comparisons indicated that the error rate of the BE group ($M = .738$) did not differ significantly from the error rate of the BS group ($M = .611; p = .457$); however, the error rate of the PE group ($M = .250$) was lower than that of the BE group ($p = .009$) and the BS group ($p = .025$). These findings are consistent with reports that children are more likely to make naming errors when they are in the process of acquiring vocabulary (Dapreto & Bjork, 2000; Gershkoff-Stowe & Smith, 1997).

**Discussion**

Overall, these findings demonstrate that children produced more vocabulary in their dominant language, as was expected. Furthermore, children from all groups produced additional items from the other language. Even children in the PE or PS group occasionally knew vocabulary from the nondominant language and produced it in a bilingual testing context. Children in all groups produced errors at a comparable rate, but children in the BS and BE groups were more likely to produce errors (reflecting a lack of knowledge of some words) in at least one of the testing contexts. The findings suggest that total and conceptual scores are potentially valuable in capturing normal performance of TD bilingual children.

In this test, the same concepts were tested using different open-ended questions in each language. Under these conditions, there was little overlap in the children’s responses. This was a distinctly different pattern than was observed in Peña et al. (2002), where the items were the same in each language and the response set contained approximately 30% overlapping items. In the current study, there were only differences between the total and conceptual scores for the BE children on the Spanish subtest. Although the total score was slightly higher, the variability was also higher. This is likely to be problematic in setting a cutoff point to accurately identify children with and without LI. Given the potential for the conceptual score to have lower variability, in Study 2, we focused on the application of a conceptual score for correct classification of TD children. One final reason to continue to explore the...
conceptual score instead of the total score is that it may be more comparable to the monolingual score of the PS or PE children.

### STUDY 2

#### Purpose

Study 2 focused on whether or not the conceptual scores of TD BS and BE children were comparable to the scores of PS or PE children when a full range of semantic test items was administered. Single-language and conceptual scores of the BS and BE groups were compared to each other and to the single-language scores of the PS and PE groups. The purpose of Study 2 was to determine if conceptual scoring could be used to classify the vocabulary development of TD children accurately.

#### Method

**Participants.** Forty TD Latino American children between the ages of 5:0 and 6:1 participated in Study 2. These children were selected from a larger pool of children who were participating in an ongoing study of semantic development in bilingual children. For this study, children were selected to match age and language background closely; none of the children who participated in Study 1 participated in Study 2. Subject selection procedures followed the same procedures outlined for Study 1. As seen in Table 2, the average ages of the groups were quite similar; the use levels of the dominant groups in the outermost columns were similar, as were the use levels of the bilingual groups.

**Materials and procedures.** The data of interest for this study were the expressive items (17 Spanish; 14 English) of the Phase 2 version of the Semantic subtest of the BESA (Peña et al., in development). Overall, the subtest contains receptive and expressive items embedded in three illustrated stories. There are six item types, including characteristic properties (a subset of the items included in Study 1), functions, analogies, linguistic concepts, similarities and differences, and comprehension of passages. The actual items for English and Spanish versions of this subtest were selected based on their difficulty levels and discrimination values from a larger set of items employed in pilot testing.

| Table 2. Average age in months and language output for participants by language group in Study 2. |
|---|---|---|---|
| PS | BS | BE | PE |
| Age (months) | 66.00 | 64.40 | 63.90 | 64.40 |
| Spanish output | 95.66 | 64.29 | 37.69 | 4.16 |
| English output | 4.34 | 35.71 | 62.27 | 95.84 |

*Note. Output represents the average percentage of time per day that children produce Spanish or English.*

Appendix B illustrates that the number of each item type is different for the two versions of the test. The Spanish and English versions are not translations but were designed to be equivalent with respect to content and range of difficulty based on the analysis of children’s responses during pilot testing (Peña et al., 2003; Peña et al., 2002). Overall, the item difficulty levels ranged from .3 to .8 for English and Spanish. The average item difficulty level for the English items was .61 and .56 for the Spanish items. The procedures for testing were identical to those of Study 1.

