Assessing Lexical Development in Bilingual Babies and Toddlers*

Barbara Zurer Pearson

University of Massachusetts

*Acknowledgments**

This research was supported in part by NIH grants RO1HD30762 and RO1DC00484 to D. K. Oller, P.I.

Abstract

The introduction of vocabulary checklists for infant acquisition has allowed us to gather detailed information about early lexical growth from a broader number of bilingual children than before and to begin exploring the relation of growth in one language to growth in the other in a range of bilingual learning circumstances. Still, no standardized instruments to date give an adequate picture of normal bilingual development. Norms and milestones based on monolingual experience underestimate bilingual abilities in that they tap only a portion of the bilinguals' knowledge and credit them with less complex conceptualizations than what they actually possess. Double-language measures, like those proposed by Pearson, Fernández, and Oller (1993) and Muñoz-Sandoval, Cummins, Alvarado, and Ruef (1998), are an improvement over single-language measures as they encompass a greater portion of the bilinguals' knowledge, but they do not address the greater complexity of bilingual mental representations. What is needed are norms derived from observations of typically-developing bilingual children, followed up with measures of concurrent and predictive validity. However, bilinguals as a group are so diverse, it will be difficult to decide which subgroup(s) would be the appropriate reference to use for a standardization. In this review of recent studies, the difficulties involved in assessing bilingual vocabularies and recommendations for clinical practice are discussed.

Key words

bilingual babies and toddlers

lexical development

Introduction

Parents eagerly await children's first words; clinicians soon tally their growing inventories; and schools test vocabulary, in particular, through to adulthood. When children are learning two languages, attention is then focused on two lexicons. At the onset of language, this attention to the words children say is based on the fact that words are perhaps the first evidence of the children's knowledge of the language they are acquiring. Before they use the syntax evident in multiword combinations, children first demonstrate their use of single words (Fenson, Dale, Reznick, Thal, Bates, Hartung, Pethick, & Reilly, 1991, p. G10). Language-specific phonology, too, appears to postdate the acquisition of a small lexical stock in a given language (Navarro, Pearson, & Oller, 1997). Beyond the earliest stages, the lexicon continues to be a locus of development in language learning and use.

For monolinguals and bilinguals alike, lexical development is both distinct from language growth more generally and, at the same time, of a piece with it. Although no one

Address for correspondence

Dr. Barbara Pearson, NIH Project Manager, Department of Communication Disorders, University of Massachusetts, 117 Arnold House, Amherst MA01003, U.S.A. Tel: 413–545–5023; Fax: 545–0803; e-mail: <bpearson@comdis.umass.edu>

can speak a language without using the words of its lexicon, knowledge of words alone is insufficient evidence of its learning or use. As Esther Dromi's 2-year-old remarked about a friend she considered a baby: "She (does) not speak, she says only words" (Dromi, 1988, p.10). Further, one can still be said to be speaking a language even if all the words are not from that language. James Joyce demonstrates in *Finnegans Wake* (1939) that one can communicate, though barely, in sentences containing an astonishing percentage of lexical items drawn from other languages and private sources of meaning. Similarly, there are clinical conditions, like Williams Syndrome or hyperlexia, in which individuals have high levels of lexical knowledge, but low levels of language or cognitive function generally (Bellugi, Wang, & Jernigan, 1994). For most people, though, it is a safe bet that greater vocabulary growth will be associated with increased age and language experience. Older groups of children will know more and longer words than groups of younger children; and people with a larger fund of vocabulary can also be shown to have a wider acquaintance with the funds of cultural knowledge in that language. So, practically speaking, lexicon can be a reasonable reflection of competency in other aspects of language and learning.

In second language (L2) learning as well, lexicon tends to covary with other aspects of language knowledge. People with larger L2 lexicons generally have a greater potential to communicate more ideas in the second language using more elaborate grammatical structures. Not surprisingly, people with a larger experience of a second language tend to have more extensive vocabularies in that language, so much so that tests of second language achievement (like those from the Educational Testing Service) routinely include vocabulary assessment, and foreign language teachers everywhere employ the vocabulary quiz as a quick and easy method of seeing who has done the lesson.

In either first or second language assessment, vocabulary testing (and scoring) can be less elaborate than the testing of other language components. Typically, one takes a target list of words and the subject recognizes or produces some percentage of them. The number correct is compared to tables of the number correct that groups of similar individuals got on the same words. With children too young to answer on their own behalf, parent checklists of child vocabulary have gained widespread acceptance (Fenson, Dale, Reznick, Bates, Thal, & Pethick, 1994; Rescorla, 1989). Even without doing a special test, one can derive a vocabulary sample from spontaneous speech: One transcribes a recorded passage of a given length, counts the number and variety of lexical items used, and then compares it to some local or general reference number.

For normative purposes, this attention to vocabulary for first language learners is based on the strong empirical correlation observed between subtests of vocabulary and full-scale tests of intelligence (Terman, 1918; Wechsler, 1974) and between vocabulary and concurrent measures of other language behaviors (Hakuta, 1987). For second language learners, vocabulary measures show a strong correlation to other skills in the language and to exposure variables (Eilers, Cobo-Lewis, Mueller Gathercole, Oller, Pearson, & Umbel, 1997; Pearson, Oller, Umbel, & Fernández, 1996), and are thus often a proxy measure for stages of acquisition for L2. While knowledge of words is theoretically independent of skill levels in other intellectual and even language functions, in practice it most often is not. So it is a widespread and accepted practice to make inferences about general cognitive or academic functioning as well as predictions about further language development based on lexical comparisons to carefully selected norming samples thought to represent the

populations of interest. The practice is often used for bilinguals as well, although there are as yet, to our knowledge, no norming samples established for bilingual populations (see also Valdes & Figueroa, 1994, p. 85).

Until quite recently, the very earliest language development was not considered predictive of future performance, and so slow acquisition of vocabulary before age two was not used as a warning sign for language delay. The general prognosis was that a child under three who had learned fewer words than his or her peers would soon "catch up." Indeed, the wide variability in normal vocal development means that some children are observed to begin using words for labeling as early as 8–10 months while a full quarter of 16-month-olds have not yet begun to do so (Fenson et al., 1991, Figure 2). Tales of famous late talkers like Einstein, along with the low correlations of early productive vocabulary to numbers of other measures (Bates, Bretherton, & Snyder, 1988), reinforced the folk wisdom that very early lexical development was not particularly predictive of later development. It is now clear, though, that a large percentage of those at the lowest 10% in vocabulary at age two remain delayed in tests two and three years later (Dale, 1991; Fischel, Whitehurst, Caulfield, & DeBaryshe, 1989; Rescorla, 1989; Thal & Bates, 1988), and it is axiomatic that intervention is more effective the earlier it is started.

This new attention to early levels of vocabulary as a warning sign for delay puts added pressure on the research and clinical communities to explore the relations of lexical growth to other language growth in bilinguals. It also becomes imperative for the growing bilingual and immigrant populations in the United States and Europe to establish standards of normal development that are based on bilingual experience. The growth in immigrant populations in almost all parts of the world and increasing acceptance of multicultural ways of living are increasing the numbers of infant and childhood bilinguals (Romaine, 1995), who come to the attention of schools and clinicians. Milestones of language development derived from monolingual experience are being used as yardsticks of normalcy without adequate reflection about how such models may be inappropriate to describe normal bilingual development. The new attention to early levels of vocabulary as a warning sign pushes back the age at which monolingual norms may be inaccurately labeling bilingual children as "delayed" or in need of special resources. Research on young bilinguals has examined ways in which aspects of bilingual knowledge and experience are not captured by standard assessment tools (Umbel, Pearson, Fernández, & Oller, 1992; Fernández, Pearson, Umbel, Oller, & Molinet-Molina, 1992). With the new attention to early language development at age two and before, it is now more urgent to begin translating the research findings into practical guidelines for educational and clinical practice.

The inadequacy of existing norms for evaluating bilingual vocabulary

There is much evidence to indicate that normative guidelines based on monolingual populations make *IN*accurate predictions for bilingual children. One reason may be that a single lexical item occupies a different status in the mind of a bilingual than it does in a monolingual lexicon. We can only speculate about what supports the observed correlation of lexicon with more general aptitude or achievement measures in first and second language learning. From a theoretical point of view, a vocabulary measure is a yardstick for the number of concepts for which the child has a lexical representation. There is no general

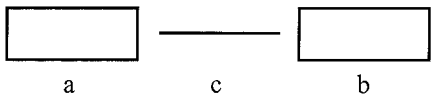

Figure 1

Schema of a lexical entry i e., a "sound-meaning pairing"
a. Representation of the word shape label (sound or sign)
b. Representation of the lexicalized concept
c Link

consensus about what a lexical representation is, but in a minimal definition it is a sound-meaning pairing on an unordered list (Chomsky, 1965). Thus, it involves, at the very least, a mental representation for the sound, a mental representation for the meaning, and a link between them. Figure 1 shows a possible model of a lexical entry in a mental lexicon, where each unit includes a "label," (for the sound or handsign for the word), and the lexicalized concept for the referent (an object, event, or relation), and some link between them.

When a monolingual checks one item on a vocabulary inventory, he or she is credited with one lexical "unit," which includes one label, one lexicalized concept, and one link between them; two units would include two of each component. A bilingual with two items checked is also credited with two units. These two units, though, include not 2+2+2 or 6 subcomponents, but 10 or 12. For two items known in both languages, there are four labels, four links, and either two or four lexicalized concepts, depending on whether the words in the two languages share the same mental representation or not. (E.g., some people report associating "bread" in English with Wonderbread and "pain" in French with a baguette. In that case, there would be two meaning representations per unit. If both the French and the English labels were associated with the same mental representation, there would be just one.) When one goes to compare how many words a child is considered to know compared to her playmate, is the key element for the comparison the number of labels, the number of lexicalized concepts, the number of links, or some combination of them all?

