LSHSS

Research Article

# Evaluating English Morpheme Accuracy, Diversity, and Productivity Measures in Language Samples of Developing Bilinguals

Irina Potapova,[a,b] Sophia Kelly,[a] Philip N. Combiths,[a,b] and Sonja L. Pruitt-Lord[a]

**Purpose:** This work explores the clinical relevance of three measures of morpheme use for preschool-age Spanish–English bilingual children with varying language skills. The 3 measures reflect accuracy, diversity (the tense marker total), and productivity (the tense and agreement productivity score [TAP score]) of the English tense and agreement system.
**Method:** Measures were generated from language samples collected at the beginning and end of the participants' preschool year. Participants included 74 typically developing Spanish–English bilinguals and 19 peers with low language skills. The morpheme measures were evaluated with regard to their relationships with other language sample measures, their ability to reflect group differences, and their potential for capturing morphological development at group and individual levels.
**Results:** Across both groups, the tense marker total and TAP scores were associated with other language measures and demonstrated both group differences and growth over time. The accuracy measure met few of these benchmarks.
**Conclusion:** The tense marker total and TAP score, which were designed to capture emerging morphological abilities, contribute valuable information to a comprehensive language assessment of young bilinguals developing English. Case examples are provided to illustrate the clinical significance of including these measures in assessment.

Substantial individual variation is characteristic of bilingual language development, including the acquisition and mastery of morphosyntactic skills (Paradis, 2005; Paradis & Crago, 2000; Paradis, Rice, Crago, & Marquis, 2008). Capturing and assessing these emerging skills is made more difficult by the dearth of clinical tools developed for culturally and linguistically diverse populations (e.g., Bedore & Peña, 2008; Caesar & Kohler, 2007; Gillam, Peña, Bedore, Bohman, & Mendez-Perez, 2013). Language sample analysis is a highly recommended assessment approach that is resistant to cultural and linguistic biases that are likely implicit in standardized assessments (Gutiérrez-Clellen, Restrepo, Bedore, Peña, & Anderson, 2000; Heilmann, 2010; Heilmann, Miller, & Nockerts, 2010; Hewitt, Hammer, Yont, & Tomblin, 2005; Rojas & Iglesias, 2009). However, to maximize the benefits of this culturally sensitive approach, appropriate language sample measures must be identified (Oetting et al., 2010; Stockman, 1996).

Presently, we explore the clinical utility of three measures of English morpheme use generated from the spontaneous language samples of preschool-age Spanish–English bilingual children with typical and low language skills. To do this, we examine whether these measures successfully track progress and/or capture differences across children with varying language abilities. Such investigations of English language measures are imperative and meet a practical need, as English is frequently used in the assessment of bilinguals in the United States (Caesar & Kohler, 2007; Gillam et al., 2013). In addition, two case examples are provided to demonstrate these measures in practice.

## *Broad Language Sample Measures*

Language sample analysis is important for evaluating and monitoring the language development of children from nonmainstream backgrounds, as formal assessments are widely regarded as inadequate for these populations (Bedore

[a]San Diego State University, CA
[b]University of California, San Diego
Correspondence to Irina Potapova: ipotapova@sdsu.edu

Editor-in-Chief: Shelley Gray
Editor: Janna Oetting
Received March 3, 2017
Revision received August 7, 2017
Accepted September 28, 2017
https://doi.org/10.1044/2017_LSHSS-17-0026
**Publisher Note:** This article is part of the Clinical Forum: Toward Accurate Identification of Developmental Language Disorder Within Linguistically Diverse Schools.

***Disclosure:*** *The authors have declared that no competing interests existed at the time of publication.*

& Pena, 2008; Caesar & Kohler, 2007; Paradis, Nicoladis, Crago, & Genesee, 2011; Stockman, 1996). Transcribed spontaneous language samples can be analyzed for many broad measures of language production (e.g., type–token ratio, percent intelligibility, grammaticality). Mean length of utterance in words (MLUw) and number of different words (NDW) are two traditional language sample measures that are clinically relevant and appropriate for use with culturally and linguistically diverse clients (Rojas & Iglesias, 2009). Mean length of utterance is associated with morphosyntactic development, and MLUw is considered a preferable measure for bilingual children because it is resistant to cross-linguistic differences in morphological richness (Gutiérrez-Clellen et al., 2000). NDW reflects the number of unique uninflected root words in the sample and is a measure of lexical diversity (Golberg, Paradis, & Crago, 2008).