**Scoring and reliability.** Responses for each item were scored with the monolingual and conceptual scoring procedures described for Study 1. On the basis of data from Phase 1 testing, a preestablished list of acceptable responses was used to determine correct responses. The procedure for calculating these score types was identical to the procedure employed in Study 1. However, raw scores for the monolingual scores and conceptual scores were converted to percentages in order to compare across languages because the number of items on the test in each language differed. Twenty percent of the response forms were rescored by a bilingual research assistant. Interrater reliability for correct scoring of responses in Spanish and English was 99%.

#### Results

The current research examined how children with varying levels of bilingualism performed on the expressive items from a battery of semantic tasks. Table 3 summarizes the performance of the bilingual and PE or PS children’s mean percentage scores and standard deviations using the monolingual and conceptual scoring systems. A three-factor repeated measures ANOVA with subtest (Spanish, English) and score (monolingual, conceptual) as within-subjects factors and group (BE, BS) as the between-subjects factor was used to compare the scores of the bilingual groups. In addition, two two-factor repeated measures ANOVAs with score (monolingual score, conceptual score) as the within-subjects factor and group (PS, BS, and BE; PE, BS, and BE) as the between-subjects factor were conducted. Effect sizes were calculated using the same procedure as in Study 1. Finally, each bilingual child’s score was compared individually to the PS and PE groups’ means with a classification analysis.

Results of between-subjects comparisons indicated that there were no significant group (BE, BS) differences, $F(1, 18) = .521, p = .480, \eta^2 = .028$, suggesting that the average score of both subtests was comparable for the bilingual groups. Additionally, results of within-subjects contrasts indicated that there were no significant differences in how children scored on the two subtests (Spanish vs. English), $F(1, 18) = .481, p = .497, \eta^2 = .026$. These effect sizes were negligible. There was a significant main effect for score, $F(1, 18) = 17.780, p = .001, \eta^2 = .497$, with a moderate effect size. Children’s conceptual scores ($M = 46.10\%$) were significantly higher than their monolingual scores ($M = 40.95\%$). There were significant interactions for Test $\times$ Group, $F(1, 18) = 6.922, p = .017, \eta^2 = .278$, and Test $\times$ Score, $F(1, 18) = 7.834, p = .012, \eta^2 = .278$. These results suggest that Spanish-English bilingual children are more likely to score higher on the conceptual score than on the monolingual score, with English being the dominant language.
.303, with moderate effect sizes. These interactions are represented in Figures 2 and 3. Post hoc analysis using Scheffé’s test for multiple comparisons indicated that the difference between the BS group and the BE group on the English subtest was greater than the difference between BS and BE performance on the Spanish subtest (M difference = 27.70%), p < .01. Whereas the BS group received higher scores than the BE group on the Spanish subtest (M = 46.50% vs. M = 36.80%), the difference was not statistically significant. This may indicate that the bilingual children, regardless of whether they were English or Spanish dominant, were highly familiar with the content presented on the Spanish version of the test.

Post hoc analysis using Scheffé’s test for multiple comparisons demonstrated that the difference between monolingual and conceptual scoring was greater for the Spanish version of the test (M difference = 7.1%) than for the English version of the test (M difference = 0.4%), p < .01. In other words, the BS and BE group provided a greater number of responses in English during the Spanish subtest than they did Spanish during the English test. Many of the items on the BESA (Peña et al., in development) include academic concepts (e.g., colors, shapes), and it is likely that the bilingual children learned these concepts in English rather than Spanish. Children may have perceived the situation as another school activity and thus responded in the language of instruction, even though examiners were fluent Spanish speakers.

Effects of language output and scoring method. Our second question addressed how bilingual children’s scores compared to the scores of PS and PE children when conceptual scoring was applied. Two two-factor repeated measures ANOVAs with score (monolingual score, conceptual score) as the within-subjects factor and group (PS, BS, and BE; PE, BS, and BE) as the between-subjects factor were conducted to compare the PS and PE groups to the BS and BE groups.