Very grossly, second language assessments most likely assume the prior existence of a conceptual organization and seek evidence of the number of *labels* for the concepts one knows in the new language. By contrast, first language assessments seek evidence about knowledge of a more complicated sort: the number of lexical representations one has—and wordshape labels are just one element in the lexical representation. Knowledge of a word in a first language assessment will implicate a greater degree of conceptual knowledge than what is tested on a second language assessment. That is, when a child does not know a first language item, it generally means she does not have a lexical entry. When she misses the same item on a second language assessment, it could be that she has no lexical entry, but might also mean that she is missing just the label for it in that language. For the monolingual, it is immaterial whether the measure is based on the full lexical representation or the label alone, as the two will be identical in number. Furthermore, no matter which factor is more crucial, a test in one language will tap the totality of the monolingual individual's lexicalized knowledge.

In contrast, a bilingual has, by definition, some knowledge of two languages. This poses three types of problems for the comparison to a norming sample implied in a monolingual vocabulary test. First, as just mentioned, a measure based on a single language will tap into only a *subset* of the bilingual child's total lexical knowledge. Even two single-language measures will quantify each language separately and will not give any measure encompassing both. What is needed is a measure of knowledge unique to each language

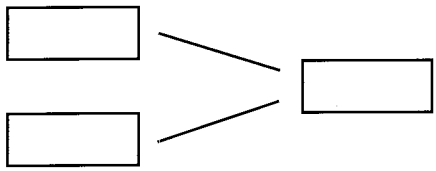

Figure 2a

Schema of bilingual lexical entry (single store model): One unit including two labels, one lexicalized concept, and two links

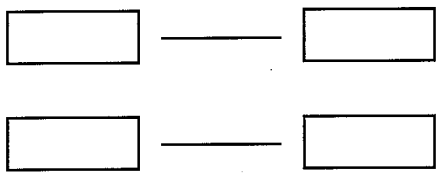

Figure 2b

Schema of bilingual lexical entry (double store model): One unit including two labels, two lexicalized concepts, and two links

separately that could not be captured in a single-language score of even the child's stronger language. Further, it is never clear with a bilingual whether the language being tested is a first language, a first language en route to becoming a "second" language, a second language, a second language en route to becoming a primary language, or some combination. Bilingual language use may represent some combination of first and second language development, or possibly a distinct mode of operation altogether (Grosjean, 1989).

Given this state of affairs, it is not clear which norms are more appropriate, first or second language norms. Infant bilinguals, in particular, are thought to be learning two languages as their first language (Meisel, 1990), and some undoubtedly are. However, despite careful terminological practices (as recommended by De Houwer, 1990, 1995; McLaughlin, 1984), some researchers claim to have observed second language strategies in a small number of bilingual children as early as two or younger (Karniol, 1990; Pearson & Fernández, 1994: p. 647). Such claims lead us to question the assumption of first language strategies for all very young bilinguals. Finally, young bilinguals may show a range of possible relations between the two languages, for example, those illustrated in Koler's (1963) conceptualizations of a bilingual's languages, and these will alter our ideas about what a lexical representation looks like and what it would "count for."

Children may have a single lexicon and a single conceptual store with two labels for each concept, one in each language, as in Figure 2a. Or they may have two separate conceptual stores with two sets of conceptual representations and two sets of word-labels associated with each concept, as shown in Figure 2b. In some cases, learning a word means creating the whole new lexical unit for it; at other times, the child would just be learning a second language label for a conceptual representation already established in the first language.

Further, a given test in a single language addresses only a portion of the child's lexical knowledge. Assuming that the bilingual child has two lexicons, each would consist of two parts: items known only in one language ("singlets") and another set of items known in both

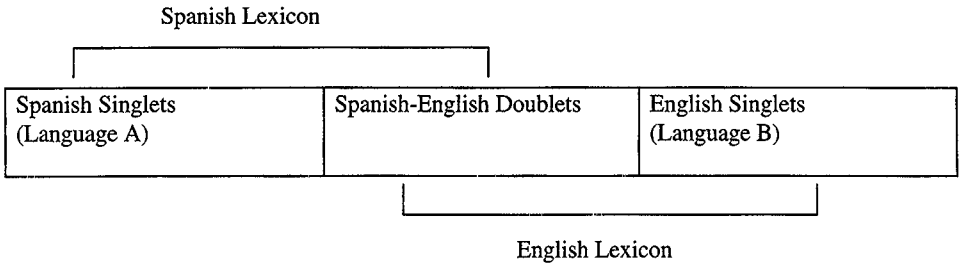

Figure 3

Relation of two lexicons of the bilingual.

("doublets"). The total lexicon, then, would comprise two sets of singlets and one set of doublets, as in Figure 3.

From this figure, one can clearly see that a measure of the child's English leaves out her singlets in Spanish; and vice versa for the Spanish. Thus either single-language vocabulary measure will be an underestimate for any bilingual child with singlets in both languages (as most children have been shown to have, Volterra & Taeschner, 1978; Pearson, Fernandez, & Oller, 1995).

So what do we know about a bilingual who knows a word on a test or a checklist? If the child knows two labels for the same object or action, one has no easy way of knowing whether the two labels represent just one or two separate mental representations. If it is a single lexical item with tags in two languages, it is likely *more elaborate* than the analogous single lexical item of a monolingual. For example, it might include information on the differences in the range of meaning between say "comida" in Spanish and "food" and "meal" in English, plus a tag to indicate which language was which. If the words are organized neurally as two separate lexical entries, each concept will involve a *greater number* of mental representations. Either way, the knowledge of a single lexical item known in two languages is more complex than knowledge of the same item in just a single language.

These considerations have more than academic import. Beyond the widespread use of vocabulary subsections in tests of intelligence (Terman, 1918; Weschler, 1974) and language or academic achievement (SAT, TOEFL, among others), vocabulary checklists are gaining popularity as screening devices for language delay, beginning at age two. Yet, these checklists have not yet been calibrated to the bilingual experience. Therefore, bilinguals from age two on may find themselves tested with instruments which tap only a portion of their knowledge and credit them for less complex conceptualizations than what they possess.

Measures of early vocabulary: Parent checklists

Two major initiatives for very early monolingual expressive vocabulary were developed in the late 1980s and have been rapidly disseminated: the MacArthur Communicative Development Inventory (MCDI, Fenson et al., 1994) and the Language Development Survey (LDS, Rescorla, 1989). They are vocabulary checklists that rely on a readily available source of information; parent reports of children's vocabularies. The forms

improve on earlier methods, like those requiring parents to keep diaries, in that parents are not asked to recall the words produced or understood by their children but just to recognize them from a standard list. They are also asked to make statements about current behaviors, and thus the potential bias of memory is removed from the task. Both forms have contributed enormously to our database of information about commonalities in children's early vocabularies. In particular, the standardization from the MCDI has allowed vocabulary measurement to be applied to a range of new applications in the lab and in the clinic.

The Language Development Survey (LDS) was designed by Rescorla in 1981 (Rescorla, 1991). It is a one-page double-sided survey consisting of a list of about 300 vocabulary words arranged alphabetically by semantic category (animals, food, vehicles, etc.), and takes about ten minutes to complete. The parent is asked to check off each word the child uses spontaneously. Its main objective is to be an efficient screening tool to discover children at risk for later language delay. Rescorla (1989, 1991) measured vocabulary levels with her instrument and then tested them for sensitivity and specificity against a "gold standard," such as the Reynell Developmental Language Scales (1977) and subsections of the Bayley (1969) and the Stanford-Binet (Thorndike, Hagen, & Sattler, 1986) and followed the children for many years to record the outcomes on subsequent measures. She focused on the 24-month mark when children were producing an average of 150 words. The guideline she developed was that children should be producing at least 50 words and should have begun to combine words by 24 months. The combining words criterion, in my opinion, needs to be subjected to further scrutiny (Pearson & Basinger, 1995), but several sources have converged on the 50-word landmark (Coplan, 1987; Paul, 1991). Rescorla experimented with different sized lists but found there to be little difference in the reported rate of delay. That is, the mean size of vocabulary was higher for a longer list, but children at the low end of the distribution marked the same few words on both longer and shorter forms (1989, p. 591). Also, Rescorla apparently invites parents to report words in any language and to add words the child says that are not on the list. Both practices give the impression that reports about any 50 words will do as a diagnostic. In fact, the guidelines are and should be based on the actual form used, and they generalize only roughly to other similar lists.

The MCDI (Fenson et al., 1994) is a much longer form, and it has been normed in a more comprehensive manner. There are 396 items on the CDI: Words and Gestures (originally called the "Infant" form for ages 8–15 months) probing both receptive and expressive vocabulary; and 680 words on the CDI: Words and Sentences, or "Toddler" form, ages 16–30 months, assessing only expressive language. (There are also many other useful sections on the MCDI, such as those seeking information about gestures and early grammar, for example, but they go beyond the scope of this paper and will not be discussed here.) The MCDI standardization covers the monthly intervals from 8 to 30 months with a norming population of 60 to 100 children at each age, comprised of equal numbers of boys and girls. Percentiles were developed from the descriptive statistics at each age to indicate whether a given number of words at a given age represents an advanced, average, or delayed level of development, with the lowest 5 or 10% giving the warning sign for delay.