Both MLUw and NDW are utilized in research and clinical settings for characterizing young bilinguals' productive language. These measures have been found to help identify language impairment in both bilingual and monolingual children (e.g., Bedore, Peña, Gillam, & Ho, 2010; Hewitt et al., 2005; Simon-Cereijido & Gutiérrez-Clellen, 2009). Conveniently, both measures can be automatically generated by transcription software after language samples have been transcribed and coded for bound and unbound morphemes. However, neither measure directly captures a child's development of tense and agreement (T/A) morpheme marking, which has been established as a salient indicator of language impairment (Bedore & Leonard, 1998; Gutiérrez-Clellen, Simon-Cereijido, & Wagner, 2008; Rice & Wexler, 1996).

## *T/A Morpheme Measures*

Difficulty with T/A morphology is a hallmark of language impairment in English-speaking children (e.g., Leonard, 2014). However, errors in morpheme marking are also to be expected in typically developing children acquiring English, whether they are young English monolinguals or children acquiring English in addition to another language (e.g., Rice, 2010). An understanding of morpheme marking in bilingual children—as well as adequate tools to measure morpheme use in this population—are needed to avoid mistakenly identifying typically developing bilingual children as having language impairment (Paradis, 2005; Paradis & Crago, 2000). Carefully characterizing morpheme use in bilingual children is also important because it is possible that influence from the native language will cause bilinguals' developmental trajectories to differ from monolingual peers with regard to sequence of acquisition or error types (Armon-Lotem, 2014; Nicoladis, Song, & Marentette, 2012; Paradis & Blom, 2016). To identify clinically relevant measures of morpheme use for preschool-age bilingual children, this study tested whether approaches based on measuring T/A morpheme accuracy, diversity, and productivity aligned with broad language measures, successfully tracked progress at group and individual levels, and reflected differences between children with varying language abilities.

### Accuracy of T/A Morpheme Marking

One traditional approach to measuring morpheme mastery is to calculate the accuracy of morpheme marking. In clinical and research settings, accuracy may be determined based on performance during a spontaneous language sample or on a probe designed to elicit targeted morphemes. In language sample analysis, accuracy measures require the transcription of a sample, followed by coding of all obligatory contexts for each morpheme of interest. Obligatory contexts are then manually reviewed to identify successful morpheme productions (e.g., the child says, "he walked" in a past-tense context), as well as morpheme omissions and other errors (e.g., the child says, "he walk" or "he walks," respectively, in a past-tense context). The number of successful morpheme productions is divided by the total number of obligatory contexts to produce a composite measure of morpheme accuracy. Composite accuracy measures thus collapse performance across multiple morphemes and reflect both correct and errored productions.

As a direct measure of morpheme use, accuracy rates have an important role in clinical decision making. Measures of T/A accuracy have been used to differentiate monolingual children with language impairment from typically developing peers (Bedore & Leonard, 1998; Gladfelter & Leonard, 2013; Rice & Wexler, 1996; Rice, Wexler, & Cleave, 1995), though these studies and others have differed in which morphemes are included in the composite and in other methodological considerations (Balason & Dollaghan, 2002). Limited available research also suggests that English T/A accuracy may differentiate bilingual children (4;5–6;5 [years;months]) with typical and atypical language development (Gutiérrez-Clellen et al., 2008).

However, accuracy may not be the most appropriate measure for all stages of language development. To illustrate, Fitzgerald, Rispoli, Hadley, and McKenna (2012) found that 41% of typically developing English monolingual children in a longitudinal sample demonstrated lower accuracy scores at 27 months of age than at 24. This phenomenon, "backtracking," can be explained by inflated accuracy rates at earlier time points due to the production of high-frequency combinations that do not require morphosyntactic processing or knowledge (e.g., *that's, it's, what's*; Guo, Spencer, & Tomblin, 2013; Rispoli, Hadley, & Holt, 2009). As bilingual children acquiring English may demonstrate similar acquisition patterns to younger monolinguals (e.g., Nicoladis et al., 2012; cf. Paradis & Blom, 2016; Rice, 2010), it is important to consider measures of morpheme accuracy in this population.

Furthermore, bilinguals' morpheme accuracy is also characterized by greater individual variability relative to age-matched monolingual peers (Gutiérrez-Clellen et al., 2008; Paradis, 2005; Paradis & Crago, 2000; Paradis et al., 2008). In addition, parallels in morpheme accuracy have been found between typically developing bilingual children and monolingual peers with language impairment, a group whose mastery of T/A marking is also delayed relative to typically developing monolinguals (Paradis, 2005). Altogether, there is motivation to investigate measures of morpheme use for bilingual children.