Spanish comparison. A repeated measures ANOVA comparing the PS, BS, and BE groups revealed a main effect for score, $F(1, 27) = 14.467, p = .001$, and a moderate effect size of $\eta^2_p = .349$, but no significant main effect for group. There was, however, a significant Score $\times$ Group interaction, $F(2, 27) = 3.384, p = .049$. $\eta^2_p = .200$. This effect size was small. Post hoc comparisons indicated

<table>
<thead>
<tr>
<th></th>
<th>PS Average</th>
<th>PS SD</th>
<th>BS Average</th>
<th>BS SD</th>
<th>BE Average</th>
<th>BE SD</th>
<th>PE Average</th>
<th>PE SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monolingual score</td>
<td>50.6</td>
<td>11.73</td>
<td>43.6</td>
<td>24.59</td>
<td>31.2</td>
<td>19.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conceptual score</td>
<td>51.7</td>
<td>13.59</td>
<td>49.4</td>
<td>21.58</td>
<td>42.4</td>
<td>19.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monolingual score</td>
<td>34.9</td>
<td>20.86</td>
<td>54.2</td>
<td>11.52</td>
<td>51.3</td>
<td>20.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conceptual score</td>
<td>37</td>
<td>20.7</td>
<td>55.6</td>
<td>11.8</td>
<td>52.0</td>
<td>20.86</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Average scores represent the average percentage of correct responses for the test. PS children were not tested in English; PE children were not tested in Spanish.

Table 3. Average and standard deviation scores by language for group in Study 2.

Figure 2. Interaction of test performance (Spanish or English) by language group in Study 2.

Figure 3. Interaction of test performance (Spanish or English) by score type in Study 2.
that the monolingual score of the BE group ($M = 31.20$) was significantly lower than the conceptual score of the BS group ($M = 49.40$), $p = .035$, and the PS group ($M = 51.70$), $p = .018$ group as well as the monolingual score of the PS group ($M = 50.60$), $p = .025$. In contrast, there were no significant differences between groups when conceptual scores were used in the comparison. Therefore, for the Spanish version of the test, PS, BS, and BE scores were comparable when conceptual scoring was used.

**English comparison.** A repeated measures ANOVA comparing the PE, BS, and BE groups revealed main effect for score, $F(1, 27) = 7.043$, $p = .013$, $\eta^2_p = .207$. The effect size was small. Generally, children scored higher when Spanish was accepted (i.e., conceptual scores) than when only English responses were counted as correct. Although there were no significant group differences or interactions, examination of the scores indicates that the BS children who were dominant in Spanish demonstrated lower performance as compared to the BE and PE groups on the English version of this test, regardless of whether conceptual scoring was used or not. This has important clinical implications for determining language fluency in order to select the appropriate comparison group because it can affect classification accuracy. Further, the fact that children tended to score higher using the conceptual score illustrates the importance of considering both languages when testing.

**Classification analysis.** To evaluate how well the conceptual score worked to compare TD groups of monolingual and bilingual children, a classification analysis was conducted. Valencia and Suzuki (2001) stated that this sort of validity analysis is critical because single scores drive clinician’s day-to-day diagnostic decisions. The purpose of this analysis was to determine how many children who were TD (as indicated by parent and teacher report) were correctly classified as TD based on test performance. Diagnosis of language or communication impairment in many school-based settings relies heavily on performance on standardized tests. One common cutoff point is a score of 1.5 SD below the mean on different measures of language or communication performance (e.g., state standards). Thus, a cutoff score was set at −1.5 SD relative to the mean scores for the Spanish-dominant and English-dominant groups. Then, classification analyses measured the number of standard deviations that each bilingual child’s score was from the PS and PE group means in order to determine how many children fell below the −1.5 SD cutoff point using conceptual scores and monolingual scores. The rationale for using the scores from the monolingual groups was to evaluate how norms based on monolingual children from the same speech community can be applied to classification.

This classification analysis focuses on the role of conceptual scoring in classifying TD children only. Following the guidelines set by Plante and Vance (1994), classification below 80% is considered poor, 80%–90% is fair, and 90% and greater is good. Table 4 displays the classification rates by language group and score type. Classification for the BE children was in the good range for all groups. For the BS group, the conceptual score resulted in a change from poor to fair classification.