Soon after the original norming project was completed, Fenson and his colleagues derived short forms of the MCDI: a 30-item screening form (Fenson, Ralston, and Sweet, 1994), an 89-word Level 1 based on the old "infant" form, and two 100-item Level 2 forms

based on the old "toddler" form (Fenson, Pethick, Renda, Cox, Dale, & Reznick, 1997). Reliability analyses based on the norms for the long forms permitted the short forms to "inherit" the norming potential of the long forms while providing an instrument that is much easier to administer. The short forms also minimize the literacy differences in the wider segment of the population that can be reached with them. The new norms being developed directly from the short forms, then, can potentially cover a wider socioeconomic range than did the original norming sample (Fenson et al., 1997).

General cautions

In using the vocabulary checklist norms, even for monolinguals, one must observe some basic cautions. Despite the name "inventory," even the long form of the vocabulary checklist on the MCDI is not an exhaustive inventory. A comparison of the terms included on the MCDI to a corpus of observational data reveals that even at the lower levels, the MCDI represents only a subset of the words children say. There are many everyday words not included on the form, "thing" and "color" (pointed out by Betty Hart) being two of the most common. In the Hart and Risley (1995) lexicons of 42 children from 9 to 36 months, even at the point when children had only 100 words each, MCDI words accounted for only two-thirds of the words observed, and the percentage diminished to about one-third as the children learned up to 600 words or more (B. Hart, personal communication). One assumes that almost all children will know some words that are not on the MCDI, but the standardization by number of words known *applies only to the standard sample on the published list*. One may like to invite parents to write down other words that their children say, but they should not be added to the number being scored.

This will be true even more strongly for the short forms. They were made by sampling from different difficulty levels on the long form, so one cannot compare 50 words checked off on the long form to 50 words on the Level 1 short form or to a similar number on one of the Level 2 forms which were engineered to include about 50% "harder" words, (i.e., words acquired later or by fewer children according to the original norming (Dale & Fenson, 1993; Fenson et al., 1994, Appendix B)). In addition, the discrepancies in the norms for overlapping ages—those found on both the infant and toddler forms—are another indication that the guideline is specific to a particular form. One can see that a 16-month-old boy in the 50% rank on the infant form would have 20 words of expressive vocabulary, whereas he would need 43 words to be at the 50% rank for the same age on the toddler form. A 16-month girl with 11 words would be in the bottom 5% on the toddler form, but above the 20th percentile on the infant form. Clearly the raw numbers of words reported must be evaluated with respect to which form they are reported on, and comparisons of number of words measured derived from different forms should be avoided.

It is also fallacious to compare groups of children on the raw number of words checked on the same form unless the children are all absolutely the same age and perfectly matched for gender. If the curve of expressive vocabulary development were linear (as it appears to be in a few segments of it but not in others), this practice would be less dangerous. Since development is not linear, increments of one or ten words, for example, contribute to the norming differently at different levels of proficiency, and those differences are lost when the raw number of words is used and the statistics for comparing groups of children on numbers of words assume linear relations. Similarly, the original norming

sample showed a small but significant difference in performance (ca. 5%) favoring girls, so much so that Fenson and his colleagues generated separate norms tables for each gender. Other authors who test for gender also report small but reliable gender differences (e.g., Patterson, 1997).

The potential problems created by these factors were illustrated in a recent research effort at the University of Miami. In a project involving typically-developing, hearing-impaired, and late-babbling children with normal hearing (Neal, 1997), parents had been asked to fill out MCDI Level 1 or Level 2 depending on the child's age and level of development (a policy that had served us well in previous projects). As a result, different numbers of girls and boys within the groups had lexical assessments based on different forms, and, as it happened, at slightly different ages. In order to be able to use the language variable, Neal used regression equations to transform female raw scores into "male-equivalent" scores. Then she changed the male-equivalent scores to "percentage" scores, dividing by the number of words on the form to reflect the number of alternatives the parents were working from. If a child knew 40 words on Level 2, that was counted as 40%, compared to 40 words on Level 1, which would count as 45%. This procedure ignores the fact that many words on Level 2 are harder than those on Level 1; thus knowing 40 Level-2 words is likely a more difficult achievement than knowing 40 Level-1 words. But Neal's goal was to err at every step toward the most conservative estimate of the difference between groups. I think we can have confidence in the significant differences she eventually found between the groups, but I would not recommend her method, and indeed, will have to change our recommendation about which level of the short form to use so that we do not get in that predicament again.

Finally, the percentiles given in the MCDI technical manuals may yet be subject to change as more information is learned about larger groups of children. As they are, the percentiles are useful for ranking individuals with respect to each other and to general growth guidelines, and the rankings of word frequency derived from the norming project (Dale & Fenson, 1993) are invaluable. Still, I think it may be premature to place too much faith on the actual percentiles of the original norming. For example, in a project of our lab in Miami, a group of 35 high-SES, typically-developing monolingual children, with average scores on a range of psychometric measures—the Bayley (1969), the Sequenced Inventory of Communicative Development (SICD, Hedrick, Prather, & Tobin, 1984), and the Wechsler Preschool and Primary Scale of Intelligence (WPPSI, 1974)—were in only the 34th percentile on the MCDI. Without a group of local monolingual controls, our conclusions about the rate of development of the bilinguals would have been different (Pearson et al., 1993).

Vocabulary tests for bilinguals

Even before the MCDI was available in English, an analogous form was being used for Italian (Caselli, Casadio, & Sanders, 1993). This is not merely a translation of the English form, but is based on the authors' knowledge of common patterns in child Italian, and it has its own norming data. A Spanish adaptation soon followed (Jackson-Maldonado & Bates, 1988; Jackson-Maldonado, Thal, Marchman, Bates, & Gutierrez-Clellan, 1993); more recently, Pearson and Rojas (1995) have made a (Miami-) Spanish adaptation of the short forms, Level 1 and Level 2A. There are also French, Finnish, German, Hebrew, Icelandic, Japanese, and Swedish forms, with others underway (Fenson, 1994), but only the Italian and

soon the Spanish (Swaine, Renda, Jackson-Maldonado, Thal, & Fenson, 1996) have independent norms.

In the absence of true norms, one must settle for approximating the quantitative guidelines learned from the English norms, applied WITH CAUTION to a language and culture they were not derived from. On general principles, one knows that a direct transfer of norms from one language to another or even just to another language community using the same language is not warranted. Tomayo (1987) documents that the same words are used with different frequency in different language communities, even those using the same language. The developers of the TVIP, for example, created one set of norms in Spain (Dunn, 1986) and then had to create a second set for Mexico and Puerto Rico, called the Adaptación Hispanoamericana (Dunn, Padilla, Lugo, & Dunn, 1986).

It can be suspected that there will be cross-linguistic differences in patterns of vocabulary acquisition from early on. One suggestion of how quantitative norming guidelines will differ across languages is given in Boysson-Bardies and Vihman (1991). The Swedish, French, and English-learning babies in their study all reached the 4-, 12- and 16-word stage at approximately the same age, whereas the Japanese babies were as much as a month or two slower. The authors note that, consistent with the phonotactics of Japanese, the words being attempted by the Japanese babies had more syllables than those attempted by the other babies. Thus, it was perhaps not surprising that the number of words learned by a given age was smaller. Bornstein, Tal, and Tamis-LaMonda (1991) have also reported lower vocal production and slower vocabulary acquisition among Japanese infants. Those authors invoke the different emphasis on verbal interaction in different cultures to explain the difference in rate. The Boysson-Bardies and Vihman study also involved only five children in each language, so it might just be an unrepresentative sample. Whatever the true explanation turns out to be, such observations highlight the dangers of translating norms from one language and culture to another, even at very early ages.

The MCDI forms in different languages permit a new range of cross-linguistic research on infants and toddlers. They can also be used in research to test bilinguals, but the application of their norms to bilinguals is perhaps even more problematic than to different languages. As with all tests, the application of monolingual norms to bilinguals should be avoided whenever possible (Valdes & Figueroa, 1994). Further, since there is evidence at older ages that expressive vocabulary levels as gauged by monolingual tests are normally quite different for typically-developing bilinguals and monolinguals (Romaine, 1995; Cobo-Lewis & Umbel, 1997), it is important to examine the same issues with the new instruments to determine whether the same principles are at work for the younger population as for the older.

In the testing of bilingual children, one sees basically two approaches: 1) test the child in one language, usually the majority language of the country, regardless of the child's background, and make comparisons to the norms for that language. In more enlightened situations, one might test in the child's stronger language, but that portion of the child's knowledge which is found in the other language is still ignored, and a noncomparable group is used for norms. 2) Some researchers have done analogous tests in two languages; usually the test is normed in one language and there is a homemade translation for the other, or the test is adapted to the other language, but there are no separate norms for it (e.g., the SICD, Hedrick et al., 1984; EOWPVT, Expressive One-word Picture Vocabulary Test, Gardner,

1979). For some very few tests there are separate normings in two languages; the Peabody Picture Vocabulary tests in English, the PPVT (Dunn & Dunn, 1981) and in Spanish, the TVIP (Dunn et al., 1986), the Woodcock (1991) and the Woodcock-Muñoz, (1995), and the Idea Oral Language Proficiency Tests (Ballard, Tighe, & Dalton, 1990) are three widely available examples. But even in that improved situation, it is not clear how to interpret the two scores for one individual.