**Diversity and Productivity of T/A Morpheme Use**

The tense marker total and T/A productivity score (TAP score) were designed to better capture morpheme use for children in early stages of English language development. Introduced by Hadley and Short (2005), the two measures reflect contrastive uses of five morpheme categories that have been extensively studied in the literature on language impairment: (a) third-person singular (-3s: *drive/3s*), (b) past tense (-ed: *walk/ed*), (c) forms of copula *BE* (cop *BE: She is fast*), (d) forms of auxiliary *BE* (aux *BE*: *I am going*), and (e) forms of auxiliary *DO* (aux *DO*: *Do you like it?*).[1]

The tense marker total awards points for different surface forms for the five morphemes of interest: -3s, -ed, aux *DO* (*do, does, did*), cop *BE* (*is, am, are, was, were*), and aux *BE* (*is, am, are, was, were*). Higher scores thus indicate an ability to use an increasing number of unique surface forms. The TAP score awards points for each T/A morpheme provided that the child demonstrated sufficiently different productions of each one. Higher TAP scores indicate an ability to use T/A morphemes in increasingly unique syntactic contexts. Both measures were designed to capture onset of the T/A system (Hadley & Short, 2005); in addition, the tense marker total can be thought of as a measure of diversity or breadth of the T/A system, whereas the TAP score can be thought as a measure of productivity or depth. In contrast with measures of T/A accuracy, the scoring protocols for both the tense marker total and TAP score were designed to award points for morpheme uses that meet specific productivity criteria to safeguard against artificially inflated scores (Hadley & Short, 2005). Further scoring details are provided in the Method section.

The tense marker total and TAP score are valuable for measuring early English T/A development. For English monolinguals, these focused measures are correlated with broad language measures that include a wider variety of grammatical forms. Hadley and Short (2005) found that the tense marker total and TAP score were correlated with traditional language sample measures (e.g., mean length of utterance) for monolingual children ages 2;0–3;0 with low language and those at risk for language impairment. Furthermore, higher TAP scores predicted progress toward T/A mastery as measured by spontaneous language samples and standardized probes (Hadley & Short, 2005; Hadley, Rispoli, Holt, Fitzgerald, & Bahnsen, 2014; Rispoli et al., 2009). In addition, these measures are clinically relevant for monolinguals. Young children (2;0–3;0) at risk for specific language impairment had lower TAP scores than their peers, and their scores increased at a slower rate over time (Hadley & Holt, 2006). The tense marker total and TAP score have also been used to differentiate between typical and atypical language development in older English monolingual children (3;0–5;6; Gladfelter & Leonard, 2013; Guo & Eisenberg, 2014). Gladfelter and Leonard (2013) found that the tense marker total correctly identified 85.19% of the children aged 4;0–4;6 (23/27 participants) and that the TAP score correctly identified 82.14% of the children aged 5;0–5;6 (23/28 participants). In summary, the tense marker total and TAP score appear to be meaningful measures for children who are in the stage of development between first use of T/A morphemes and mastery of the T/A system.

However, these promising measures had not yet been considered in the context of dual language exposure. This study thus investigates whether these measures of diversity (i.e., the tense marker total) and productivity (i.e., the TAP score) can serve similar purposes for developing bilingual children who, like English monolinguals, undergo the process of acquiring English T/A morphemes.

## Present Study

This research was motivated by the need for appropriate measures of English language development for young bilinguals in the United States (Bedore & Peña, 2008; Bedore et al., 2018; Gillam et al., 2013). The present goal was to consider measures of English morpheme use in preschool-age Spanish–English developing bilinguals with typical language and with low language skills. Three morpheme measures from spontaneous language samples were considered: a composite T/A accuracy measure, the tense marker total, and the TAP score. First, we evaluate these measures on the basis of convergence with established measures; next, we explore their clinical potential to capture group differences and growth over time.

We ask, for preschool-age developing bilinguals with varying language skills, the following:

1. Do morpheme measures reflecting accuracy, diversity, and productivity relate to broad language sample measures?
2. Do morpheme measures reflecting accuracy, diversity, and productivity capture differences across groups and over time?
3. Do morpheme measures reflecting accuracy, diversity, and productivity successfully capture growth at the individual level (i.e., minimize backtracking)?