<table>
<thead>
<tr>
<th></th>
<th>Spanish conceptual score</th>
<th>Spanish monolingual score</th>
<th>English conceptual score</th>
<th>English monolingual score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BS</td>
<td>80</td>
<td>70</td>
<td>90</td>
<td>80</td>
</tr>
<tr>
<td>BE</td>
<td>80</td>
<td>50</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

In both Spanish and English, there were fewer bilingual children with scores that were greater than −1.5 SD below the mean when the conceptual score was used than when the monolingual score was used. This suggests that employing a conceptual score would result in reduced misclassification of TD children. However, as seen in Table 4, the difference was greater for the children in Spanish than in English. Consequently, conceptual scoring increased the classification accuracy of the English subtest and may allow for the use of the English version of the test with bilingual children over a range of language proficiencies with high degrees of clinical accuracy. If English dominance is clearly established before testing, administration of only the English version of the test along with conceptual scoring may be sufficient in estimating semantic skills with this test battery. For the Spanish subtest items, with conceptual scoring, classification accuracy increased from 50% to 80% for the BE group and from 70% to 80% for the BS group. The increased diagnostic accuracy observed for both groups with the use of conceptual scoring highlights its utility regardless of language fluency.

**GENERAL DISCUSSION**

It has long been recommended that clinicians fully evaluate both of a bilingual’s languages to help alleviate the problem of misidentification of LI in children from bilingual backgrounds. However, to interpret the findings of a dual-language evaluation accurately, information is needed for normative comparison. The purpose of these studies was to explore the response patterns of bilingual children on tasks of semantic development that were developed specifically for this population and to determine if a monolingual or conceptual scoring system would more adequately classify the performance of TD children from a variety of bilingual backgrounds.

**Implications for Understanding Distributed Vocabulary Knowledge**

Study 1 focused on the nature of children’s responses to a set of items designed to tap into the knowledge base of children from bilingual backgrounds and how the responses could be scored using total, monolingual, and conceptual scoring. The questions used in the task were open ended...
but tapped the same concepts in both languages. Thus, depending on the focus of the child’s responses, there was the possibility of observing overlapped responses or not. With such questions, relatively few overlapped responses were observed (as manifested in the similarity between the total scores and the conceptual scores in Study 1). This is consistent with the idea that children map concepts based on the contexts in which they learn them. One example of this was that children could have generated similar responses about trucks (in Spanish) and school buses (in English). In fact, children gave more description of trucks in Spanish and they gave more information about riding buses in English. These findings highlight the role of context of learning in the children’s responses.

Children from bilingual backgrounds, even those in the PS and PE groups, sometimes used vocabulary from each of their languages. In both studies, children had slightly higher conceptual scores than monolingual scores (see Tables 3 and 4). This was especially evident in the results of Study 2, which sampled a wide range of question types. When children code switched, they did so to add information and they rarely repeated a response. This pattern of responses suggests that bilingual children may operate under similar pragmatic and lexical constraints as do monolingual children (Davidson, Jergovic, Imami, & Theodos, 1997; Frank & Poulin-Dubois, 2002). This may help us understand the nature of children’s lower scores when they are given translated language tests. Children may view such repetition as a request for new information. Thus, they may not display their overlapping knowledge.

In the studies reported here, all children were participating in bilingual education programs that support transition to English. In this context, it is particularly likely that children would demonstrate greater knowledge of academic concepts in English, for example. Children in other types of bilingual educational programs may demonstrate a different distribution of vocabulary. For example, children in two-way immersion programs may have a more balanced distribution of academic vocabulary across their two languages. However, given that children in all groups demonstrated some code-switched responses, it is predicted that children in other types of bilingual education programs would code switch as well. Thus, they would be likely to benefit from a scoring system that used conceptual scoring.