Without norms in the second language, authors testing bilinguals have faced a dilemma in reporting the scores. Rosenblum and Pinker (1983), for example, tried to derive scores for both Hebrew and English from one administration of the PPVT, with half of the items given in English and the second half given in Hebrew (or vice versa). To compensate for the fact that there were half as many items as in a standard administration, they doubled the number of correct responses and added that number to the basal to find the "standardized" score for each language. They used the two scores to evaluate language balance in the children, but reported only the English scores, as they could not consider the nonstandardized Hebrew scores as equivalent to the "almost standardized" PPVT in English.

Doyle, Champagne, and Segalowitz's (1978) approach was similarly nonstandard. They used the PPVT and a "relatively unstandardized" French version of it that had been developed for the St. Lambert experiment (Lambert & Tucker, 1972), which they then adapted for a Quebec population. They apparently gave both tests to the children, but reported only the child's better language. In the first year of the experiment, not all the children, whose ages ranged from 17 to 42 months, were within the age range of the standardization (2;6 to adult), so the authors reported "mean number of words understood" by each group. Since the monolingual children were matched to the bilinguals by dominant language and age (among other variables), the scores reported included some French and some English scores in the average for each group. The difference was 8.6 words between the monolinguals and bilinguals, which was shown to be significant by a *t*-test. It is doubtful, however, that eight words of raw score means the same thing for 42-month-olds as it does for 17-month-olds. The next year, the children were older and so "standard" scores could be derived, 97 for the bilinguals and 109 for the monolinguals, again using the English norms for an undisclosed number of French subjects. In a subsequent regression analysis, the authors tried to compensate for the possibility that some of the bilingual children "were dividing their efforts between two languages" (p. 17), so they created a new variable, a total language score by "adding IQ scores" for French and English to get a PPVT total.

Investigators studying Spanish-English bilinguals have been in a more fortunate position as there are at least three tests that have "equivalent" separately normed versions in the two languages. These tests have been useful to show the shortcomings of using a simple monolingual measure for a bilingual child, and they point the way toward capturing what Doyle, Champagne, and Segalowitz were seeking, a measure encompassing knowledge in both languages. The Woodcock company is in the process of making it even easier to arrive at a double-language score, combining English with 13 other languages. They have developed the Bilingual Verbal Ability Tests (Muñoz-Sandoval, Cummins, Alvarado, & Ruef, 1998), which are scheduled for release in 1998. This adaptation of the Woodcock Verbal Proficiency Battery (Picture Vocabulary, Synonyms, and Antonyms, and Verbal Analogies) attempts to clarify the relation between the child's languages and let knowledge from more than one language contribute to the raw score.

Still, a double-language or composite measure, though an improvement over a single-language measure, has an undefined relation to a monolingual standardized score. How to bridge the distance between a double-language score and a monolingual norm is still an open question.

Measuring singlet and doublet vocabulary in bilinguals

In considering how to derive measures of development for bilinguals, the Language Project of the University of Miami determined to collect several pieces of information about bilinguals' lexical knowledge in order to derive not just one score, but a set of scores. Using this approach, we were able to greatly expand our knowledge of early bilingual lexical growth from a variety of perspectives. Until recently, most of our knowledge of early bilingual lexical acquisition had been gleaned from case studies, mostly of the children of linguists, a group whose language development is not considered typical. In that literature (see Pearson, Fernández, & Oller, 1995, for a review), there are several very complete records of individual growth, most notably those by Leopold (1939) and now Quay (1993b, and Deuchar & Quay, in preparation). But without a broader base of information about large groups of children, it has been very difficult to gain a perspective on what is typical and what is exceptional in early bilingual development. Therefore, from 1990–1992, we followed 25 babies from ages 3 to 30 months (and some of them beyond that). We recorded play sessions in each language monthly, obtained Bayley (1969), SICD (Hedrick et al., 1984), and Peabody (Dunn & Dunn, 1981; Dunn et al., 1986) scores, and collected multiple MCDIs from the parents.

For the lexical part of the study, we wanted to be able to ascertain when a test item was known in one language only, in both languages, or in neither. In the work reported in Pearson, Fernández, and Oller (1993), we made two single-language measures—English Vocabulary and Spanish Vocabulary—and two double-language measures—Total Conceptual Vocabulary (the "lexicalized concepts" above) and Total Vocabulary. This last corresponded to "labels" above, that is, the sum of English plus Spanish less the words reported for which the children had only one label that served them in both languages (e.g., [wawa] for both "water" and "agua." These last were estimated in consultation with the parents.) An illustrative example on how these scores were arrived at follows (from Pearson, Fernández, & Oller, 1995, p. 354). Consider a child with this vocabulary at 1;2:

ENGLISH	SPANISH
mama	***mama***
bear	***oso***
duck	*abuela*
more	*agua*
daddy	*sí*
no	*araña*

The Spanish vocabulary score is 6 words, and the English vocabulary is six. If *mama* were reported by the mother to be pronounced the same in both languages, the child's Total Vocabulary would be 11, 6 English plus 6 Spanish minus 1. Doublets would be *mama* and *bear-oso*, equalling 3, and singlets the other 8 words. The Total Conceptual score would be 6 in English plus 4 in Spanish (or vice versa), or 10. (The details of the doublet and singlet

percentage scores need not concern us here.) This broadened notion of bilingual vocabulary "scores" provides flexibility in evaluating a bilingual's lexical knowledge and illustrates the potential complexity involved in doing so.

With a fuller set of measures, we reasoned, we could calculate different scores depending on the purpose of the measurement. If we were comparing the number of lexicalized concepts different children knew, we would be able to access all of the lexical units the bilingual child knew (in both languages). If it was important to know the number of labels for concepts the child knew in English, we could pull that up, but would also have a better sense of how much of the child's knowledge was being ignored to get that figure.

Work with the PPVT and the TVIP paired tests with Spanish-English bilingual first-graders showed that there were both singlets and doublets in most children's receptive vocabularies (Umbel et al. 1992). Even children with a very small lexicon in one of their languages were seen to have some singlets in that language. Subsequent work (Pearson Andrews De Flores, Tu, & Cobo-Lewis, in preparation) has shown singlets present even in adult vocabularies, but the percentage declined from around 50% at first grade (age six) to around 30% at fifth grade (age 11) and 10% at college age. Using the two tests of receptive vocabulary, we derived standard scores in two languages as well as a rough estimate of the proportions of singlets and doublets for each child.

The estimation of doublets using the PPVT and the TVIP could be made only for a portion of the words on the test. Most of the items the children responded to were not pictured on both tests so we have no way of knowing if the child had a label for that concept in both languages, or indeed, whether it was a concept that is not lexicalized in both of the languages (like many food names, especially, which exist in only one culture or the other). Our procedure to estimate a child's doublet knowledge was as follows: First we identified all the "doublet opportunities," that is, those items that were pictured on both the PPVT and the TVIP. Since there are only 125 test questions on the TVIP as compared to 175 on the PPVT, the maximum number of doublets would be 125. Of that 125, only 63 use the same pictures, trying to elicit words which are translation equivalents of each other. But children rarely see all the plates of both tests. They begin at an age-defined point to establish a "basal." When they miss six out of eight items in a row, they are said to reach a "ceiling," and testing stops. So children rarely see the beginning or the end of either test. To get each child's doublet percentage, we first had to find the denominator, the number of doublet opportunities that the child was shown. The numerator was the number she or he actually got right in both languages. With the college students, we were less interested in the standard scores and so all students saw all the plates on both tests. Even for them, half of the TVIP and two thirds of the PPVT items were not included in this analysis.

The procedure could not be used with a preschool sample who had also been administered both tests (Fernández et al., 1992). The preschoolers in our sample were shown only around 20 of the test plates before a ceiling was reached and testing was terminated; therefore, the number of doublet opportunities they saw or answered was too small for meaningful statistics. Still, case studies indicate that even very young children know some of their words in both languages (Leopold, 1939; Quay, 1993a, 1993b; even Volterra & Taeschner, 1978), so we were encouraged to find another means to document them.

The MCDI offers a better opportunity to assess singlet and doublet vocabulary with a younger population, as regards receptive vocabulary up to 16 months and expressive

vocabulary up to 30 months. In creating an algorithm for finding translation equivalents, we were able to pair about 80% of the words on the two forms. Thus, for at least the major portion of the children's tested vocabulary we could tell whether the child knew the item in one or both languages. First, the parents or caregivers filled out two forms for the child. If the parent was bilingual, he or she filled out both; if not, then one parent filled out one form and the other filled out the other. After we entered the two forms into the computer, we made several passes through them to derive the English, Spanish, Total Conceptual, and Total Vocabulary scores, as described above.

Our procedure, implemented in Lotus, was very labor-intensive. Virginia Marchman at the University of Texas is currently working on automating it from electronic answer sheets, but even automated, the problem of interpretation is still unresolved. Which of the four scores calculated above, if any, is the proper basis for determining the standard score from the norms tables?