Given that the tense marker total and TAP score were designed to capture early stages of English T/A development and were proven relevant for assessing language in preschool-age monolinguals, we expected these measures to also be appropriate for preschool-age developing bilinguals who are learning English. Specifically, we expected the tense marker totals and TAP score to be higher for the typically developing bilingual group than for the low language group. We also expected scores to be higher at the end of the school year than at the beginning. Finally, we expected these measures to result in minimal backtracking when examined at the individual level (Rispoli et al., 2009).

[1]Each surface form of the auxiliary and copula verb paradigms (e.g., *am, is, are, was, were,* and *be*) is its own morpheme, and as such, it is appropriate to refer to these paradigms as morpheme classes or categories (Hadley et al., 2014). However, for brevity, we use the term "morpheme" to refer to -3s, -ed, cop *BE,* aux *BE,* and aux *DO*.

## Method

### *Participants*

Preschool-age Spanish–English developing bilinguals were identified from an ongoing community-based research project. For inclusion in this study, each participant was required to (a) be exposed to Spanish at home at least 30% of the time (Pearson, Fernandez, Lewedeg, & Oller, 1997), (b) score within normal limits on a nonverbal cognition measure, (c) complete language samples at the beginning and end of his or her preschool year, and (d) produce at least 10 complete and intelligible utterances in each language sample. In total, 93 children (mean age = 4;2, *SD* = 5.05 months) met these criteria and were included in the study.

Per parent report, participants were exposed to Spanish 72.35% of the time (*SD* = 20.11, range 40–100) at home, on average. Scores from the Figure Ground and Form Completion subtests of the Leiter International Performance Scale–Revised (Roid & Miller, 1997), a nonverbal cognition measure, were in the normal range, with an average score of 11.60 (*SD* = 1.93, range 7–16). In addition, maternal education, reported by the parents of 72 participants, was 10.01 years (*SD* = 2.90, range 3–16) on average. This maternal education range could be considered indicative of the entire sample as all of the children were enrolled at the same preschool site and the school setting required below-poverty standards to participate.

Eligible participants were then assigned groups based on language ability: developing bilingual children with typical language development (BiTD) and developing bilingual children with low language skills (BiLL). Group membership was determined by parent report (Gutiérrez-Clellen & Kreiter, 2003; Restrepo, 1998). Parents completed language questionnaires in their preferred language to provide information regarding their child's language experience and development, answering questions such as, "Do you or did you ever have any concerns about your child's speech and/or language?" Children were considered for the BiTD group if the parent reported no concerns. Conversely, participants with reported concerns were considered for the BiLL group. The BiTD group included 74 children (39 boys, 35 girls; mean age = 4;2, *SD* = 5.2 months); the BiLL group included 19 children (12 boys, seven girls; mean age = 4;1, *SD* = 4.5 months).

Subsequent comparison of language sample performance revealed expected group differences across a variety of English language sample measures at the beginning of the year. The BiTD group outperformed the BiLL group on MLUw and NDW, as well as total number of utterances, number of complete and intelligible utterances, type–token ratio, and percent intelligibility (all *p*s < .038, as evidenced by one-tailed *t* tests; see Table 1). By the end of the year, the BiTD group continued to demonstrate significantly higher MLUw. Importantly, the two groups were comparable on a number of factors that may be relevant to performance, including age, Spanish exposure at home, and maternal education (all *p*s > .464; see Table 1).

### *Procedure*

Data were collected in coordination with a community-based research study under the direction of the final author. Information about the study, consent forms, and language questionnaires were sent home with each child in English and Spanish through collaborative efforts with teachers and classroom personnel at a local preschool. Children whose parents returned signed consent forms were eligible for the larger study, which included participation in onsite data collection at the beginning and end of the academic year. To administer an assessment battery for the larger project, including collecting the language samples used for this research, multiple sessions were planned for each participant. Session length was determined by child engagement, with an upper limit of 40 min. During sessions dedicated to language sample collection, no other assessments or measures were completed. Data were collected by graduate students in speech-language pathology who were trained to administer the standardized assessments accurately, to collect spontaneous language samples, and to monitor child engagement. All children were tested individually, and child assent was obtained before each session. Each wave of data was collected in the span of 2–4 weeks.