Applying Conceptual Scoring

In Study 2, conceptual scores yielded better classification of TD performance than did monolingual scores. This suggests that conceptual scores hold promise as a holistic score upon which normative data could be generated. However, there are differences between the ways that children were tested in the current studies as compared to past studies that have employed conceptual scores (e.g., Marchman & Martínez-Sussman, 2002; Pearson et al., 1993), and our earlier work in which conceptual scores were employed (Peña et al., 2002). These differences center around (a) differences in the amount of overlapped material on the test, (b) the languages used in testing by the children and by the examiners, and (c) the potential effects of level of bilingualism.

In past studies in which conceptual scoring was applied, there was a large amount of overlap between the test items in the two languages; therefore, the conceptual score was based on the child’s displayed knowledge of the two lexical items. In the two studies presented here, the conceptual score was based on the production of correct items regardless of language of response. The conceptual scores were derived in each language from separate tests administered to the bilingual children. If it were possible to obtain similar results using a single test administration, this would potentially be more efficient because it would not require administration of the test battery in both languages.

Test selection might influence the application of conceptual scoring. For the conceptual score to function as it did in these studies, the test items must be selected to be representative of the knowledge base of children from bilingual backgrounds.

In Studies 1 and 2, the semantic tasks were administered in both languages by bilingual testers. Children were not instructed regarding language of response but were allowed to code switch. Thus, the bilingual children may have felt free to code switch in order to use their full range of vocabulary knowledge. The examiner had the knowledge to credit the child’s responses accurately regardless of the language in which the child responded. If a single test was administered by a bilingual examiner or by a monolingual examiner in one language, children might not draw on their knowledge of both languages in the same way that they do when they are tested in two languages. On one hand, this could be problematic because it would result in a different range of scores from monolingual and bilingual testing situations. Note, for example, that the Spanish-dominant bilingual children tended to code switch less when they were tested in English. On the other hand, it is problematic for clinicians, especially monolingual clinicians, who need to decide how to score responses across languages.

Overall, the findings of Study 2 suggest that conceptual scoring has a role to play in diagnostic decision making with single-language tests. However, it is necessary to specifically explore how the advantages of bilingual testing could be incorporated in a single-language situation. Examiners would need to have sufficient information to know when a response could be credited, and children would need to know that they could code switch.

In Study 2, the BS and BE children used their more dominant language approximately 65% of the time and their less dominant language approximately 35% of the time. The PS and PE children were using very little of their non-dominant language (approximately 4%). All groups, however, responded in both languages, as indicated by the fact that conceptual scores were higher than monolingual scores, as shown in Table 3. Additional information is needed to determine if the cutoff points (e.g., –1.5 SD, –2 SD) used when determining if children qualify for special education services are the most effective for school-age children, particularly those who are bilingual. Establishment of a cutoff point was not the focus of the current study. Ideally, cutoff points should be determined by comparing typical
and clinical (in our case, language impaired) performance using classification analyses such as discriminant functions analysis or logistic regression (Gray, Plante, Vance, & Henrichson, 1999; Hosmer & Lemeshow, 1989; Huberty, 1994; Sicoly, 1992). Further study is needed to determine across what range of bilingualism the conceptual score is stable and the extent to which it can be used to improve classification as it did in the current study.

Limitations

There are several limitations to the current study. The current study focuses on the correct classification of TD children. This work needs to be extended to explore correct classification of children with LI. Another difficulty in extending the current findings is that in this study, bilingual children were tested by bilingual personnel. It was the children’s awareness that the examiners were bilingual that permitted them to code switch. If children were tested in only one language or were not tested by bilingual personnel, they may be less likely to code switch. Given that children from most of the groups did code switch, it may be that if children do not respond using their whole range of knowledge, their scores may underestimate performance/knowledge. This would mean that it might be difficult for nonbilingual personnel to assess children without collaborating with bilingual personnel. Finally, it should be noted that this procedure is potentially useful for the diagnosis of LI, for the BESA in particular, but would not eliminate the need for assessment in both languages to determine intervention priorities.