University of Miami Bilingual Lexicon Project: Key findings

In the University of Miami Language Project's lexical study of 25 Spanish and English bilingual-learning children between the ages of 8 and 30 months, individual graphs plotted the four measures for 18 of the children who had multiple MCDIs. These allowed us to confirm that the bilingual children followed the broad outlines of early lexical development as described for monolinguals in Fenson et al. (1994). Their growth was also similar to a group of 35 typically-developing monolingual controls being followed with similar regularity in a related study (Pearson et al., 1993). Looking at their total growth, we see that most of the bilingual children, like the monolinguals, started using words slowly around 12 to 14 months of age, and that somewhere around the middle or end of the second year, their growth accelerated. Only one of the 25 bilingual children was observed to be outside the range from Fenson et al. (1994) for producing an average of 50 words by about 18 months and at least 100 or more words of Total Vocabulary by 27 months. Thus, the broad quantitative expectations of vocabulary growth were met in the sample (Pearson & Fernández, 1994). Also, about two-thirds of the children showed a spurt in one language, or in the two languages combined, although none showed a spurt in each of the two languages at once. (The rate of growth required for such a simultaneous spurt is quite high and was shown by only three children in the cohort. In addition to learning that many words, they would have had to have their growth more evenly balanced between their languages in order to call their growth in each language a "spurt.")

How much growth was observed in each language in any given two-month period followed remarkably closely (r=.81) on the parents' estimates of how much of each language the child was hearing (Pearson, Fernández, Lewedag, & Oller, 1997). In the two or three cases where the proportions of language input changed dramatically from one observation point to the next, the proportion in the child's lexicon of each language appeared to follow the change, but with a two- or three-month delay before an increase in one language and a corresponding decrease in the other was observed.

The majority of the children appeared to concentrate growth in one language at a time. Only three children showed parallel growth in the two languages, while four appeared to

have almost all their words in one language and almost none in the other. The growth slope for the one language looked much like the slope of a monolingual's while the other showed very slow but consistent increases. The other eleven children observed at relatively close intervals showed a "complementary" pattern: When one language grew, the slope of the other language flattened or even dipped. Then the slower language would accelerate, while the other appeared to plateau (Pearson & Fernandez, 1994).

A key question answered by this research was the extent to which children learned translation equivalents (TEs) in their earliest vocabularies. A widely-cited claim from Volterra and Taeschner (1978) suggested that children would initially reject doublets, or at least fail to learn them. In this sample, doublet vocabulary, test items reported as known in both languages, was observed across the whole age range of the MCDI, (8 to 30 months) in 23 of 24 children, although children varied in the number of equivalent terms they had (Pearson, Fernández, & Oller, 1995). Only two children showed what appeared to be a doublet strategy, that is, an apparent preference for learning words in the second language that were already known in the first. Everyone who was not actually losing words in a language was adding some doublets. There was an average of 30.8% doublets overall (with a range from 0 to 100), and contrary to the claim in Volterra and Taeschner (1978), the mean percentage changed very little between 2 and 500 words.

Persisting problems in measuring vocabulary in young bilinguals

The application of a set of four measures of bilingual vocabulary represents an advance over a single measure, but the issue is far from resolved. For instance, one must remember that the measurement of singlets and doublets is very approximate. In our work, it was accomplished with a mapping between the Spanish and English long forms of the MCDI (Pearson & Fernández, 1992), but creating such a mapping is fraught with difficulty. Two languages do not carve up the world in the same way, so that even words used as translations will rarely be direct equivalents. Beyond the differences in the languages, there are idiosyncrasies on the forms of the MCDI. "Araña," for example, is included in the Spanish form, but "spider" is not found on the English form. Nonetheless, the rough approximation we achieved was adequate for our purpose. It helped us demonstrate that there is more to be measured than the single-language scores are able to capture. But no claims are made that our pairings are altogether correct, or that they could be used as the basis of a normed test.

Indeed, according to the guidelines of the American Psychological Association (AERA/APA, 1985) and Valdes and Figueroa (1994), there are no clearly appropriate measures of the language capabilities of young minority language children. The work done by our lab within this framework is exploratory, and aimed at demonstrating weaknesses in current measurement practices by showing possible avenues where they might be improved. Pearson et al. (1993, p.117) state that from the 60 children we observed we found no statistical basis for concluding that the bilingual children were slower to develop expressive vocabulary before the age of 30 months than were the monolinguals. Using the MCDI norms for English and the four measures outlined above, the bilinguals were lower, but not statistically lower than monolinguals when the bilinguals were measured in their *stronger* language. With the Total and Total Conceptual Vocabulary measures, there was a remarkable

degree of equivalence between monolingual and bilingual measures, considered either as percentiles when compared across a range of ages, or as a raw number of words, when the comparison was between children of the same gender and age in months.

However, the double-language measures are not without flaws. The biggest problem is the impossibility of using a checklist of the same size for one language on the one hand and two languages on the other. As discussed above, the scores (especially the scores in the middle and upper ranges) are a function of the size of the list they are drawn from (Rescorla, 1989). Equally damaging is the practice of using English norms for the Spanish word lists and also for double-language measures, as even we did in the 1993–95 work of the University of Miami Language Project described above. Our reports clearly state (e.g., 1993, p.104) that "the double-language measures have *no normative baseline* [emphasis added] [and] the percentiles derived were for purposes of illustration and comparison only." However, those cautions are often overlooked.

When researchers do not acknowledge these limitations, they risk overstating their findings. Rimel and Eyal (1996), for example, used the Pearson et al. 1993 framework to evaluate early vocabularies of 19 Hebrew-English learning children ages 18 to 30 months, as compared to 20 Hebrew-learning children. They used the MCDI and an unnormed Hebrew adaptation of it by Mai-Tal, Dromi, and colleagues. The study claims to show that the bilingual children have significantly lower single-language scores, but double-language scores which are comparable to monolinguals'. From the figures and charts in their article, this appears to be true, but despite their statistical testing, one cannot have full confidence in their results.

For one thing, the authors leave out crucial information about the language balance or language exposure variables of the children, and they fail to accommodate how language exposure variables may bias the single language evaluations of the bilinguals. Consider two children: one hears mostly English in his environment and the other is exposed to Hebrew approximately 80% of the time. Child A knows 100 words in English and 10 words in Hebrew; Child B knows 10 words in English and 100 words in Hebrew. Their average number of words in both languages is 55. Will one want to say that the bilingual children were significantly lower in each language than a monolingual average of say 105? It is unlikely that all the children in the study were as unbalanced as in this example, but the ranges reported, 21 to 456 in Hebrew and 31 to 593 in English, do not rule out the possibility. The authors do not appear to have averaged only the stronger single-language scores of the children for the single-language comparison, as we suggested would be more appropriate.

Rimel and Eyal also compare the mean number of words aggregated across the whole age range. The similarity of the bilinguals' Total Vocabulary to the monolingual score is indeed striking (371 words for the monolinguals' Hebrew vs. the bilinguals' 387 Total (TV) or 307 Total Conceptual, TCV), $p > .1$ for both; however, we have no indication that the similarity is consistent across the whole age range. The authors give no descriptive statistics by age to confirm that the general pattern of equivalence was not achieved by an interaction in the two sets of scores. Furthermore, a difference of 65 words will represent a greater discrepancy between groups at the earlier ages in their age range than at the end, so even though the authors report that they have matched the bilinguals by age and gender, they are not making the comparison *at* each age as they would be if they were using percentile norms.

One might also wonder how representative the groups were. The authors accepted for the study only children whose parents reported at least 20 words in each language. That criterion ends up being quite stringent for bilingual 18-month-olds, who are expected to have 40 words. (Monolinguals who have 40 words at 18 months are above the 30th percentile at that age, according to the English MCDI.) Whether or not one would quarrel with the double standard in the criterion, it is quite likely that the authors found a larger percentage of potential bilingual subjects ineligible, and thus they may have tested an unrepresentative sample of only the strongest bilinguals. Finally, the authors give little information about the nature of the forms beyond the fact that the Hebrew form has 70 fewer words on it. We do not know how Rimel and Eyal accomplished the mapping of one form onto the other to assess doublets and singlets for the TCV measure, but it is clear that the bilinguals had many more alternatives to mark on their forms. The highest Total Score (the maximum) for the bilingual group was 370 points higher than the highest possible monolingual score, and even the highest TCV score for the bilinguals was 12 points higher than the highest possible score for a monolingual. It is true that bilinguals *do* have more words to choose from in the real world, but this inequality does not make a good basis for psychometric conclusions. The authors further compromise the notion of a standard set of words by inviting parents to add words not found on the forms (p. 213). The checklist gives an estimate of the "true" vocabulary, and we know the estimate is sensitive to the number of alternatives, so it is important to try to keep their number constant.

Similarly, Patterson, in the *American Speech, Language and Hearing Association Special Interest Division 14 Newsletter* (Communication Disorders and Sciences in Culturally and Linguistically Diverse Populations, 1997), reports a mean number of words for a group of seventy-two 21- to 27-month-olds whose parents used a bilingual adaptation of the Language Development Survey (LDS) designed, we surmise, to find Total Vocabulary. (The form lists items with both English and Spanish translations for all but 10 items, which are considered to be the same in both languages, e.g., "pizza," or "McDonald's.") Patterson is looking for "sufficient data on reported vocabulary for clinical purposes" (p. 11). Based on that, and the fact that she uses the LDS, which is principally a screening form, it can be inferred that she is seeking a "cut-off" point—a warning sign for language delay. Indeed, it will be very interesting to learn what number of words the lowest 10 or 15% of this 24-month bilingual sample have checked off on that instrument, and what percent of the bilingual children at 24 months fall below the 50-word guideline (and no multiword combinations) derived from Rescorla's 300-word monolingual form.