### *Measures*

All measures of interest were generated from language samples collected at the beginning and end of an academic year (Time 1 and Time 2, respectively). Each language sample was elicited following a set play protocol, using a toy car, garage, and picnic sets, as well as a standard set of pictures for story retells. The digitally recorded language samples were orthographically transcribed and coded by trained research assistants following Systematic Analysis of Language Transcripts (SALT; Miller & Iglesias, 2012) conventions. In addition, the use of Spanish, as well as its potential impact on the measures of interest, was considered. Using a generous criterion (e.g., including utterances with only a single Spanish element; e.g., "put all your cintos"), only 6.05% of all complete and intelligible utterances included Spanish. Critically, the presence of Spanish utterances did not impact the calculation of the three English morpheme measures. On average, Time 1 samples included 99.32 complete and intelligible child utterances (*SD* = 59.43), and Time 2 samples included 145.35 (*SD* = 70.34). All measures were computed for all language samples at both testing points.

#### Broad Language Sample Measures

MLUw and NDW were automatically generated using SALT. The use of MLUw, as opposed to mean length of utterance in morphemes, is consistent with related research in bilingual children (e.g., Blom, Paradis, & Duncan, 2012; Paradis & Kirova, 2014). In addition, both MLUw and NDW are considered culturally sensitive and have been recommended for use with Spanish–English bilingual children (Rojas & Iglesias, 2009).

**Table 1.** Participant characteristics for BiTD and BiLL participant groups.

| | | | | | Broad language sample measures | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Background | | | MLUw | | NDW | | Total utterances | | Complete and intelligible utterances | | % Intelligibility | |
| Group | | Age | % Spanish heard | Maternal education | Time 1* | Time 2* | Time 1* | Time 2 | Time 1* | Time 2 | Time 1* | Time 2 | Time 1* | Time 2 |
| BiTD | *M* | 50.36 | 73.12 | 9.95 | 2.51 | 3.38 | 93.81 | 133.07 | 145.26 | 176.61 | 105.66 | 144.66 | 86.20 | 91.29 |
| | *SD* | 5.18 | 20.22 | 2.98 | 0.76 | 0.84 | 46.77 | 48.21 | 77.24 | 85.30 | 60.80 | 73.61 | 13.37 | 13.20 |
| BiLL | *M* | 49.47 | 69.35 | 10.27 | 2.16 | 2.95 | 68.89 | 130.16 | 109.00 | 183.42 | 74.63 | 148.05 | 80.16 | 92.21 |
| | *SD* | 4.54 | 19.97 | 2.66 | 0.73 | 0.74 | 41.75 | 43.39 | 63.32 | 73.55 | 47.42 | 57.48 | 11.75 | 6.45 |

*Note.* BiTD = bilingual with typically developing language; BiLL = bilingual with low language skills; MLUw = mean length of utterance in words; NDW = number of different words.
*$p < .05$.

**Accuracy of T/A Marking**

A composite measure of T/A accuracy (Pruitt & Oetting, 2009; Rice & Wexler, 1996, 2001; Rice, Wexler & Hershberger, 1998) was calculated to reflect productive marking in obligatory contexts. To most appropriately match the tense marker total and TAP score, the composite accuracy measure was calculated based on use of -3s, -ed, cop *BE,* aux *BE,* and aux *DO* in obligatory contexts, with overregularizations considered successful uses (Gladfelter & Leonard, 2013). Thus, the composite T/A accuracy measure was calculated by dividing the total number of correct uses and overregularizations by the total number of correct uses, overregularizations, omissions, and other errors. This proportion was multiplied by 100 to yield a percentage.

To facilitate scoring of morpheme accuracy, SALT morpheme codes were used to extract utterances containing obligatory contexts for the five morphemes of interest. Trained graduate research assistants categorized each obligatory context as a correct use (e.g., *play/ed* in a past-tense context), an overregularization (e.g., *break/ed* in a past-tense context), an omission (*e.g., play* in a past-tense context), or as an "other error" (including agreement, such as *they play/3s,* and tense errors, such as *he play/3s* in a past-tense context).

**Diversity and Productivity of T/A Morpheme Use**

Tense marker totals and TAP scores were generated following protocol outlined in Hadley and Short (2005), awarding points for contrastive uses of -3s, -ed, cop *BE,* aux *BE,* and aux *DO.* The tense marker total awards 1 point for each possible surface form of the five morphemes of interest, for a maximum score of 15. The TAP score awards up to 5 points for sufficiently different uses of each T/A morpheme, for a maximum score of 25. For -3s and -ed, sufficiently different uses are determined by the production of different lexical verbs (e.g., *want/3s* and *need/3s*). For the copula and auxiliary verbs, sufficiently different uses are characterized by the presence of different subjects (e.g., *the baby is* and *the mommy is*) or different surface forms (e.g., *the baby is* and *the baby was*). In other words, children are not able to earn points for repeated productions (e.g., *the baby is* produced multiple times).