In summary, these findings show that when bilingual children are tested with equivalent tasks but not translated tasks, they show relatively less overlapped vocabulary than they do when they are specifically asked the same questions in each language. Conceptual scores yielded from such testing were useful for correctly classifying TD children; further testing is needed to evaluate this procedure for children with LI.

ACKNOWLEDGMENT

This research was supported in part by Grant N01-DC-8-2100 from the National Institutes of Health (National Institute of Deafness and other Communication Disorders).

REFERENCES


Limitations

There are several limitations to the current study. The current study focuses on the correct classification of TD children. This work needs to be extended to explore correct classification of children with LI. Another difficulty in extending the current findings is that in this study, bilingual children were tested by bilingual personnel. It was the children’s awareness that the examiners were bilingual that permitted them to code switch. If children were tested in only one language or were not tested by bilingual personnel, they may be less likely to code switch. Given that children from most of the groups did code switch, it may be that if children do not respond using their whole range of knowledge, their scores may underestimate performance/knowledge. This would mean that it might be difficult for nonbilingual personnel to assess children without collaborating with bilingual personnel. Finally, it should be noted that this procedure is potentially useful for the diagnosis of LI, for the BESA in particular, but would not eliminate the need for assessment in both languages to determine intervention priorities.

In summary, these findings show that when bilingual children are tested with equivalent tasks but not translated tasks, they show relatively less overlapped vocabulary than they do when they are specifically asked the same questions in each language. Conceptual scores yielded from such testing were useful for correctly classifying TD children; further testing is needed to evaluate this procedure for children with LI.

ACKNOWLEDGMENT

This research was supported in part by Grant N01-DC-8-2100 from the National Institutes of Health (National Institute of Deafness and other Communication Disorders).
APPENDIX A. CHARACTERISTIC PROPERTY STIMULUS ITEMS

**English items**

Tell me three things about invitations.
Describe what a birthday cake looks like.
Tell me three things about a school bus.
Tell me what a dog looks like.
This is Anna. Tell me what she looks like.
What shape is the present?

**Spanish items**

Describe esta pelota. “Describe this ball.”
Describe esta flor. “Describe this flower.”
Describe esta muñeca. “Describe this doll.”
Describe esta troca. “Describe this truck.”
Dime como parece un lapiz. “Tell me what a pencil looks like.”
## APPENDIX B. FREQUENCY OF EXPRESSIVE ITEMS AND EXAMPLES OF EACH QUESTION TYPE

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th></th>
<th></th>
<th>Spanish</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># items</td>
<td>Example</td>
<td></td>
<td># items</td>
<td>Example</td>
<td></td>
</tr>
<tr>
<td>Characteristic properties</td>
<td>2</td>
<td>Tell me what she looks like.</td>
<td></td>
<td>4</td>
<td>Dime cómo es la pelota. (Tell me what this ball is like.)</td>
<td></td>
</tr>
<tr>
<td>Functions</td>
<td>3</td>
<td>What do you do with scissors?</td>
<td></td>
<td>6</td>
<td>Qué se hace con un bate? (What do you do with a bat?)</td>
<td></td>
</tr>
<tr>
<td>Analogies</td>
<td>2</td>
<td>Hot is to warm as cold is to _____.</td>
<td></td>
<td>1</td>
<td>Sandwich va con plato como limonada va con _____. (Sandwich is to plate as lemonade is to glass.)</td>
<td></td>
</tr>
<tr>
<td>Linguistic concepts</td>
<td>2</td>
<td>Tell Ana where Diego put the present with the big red bow.</td>
<td></td>
<td>4</td>
<td>Dime qué hicieron ellos después de comer. (Tell me what they did after they ate.)</td>
<td></td>
</tr>
<tr>
<td>Similarities and differences</td>
<td>4</td>
<td>What is different about these two invitations?</td>
<td></td>
<td>2</td>
<td>¿Qué es lo que es diferente entre estos pantalones? (What is different about these two pairs of pants?)</td>
<td></td>
</tr>
<tr>
<td>Comprehension of passage</td>
<td>1</td>
<td>What are Diego’s mother and sister going to do with the cake?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>