Unfortunately, the preliminary report does not include that information. Instead, a mean is given at three ages, 21–22 months, 23–25, and 26–27, even though there is no information from the LDS giving descriptive statistics based on comparison samples of any ages except 24 months. If there were such information, we know from Rescorla's 1989 and 1991 articles that such statistics would no doubt be very specific to the number of words on the form and thus the Spanish/English total would be in an undefined relation to them. The 23–25 month mean for the bilinguals on this form (M=124, no SD given) is near the 125–150 word range that Rescorla found on the several forms in her study. Since this mean number of words is derived from a different length form, however, a similarity of the number may be accidental. Also, it may be more useful to report the median than the mean. We know from Fenson et al. (1991) that there is extreme variability expected in the scores with a greater percentage of

lower scores than higher, at least at some ages. Patterson gives the range at 21–22 months as 7 to 525. If we imagine a series of scores at that age for the purpose of an illustration (say, 7, 15, 35, 100, and 525), we can see the dramatic difference between these two alternative forms of the average. The mean for these five numbers would be 136 as opposed to the median of 35. If at least a standard deviation were given for the mean, in this case 220, the reader could also see that the mean was being heavily influenced by a single high score.

Finally, unlike Rescorla's reports, where the results from the screening form are checked against independent measures of language development (the Reynell and the picture naming sections of the Bayley and the Stanford-Binet in her case), Patterson does not report any concurrent or subsequent measures on the children to evaluate the information from the vocabulary screening. Further, girls and boys are aggregated in the same mean, even though Patterson corroborates with a *t*-test that there is a significant gender difference, thus implying that the genders need to be reported separately.

Wanted: Bilingual norms

Despite the problems mentioned above, Patterson is doing precisely what needs to be done for assessing early bilingual vocabularies: creating a quantitative baseline of what is observed in children who have no known language problem (other than single-language vocabulary levels below the levels observed in monolinguals). Rather than abandon vocabulary as a diagnostic tool, because it is too much affected by "experience and cultural background" as Restrepo suggests in the same issue of the ASLHA newsletter (1997), I feel we need to work with the present tools to make them more useful. The essential question is whether the same vocabulary size in a given language will give the same prediction of later growth for a two-year-old bilingual as for a monolingual.

There is little reason to doubt that vocabulary will be as useful an indicator of language level for bilinguals as for monolinguals (Pearson et al., 1996), but the specific prediction will be different for the different groups. My speculation, based on studies of older bilinguals on tests of scholastic aptitude and reading achievement (Pearson, 1993; Cobo-Lewis & Umbel, 1997), is that equivalent achievement will be predicted by lower levels of vocabulary in bilinguals. But this is ultimately an empirical question answerable only by closely observing large numbers of bilingual children and pairing vocabulary measures with concurrent and subsequent measures of other language skills.

In constituting a bilingual reference group, one should heed Patterson's suggestion to "explore the relationship between reported expressive vocabulary and family background variables" (1997, p.11). Just as gender and socioeconomic status are known to influence vocabulary levels in monolinguals (Hart & Risley, 1995), the bilingual group must add another variable, namely, language of the home. As differences between bilingual groups are often greater than the differences observed between some bilingual groups and comparable monolinguals (Cobo-Lewis & Umbel, 1997), the field needs more than one set of bilingual norms. At the very least, we need one set for children learning two languages from birth with fairly equal exposure to both languages in conditions of social advantage and another for children with equal exposure from lower socioeconomic levels. At older ages, one could use as reference points one set of children exposed equally to English and Spanish in the home and another set with only the minority language in the home, who are first exposed to the majority language at school and who continue to get most of their

exposure to that language outside the home. This will not begin to describe all the individual bilingual children and their language learning circumstances, but I think it will provide anchor points from which to judge children with other language backgrounds. It is also important not to collapse all the data into one number per child, but to preserve the notion of four scores: Language A, Language B, Total, and Total Conceptual Vocabulary. With this conceptual basis, we can better organize the more complicated data on bilingual lexicons and can use those data to answer questions from many different perspectives, depending on the purpose of our assessments.

In addition to improving our ability to measure expressive vocabulary, future research on bilingual lexical development should focus on receptive measures which can be used at earlier ages than those that are currently available. There is a gap between the upper bound of the MCDI Level 1 at 18 months and the lower bound at 30 or 36 months of the Peabody, Idea, and Woodcock-Muñoz (and similar tests in other languages). Indeed, assessment of comprehension may be as important as assessing production for identifying early delay at age two (Thal, Tobias, & Morrison, 1991; Thal & Tobias, 1992). The results from Pearson, Fernández, and Oller (1993), suggest that young bilinguals' receptive vocabulary skills in two languages may be equivalent to monolinguals' in each language and thus superior to monolinguals' in one. However, our analysis was based on the MCDI Infant form for only 12 subjects. Unlike for the expressive vocabulary measures, where correlations to the children's output in spontaneous speech samples let us corroborate the MCDI rankings, we had no independent means of reliably checking receptive scores. Hence, the findings were reported as tentative. Conceivably, the intermodal preferential-looking paradigm (Hirsh-Pasek & Golinkoff, 1996) could be adapted for vocabulary testing. Arriaga, Xu, and Carey (1996) report using a "pragmatically natural" looking paradigm with 10- and 14-month-olds to validate parent report of comprehension vocabulary. Only a few words could be tested and not all results were interpretable, but the technique shows promise both as a corroboration of checklists and as a diagnostic tool in its own right. Another promising possibility is a 2-choice PPVT-type protocol like the ones Thal and Fenson and their colleagues in San Diego have been exploring (D. Thal, personal communication). But we are still far from having a practical version of such a task suitable for toddlers with insufficient powers of attention for standard testing procedures. We are even further from having such tests with norms derived from typically-developing bilingual populations.

With respect to early assessment of language delay, there is hope that assessments of babbling onset and early social communicative behavior from scales like the Early Social-Communication Scales (ESCS, Mundy & Hogan, 1996) may soon be developed. The ESCS is a standardized "interview" where the infant sits in a chair or the caregiver's lap and engages in a sequence of object manipulations and turn-taking in response to toys presented by the examiner, who is seated directly across a table. The recorded protocols yield frequency, duration, and timing information about the child's responses to the objects and to the examiner. They are coded for behaviors involving eliciting attention, sharing joint attention, or responding to bids for attention within the context of social interaction or behavior regulation. Such assessments will be independent of vocabulary and independent of any particular language.

Research on children dubbed "late-babblers" suggests that delay in the production of mature (or "canonical") syllables, which is usually accomplished between 5 and 10 months

of age in children with normal hearing (Oller & Eilers, 1988), may indicate risk factors for language development that have been ignored until now. Oller and Eilers' lab has been following several cohorts of these children identified from a large-scale telephone screening program which relied on parent report and then confirmed the parents' judgments in the laboratory (Eilers, Neal, Oller, & Cobo-Lewis, in submission). As Fenson, Bates and their colleagues and Rescorla have shown for vocabulary checklists, preliminary reports show that parents are reliable judges of mature babbling, as exemplified by [bababa] or [dada], and delay in babbling appears to be associated with lower vocabulary at later ages (Neal, 1997).

The babbling measure will be a promising diagnostic tool for use with bilinguals. At least one study (Oller, Eilers, Urbano, & Cobo-Lewis, 1997) tested canonical babbling onset in bilinguals and found the mean age of onset to be within days of the monolingual mean. We know of no other similar analyses, but also no reports to the contrary. Conceivably, then, clinicians will soon have at their disposal a diagnostic sign which is not influenced by the number of languages being learned. It appears to also have the advantage of not being influenced by the social conditions in which the languages are being learned (Eilers, Oller, Levine, Basinger, Lynch, & Urbano, 1993).

Other research is underway to test the association of verbal development with late babbling as well as other nonverbal communicative behaviors. Particularly promising is the ability of the child to follow gaze at six months (Morales, Mundy, & Rojas, in press), to use gesture in communicative bids at 18 months (Neal, 1997), and to respond to communicative bids as tested within the ESCS (Mundy & Hogan, 1996). If these behaviors can predict later vocabulary and other language measures at 36 months, they may then have clinical value beyond their current uses. Kukkamaa, Pearson, and others are currently comparing these measures to 24-month vocabulary to see which behavior or constellation of behaviors can equal or surpass the correlation of 24-month vocabulary to 36-month measures for monolinguals. If a nonverbal communication behavior is found to have diagnostic value, it will probably still be somewhat "culture-specific" in that different societies have different strategies for how children are socialized and what behaviors they try to coax out of infants. But it may be less sensitive to the question of how many languages individuals within the same culture are learning than single-language vocabulary appears to be, and thus may have more general application.

Clinical implications

Vocabulary remains the easiest language skill to test and score, so vocabulary measures will, no doubt, continue to be an important screening and assessment tool. The introduction of forms to standardize parent reports has been an enormous advance for assessing children's earliest lexical development. These checklists or inventories allow relatively reliable assessments to be made from the onset of first words. Since the forms are so new, however, many important implications of their use have not yet been explored. Therefore, clinicians and researchers must be very careful when using them for situations beyond those they were designed for and tested with.

Several potential abuses of their underlying assumptions have been observed, and lead to these recommendations.

1. **Do not compare the "raw number" of words observed across individuals unless the children are all the same age and gender, and their tallies were elicited by the same instrument.**

Some advantage over using raw numbers can be gained by transforming a raw score into a percentile score (Fenson et al., 1991; Fenson et al. 1997), but these, too, are subject to some common misuses.

2. **Do not mix percentiles from different forms—even Level 1 and Level 2 MCDIs—as if they were the same score.**

Different percentiles derived for overlapping ages on different instruments illustrate the dangers of trying to use percentiles from different forms as if they were the same measure. Perhaps more work will be done in the future to unify the psychometric properties of the two forms, but for now, Level 1 percentiles and Level 2 percentiles need to be treated as separate variables.

3. **When making comparisons of raw numbers of words observed (i.e., for groups perfectly equated for gender, age, and elicitation form), one should generally report the median rather than, or as well as, the mean to show central tendency.**

4. **AVOID USING MONOLINGUAL NORMS FOR BILINGUAL CHILDREN WHENEVER POSSIBLE.**

If it must be done, remember that the bilinguals are being assessed for only a subset of their knowledge and credited for less complex conceptualizations than that which they possess. Even a Total Language score or a "bilingual administration" of a test, like the new Bilingual Verbal Ability Tests, is counting a bilingual lexical entry in a mental lexicon the same as a monolingual one. Unless or until it is shown otherwise, it is reasonable to assume that a given level of vocabulary for a bilingual will NOT stand in the same relation to other measures as for a monolingual. In particular, one cannot assume that the number which characterizes language delay at age two—whatever it eventually ends up to be—will be equally valid for bilinguals. (It most likely will not be.)

5. **Until bilingual norms are developed—and they are still a long way off—one should collect several pieces of information about each bilingual's lexical knowledge.**

Even though there are no guidelines for interpreting multiple scores, they are useful in reminding ourselves and others of the *multidimensional* nature of the bilingual's competence.

The failure of the educational community to have developed bilingual norms for tests—vocabulary or other—is not just a simple oversight. It can be attributed in part to the status of bilinguals, politically and conceptually. While there have probably been bilingual individuals since shortly after Babel, they have until recently been considered as monolinguals first, just ones who also had this interesting and useful skill that let them move between different language communities. Thanks to the increasing numbers and visibility of bilinguals in many communities around the world and the intense scholarship devoted to them in the last two decades, people are beginning to appreciate the point so well articulated by Grosjean (especially 1989) that there is something fundamentally special and/or different about bilinguals and their psycholinguistic processing. This new conception of bilinguals as a distinct group, along with the postwar emergence of a large and powerful testing industry, may explain why the demand for tests for bilinguals is a recent phenomenon.

Bilinguals, though, are not a single distinct group, but rather many distinct groups. In a recent study of monolingual and bilingual schoolchildren in Miami, bilinguals were tested in a factorial design with language of the home and language of the school as separate factors (along with socioeconomic status and age, Eilers et al., 1997). For many variables, differences in performance among the several bilingual subgroups were as great as or greater than those observed between some of the bilingual groups and the monolinguals. It is not obvious, though, to know which bilingual group or groups can serve as the reference for the others.

As we develop a better idea of the ways in which bilingual conceptual organization may differ from monolinguals', we can begin to operationalize them for assessment purposes. This new knowledge will need to guide our efforts in developing bilingual norms. In the meantime, it is important to appreciate what Grosjean (1989) means when he says a bilingual "is not two monolinguals in one person." One cannot expect to describe a bilingual performance standard by describing performance equivalent to two monolinguals. One does not expect a bilingual to be two people—not in a photo, and not on a checklist.

References

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, AMERICAN PSYCHOLOGICAL ASSOCIATION (1985). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association.

ARRIAGA, R. I., XU, F., & CAREY, S. (1996, April). *A pragmatically natural assessment of early noun comprehension*. Poster presented at the 10th International Conference on Infant Studies, Providence RI.

BALLARD, W. S., TIGHE, P. L., & DALTON, E. F. (1990). *Pre-IPT, IPT-1, and IPT-2. IDEA Oral Language Proficiency Test*. San Francisco: Ballard and Tighe.

BATES, E., BRETHERTON, I., & SNYDER, L. (1988). *From first words to grammar. Individual differences and dissociable mechanisms*. NY: Cambridge University Press.

BAYLEY, N. (1969). *Bayley Scales of Infant Development*. New York: New York Psychological Corporation.

BELLUGI, U., WANG, P., & JERNIGAN, T. (1994). Williams syndrome: An unusual neuropsychological profile. In S. Broman & J. Grafman (Eds.), *Atypical cognitive deficits in developmental disorders: Implications for brain function* (pp. 23–56). Hillsdale, NJ: Lawrence Erlbaum.

BORNSTEIN, M. H., TAL, J., & TAMIS-LAMONDA, C. S. (1991). Parenting in cross-cultural perspective: The United States, France, and Japan. In M. H. Bornstein (Ed.), *Cultural approaches to parenting* (pp. 69–90). Hillsdale, NJ: Erlbaum.

BOYSSON-BARDIES, B. de, & VIHMAN, M. M. (1991). Adaptation to language: Evidence from babbling and first words in four languages. *Language*, **67**, 297–319.

CASELLI, M. C., CASADIO, P., & SANDERS, L. (1993, July). *A parent report study of lexical and grammatical development in Italian*. Paper presented at the 6th International Congress for the Study of Child Language, Trieste, Italy.

CHOMSKY, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

COBO-LEWIS, A. B., & UMBEL, V. M. (1997). *Comparisons among monolinguals and bilinguals in standardized assessments of English proficiency*. Poster presented to the Society for Research in Child Development biennial meeting, Washington DC.

COPLAN, J. (1987). *The early language milestone scale*. Austin, TX: Pro-Ed.

DALE, P. S. (1991). The validity of a parent report instrument of child language at 20 months. *Journal of Child Language*, **16**, 239–250.

DALE, P. S., & FENSON, L. (1993) *LEX: A lexical development norms database [Computer program]* Seattle, WA: University of Washington, Department of Psychology.

DE HOUWER, A. (1990). *The acquisition of two languages from birth. A case study.* Cambridge, U.K.: Cambridge University Press.

DE HOUWER, A. (1995). Bilingual language acquisition. In P. Fletcher & B. MacWhinney (Eds.), *The handbook of child language* (pp. 219–250). Oxford: Basil Blackwell.

DEUCHAR, M., & QUAY, S. (in preparation). *Bilingual acquisition: Theoretical implications of a case study.* Oxford: Oxford University Press.

DROMI, E. (1988). *Early lexical development.* Cambridge, U.K.: Cambridge University Press.

DOYLE, A., CHAMPAGNE, M., & SEGALOWITZ, N. (1978). Some issues in the assessment of linguistic consequences of early bilingualism. In M. Paradis (Ed.), *Aspects of bilingualism.* (pp. 13–21). Columbia, SC, Hornbeam Press.

DUNN, L. M. (1986). *Test de Vocabulario en Imagenes Peabody: Adaptación Española.* Madrid, Spain: MEPSA (Calle Francos Rodriquez, 47, Madrid 28039).

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

DUNN, L., PADILLA, E., LUGO, D., & DUNN, L. (1986). *Test de Vocabulario en Imagenes Peabody—Adaptación Hispanoamericana.* Circle Pines, MN: American Guidance Service.

EILERS, R. E., COBO-LEWIS, A., MUELLER GATHERCOLE, V. C., OLLER, D. K., PEARSON, B. Z., & UMBEL, V. M. (1997, April). *Language and literacy in bilingual children.* Poster Symposium presented at the Society for Research in Child Development biennial meeting, Washington DC.

EILERS, R. E., NEAL, A. R., OLLER, D. K., & COBO-LEWIS, A. B. (in submission). *Late onset babbling as an early marker of abnormal development.* Department of Psychology, University of Miami, Coral Gables FL.

EILERS, R. E., OLLER, D. K., LEVINE, S., BASINGER, D., LYNCH, M. P., & URBANO, R. (1993). The role of prematurity and socioeconomic status in the onset of canonical babbling in infants. *Infant Behavior and Development*, **16**, 297–315.

FENSON, L. (1994). *Adapted versions of the MacArthur CDI.* Unpublished manuscript. Department of Psychology, San Diego State University, San Diego CA.

FENSON, L., DALE, P. S., REZNICK, J. S., THAL, D., BATES, E., HARTUNG, J. P., PETHICK, S., & REILLY, J. S. (1991). *Technical manual for the MacArthur Communicative Development Inventories.* San Diego, CA: San Diego State University.

FENSON, L., DALE, P. S., REZNICK, J. S., BATES, E., THAL, D. J., & PETHICK, S. J. (1994). *Variability in early communicative development.* Monographs of the Society for Child Development, **59** (5).

FENSON, L., PETHICK, S., RENDA, C., COX, J., DALE, P. S., & REZNICK, J. S. (1997). *Technical Manual and User's Guide for MacArthur Communicative Development Inventories: Short Form Versions.* Department of Psychology, San Diego State University, San Diego CA.

FENSON, L., RALSTON, J., & SWEET, M. (1994). *A screening form for detecting expressive language delay in toddlers* (Draft). Department of Psychology, San Diego State University, San Diego CA.

FERNÁNDEZ, M. C., PEARSON, B. Z., UMBEL, V. M., OLLER, D. K., & MOLINET-MOLINA, M. (1992). Bilingual receptive vocabulary in Hispanic preschool children. *Hispanic Journal of Behavioral Sciences*, **14**, 268–276.

FISCHEL, J., WHITEHURST, G., CAULFIELD, M., & DEBARYSHE, B. (1989). Language growth in children with expressive language delay. *Pediatrics*, **82**, 218–227.

GARDNER, M. F. (1979). *The Expressive One-word Picture Vocabulary Test (EOWPVT).* Novato, CA: Academic Therapy Publications.

GROSJEAN, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language*, **36**, 3–15.

HAKUTA, K. (1987). Degree of bilingualism and cognitive ability in mainland Puerto Rican children. *Child Development*, **58**, 1372–1388.

HART, B., & RISLEY, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Paul Brookes Publishing.

HEDRICK, D. L., PRATHER, E. M., & TOBIN, A. R. (1984). *Sequenced Inventory of Communicative Development (SICD) Test Manual, Revised*. Seattle: University of Washington Press.

HIRSH-PASEK, K., & GOLINKOFF, R. M. (1996). The intermodal preferential looking paradigm: A window onto emerging language comprehension. In D. McDaniel, C. McKee, & H. Smith Cairns (Eds.), *Methods for assessing children's syntax* (pp. 105–124). Cambridge, MA: MIT Press.

JACKSON-MALDONADO, D., & BATES, E. (1988). *Inventario del Desarrollo de las Habilidades Communicativas* [Communicative Development Skills Inventory]. San Diego, University of CA, Center for Research on Language.

JACKSON-MALDONADO, D., THAL, D., MARCHMAN, V., BATES, E., & GUTIERREZ-CLELLAN, V. (1993). Early lexical development in Spanish-speaking infants and toddlers. *Journal of Child Language*, **20**, 523–549.

JOYCE, J. (1939). *Finnegans wake*. NY: Viking Press.

KARNIOL, R. (1990). Second-language acquisition via immersion in daycare. *Journal of Child Language*, **17**, 147–170.

KOLERS, P. (1963). Interlingual word associations. *Journal of Verbal Learning and Verbal Behavior*, **2**, 291–300.

LAMBERT, W., & TUCKER, G. R. (1972). *Bilingual education of children: The St. Lambert experiment*. Rowley, MA, Newbury House.

LEOPOLD, W. F. (1939). *Speech development of a bilingual child: A linguist's record (Vol. 1)*. Evanston, IL, Northwestern University Press.

McLAUGHLIN, B. (1984). *Second-language acquisition in childhood. Vol. 1. Preschool children (2nd ed.)*. Hillsdale, NJ: Lawrence Erlbaum.

MEISEL, J. (Ed.) (1990). *Two first languages: Early grammatical development in bilingual children*. Dordrecht: Foris.

MORALES, M., MUNDY, P., & ROJAS, J. (in press). Following the direction of gaze and language development in 6-month-olds. *Infant Behavior and Development*.

MUNDY, P., & HOGAN, A. (1996). *A preliminary manual for the abridged Early Social Communication Scales*. University of Miami, Coral Gables, FL and Douglas Hospital Center, Quebec CA.

MUÑOZ-SANDOVAL, A. F., CUMMINS, J., ALVARADO, C. G., & RUEF, M. L. (1998). *Bilingual Verbal Ability Tests, Comprehensive Manual* (prepublication draft). Chicago: Riverside Publishing.

NAVARRO, A. M., PEARSON, B. Z., & OLLER, D. K. (1997, November). *Identifying the language spoken by 26-month-old monolingual- and bilingual-learning babies in a no-context situation*. Paper presented at the 22nd Annual Boston University Conference on Language Development, Boston MA.

NEAL, A. R. (1997). *The verbal and nonverbal communication skills of infants with late onset of canonical syllables*. Unpublished Master thesis, Department of Psychology, University of Miami, Coral Gables FL.

OLLER, D. K., & EILERS, R. E. (1988). The role of audition in babbling. *Child Development*, **59**, 441–449.

OLLER, D. K., EILERS, R. E., URBANO, R., & COBO-LEWIS, A. B. (1997). Development of precursors to speech in infants exposed to two languages. *Journal of Child Language*, **24**, 407–426.

PATTERSON, J. L. (1997). Expressive vocabulary of bilingual toddlers: Preliminary findings. *American Speech, Language, and Hearing Association (ASLHA) Special Interest Division 14 Newsletter*, **3**(1), 10–11.

PAUL, R. (1991). Profiles of toddlers with slow expressive language development. *Topics in Language Disorders*, **11**, 1–13.

PEARSON, B. Z. (1993). Predictive validity of the Scholastic Aptitude Test for Hispanic bilingual students. *Hispanic Journal of the Behavioral Sciences*, **15**, 342–356.

PEARSON, B. Z., ANDREWS DE FLORES, P., TU, S., & COBO-LEWIS, A. B. (in preparation). *Doublet vocabulary in bilingual vocabularies at successive ages, infant to adult.* Unpublished manuscript, Department of Psychology, University of Miami, Coral Gables FL.

PEARSON, B. Z., & BASINGER, D. (1995, April). *Criteria for delay in three populations of two-year-olds.* Poster presented to the Society for Research in Child Development, Indianapolis IN.

PEARSON, B. Z., & FERNÁNDEZ, S. (1992). *Rationale for English-Spanish CDI mapping.* Unpublished manuscript, University of Miami, Mailman Center for Child Development, Miami FL.

PEARSON, B. Z., & FERNÁNDEZ, S. C. (1994). Patterns of interaction in the lexical development in two languages of bilingual infants. *Language Learning*, **44**, 617–653.

PEARSON, B. Z., FERNÁNDEZ, S., LEWEDAG, V., & OLLER, D. K. (1997). Input factors in lexical learning of bilingual infants (ages 10 to 30 months). *Applied Psycholinguistics*, **18**, 41–58.

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.

PEARSON, B. Z., FERNÁNDEZ, S., & OLLER, D. K. (1995). Cross-language synonyms in the lexicons of bilingual infants: One language or two? *Journal of Child Language*, **22**, 345–368.

PEARSON, B. Z., OLLER, D. K., UMBEL, V. M., & FERNÁNDEZ, M. C. (1996, October). *The relation of lexical knowledge to measures of literacy and narrative discourse in monolingual and bilingual children.* Poster presented at the Second Language Research Forum, Tucson AZ.

PEARSON, B. Z., & ROJAS, J. (1995). *McDI Short Forms, Level 1 and Level 2, Spanish Adaptations.* University of Miami, Language Project, Coral Gables FL.

QUAY, S. (1993a). *Bilingual evidence against the principle of contrast.* Linguistics Society of America annual meeting, Los Angeles CA.

QUAY, S. (1993b). *Language choice in early bilingual development.* Unpublished doctoral dissertation. Cambridge, U.K.: Cambridge University.

RESCORLA, L. (1989). The language development survey: A screening tool for delayed language in toddlers. *Journal of Speech and Hearing Disorders*, **54**, 587–599.

RESCORLA, L. (1991). Identifying expressive language delay at two. *Topics in Language Disorders*, **11**(4), 14–20.

RESTREPO, M. A. (1997). Guidelines for identifying primarily Spanish-speaking preschool children with language impairment. *ASLHA Special Interest Division 14 Newsletter*, **3**(1), 11–12.

REYNELL, J. (1977). *Reynell Developmental Language Scales (Revised).* Los Angeles: Western Psychological Corporation.

RIMEL, A., & EYAL, S. (1996). Comparison of data on the lexical knowledge of bilingual versus monolingual toddlers collected by their parents using the MacArthur Communicative Development Inventory (CDI). *Speech and Hearing Disorders*, **19**, 212–219.

ROMAINE, S. (1995). *Bilingualism* (2nd edition). Oxford: Blackwell.

ROSENBLUM, T., & PINKER, S. (1983). Word magic revisited: Monolingual and bilingual children's understanding of the word-object relationship. *Child Development*, **54**, 773–780.

SWAINE, K. V., RENDA, D., JACKSON-MALDONADO, D., THAL, D., & FENSON, L. (1996, April). *Norms for the Spanish-language version of the MacArthur CDI.* Poster presented at the 10th International Conference on Infant Studies, Providence RI.

TERMAN, L. M. (1918). The vocabulary test as a measure of intelligence. *Journal of Educational Psychology*, **9**, 452–466.

THAL, D. J., & BATES, E. (1988). Language and gesture in late talkers. *Journal of Speech and Hearing Research*, **31**, 115–123.

THAL, D. J., & TOBIAS, S. (1992). Communicative gestures in children with delayed onset of oral expressive vocabulary. *Journal of Speech and Hearing Research*, **35**, 1281–1289.

THAL, D. J., TOBIAS, S., & MORRISON, D. (1991). Language and gesture in late talkers: A 1-year follow-up. *Journal of Speech and Hearing Research*, **34**, 604–612.

THORNDIKE, R. L., HAGEN, E. P., & SATTLER, J. M. (1986). *Stanford-Binet Intelligence Scale Fourth Edition*. Chicago: Riverside.

TOMAYO, J. M. (1987). Frequency of use as a measure of word difficulty in bilingual vocabulary test construction and translation. *Educational and Psychological Measurement*, **47**, 893–902.

UMBEL, V. M., PEARSON, B. Z., FERNÁNDEZ, M. C., & OLLER, D. K. (1992). Measuring bilingual children's receptive vocabularies. *Child Development*, **63**, 1012–1020.

VALDES, G., & FIGUEROA, R. (1994). *Bilingualism and testing. A special case of bias*. Norwood, NJ, Ablex.

VOLTERRA, V., & TAESCHNER, T. (1978). The acquisition and development of language by bilingual children. *Journal of Child Language*, **5**, 311–320.

WECHSLER, D. (1974). *Manual of the Weschler Intelligence Scale for Children—revised*. NY: Psychological Corporation.

WOODCOCK, R. W. (1991). *Woodcock Language Proficiency Battery: English form—revised*. Chicago: Riverside.

WOODCOCK, R. W., & MUÑOZ-SANDOVAL, A. F. (1995). *Woodcock Language Proficiency Battery: Spanish form—revised*. Chicago: Riverside.