For both the tense marker total and TAP score, points are awarded for correct uses (e.g., *play/ed* in a past-tense context) and for overregularizations (e.g., *break/ed* in a past-tense context), as both are indicative of productive use. Conversely, no points are awarded for other errors, including morpheme omissions, T/A errors, and productions with null subjects. Furthermore, scoring restrictions for copula and auxiliary verbs were established to ensure that the scored productions reflect grammatical encoding, as opposed to direct activation of common forms (Hadley & Short, 2005; Rispoli & Hadley, 2011). Contracted copula and auxiliary verbs are scored when used with nouns (e.g., *baby/'s hungry*), but not with pronouns (e.g., *she/'s hungry*; Hadley & Short, 2005; Rispoli et al., 2009). Uncontracted forms (e.g., *baby is hungry; she is hungry*) are always eligible for scoring.

Scoring procedures for the tense marker total and TAP score are demonstrated—and contrasted with morpheme accuracy—using the abbreviated language sample in Appendix A. The tense marker total for this abbreviated sample is 6: 1 point each for Utterances 1 (cop *BE, is*), 3 (aux *BE, is*), 7 (-3s, *looks*), 11 (cop *BE, am*), 12 (aux *DO, does*), and 13 (-ed, *play/ed*). The TAP score, which awards additional points for sufficiently different uses of the same surface form, is 8: 1 point for each of the utterances above, as well as additional points for Utterances 5 (auxiliary *BE, is*) and 15 (-ed, *break/ed*). Note that Utterances 6 and 14 did not contribute to the TAP score, as neither meets the criterion for sufficiently different use: Utterance 6, which includes an aux *BE* form contracted to a noun, repeats a subject/surface form (*Daddy/'s* was awarded a point in Utterance 5), and Utterance 14 repeats a lexical verb (*play/ed* was awarded a point in Utterance 13). Following scoring procedures for both measures, no points were awarded for errored productions (Utterances 4 and 9) or for forms contracted to pronouns (Utterance 2).

In contrast to the tense marker total and TAP score, a measure of morpheme accuracy would take into consideration all utterances in the abbreviated sample. The number of correct productions and overregularizations, 11, would be divided by the total number of obligatory contexts, 15, and multiplied by 100 to yield an accuracy rate of 73.33%. Unlike the diversity and productivity measures, this approach both rewards productions in repeated contexts (e.g., Utterances 5 and 6; Utterances 13 and 14) and reflects both successful and errored productions.

To facilitate scoring, SALT codes were used to extract utterances with relevant T/A morphemes. The samples were hand scored by trained research assistants and the first and second authors to identify contrastive uses of the five target morphemes.

**Reliability**

Steps to ensure data reliability were taken at each level of transcription, coding, and scoring. All research assistants received training relevant to their assignment (transcription, coding, and/or scoring) and completed sample tasks to a satisfactory criterion prior to contributing to data processing. An adapted consensus procedure was utilized for transcription and coding (e.g., Eisenberg, Guo, & Germezia, 2012). After a trained research assistant transcribed a language sample, a second research assistant independently reviewed the transcript while listening to the corresponding audio file. All transcribers were instructed to mark an utterance as unintelligible if they were not able to transcribe the utterance after listening to the audio three times. Research assistants trained in coding protocol then coded the agreed-upon transcriptions for bound and unbound morphemes following established lab procedures. As with transcription, each sample was then independently reviewed for coding conventions by a second trained research assistant. Disagreements were resolved by referencing training materials

or by appealing to a third transcriber/coder (the first or final authors) if needed. Coding was further reexamined during subsequent scoring procedures, as calculating the composite accuracy measure, tense marker total, and TAP score required the manual review of utterances containing T/A morphemes (Hadley & Short, 2005). Over 50 samples across the two testing points were independently scored for the composite T/A accuracy measure; average reliability was 94.75%. Over 25 samples across the two testing points were independently scored for tense marker totals and TAP scores; average reliability was 95.73%. All research assistants were blind to group status and children's performance on other measures.

## Results

*Do morpheme measures reflecting accuracy, diversity and productivity relate to culturally sensitive broad language sample measures?*

Correlational analyses were conducted to test for convergence between the three morpheme measures of interest (composite accuracy, tense marker total, and TAP score; see Table 2) and the two culturally sensitive broad language sample measures (MLUw and NDW) at each time point for each group. Correlation coefficients of .2, .4, .6, and .8 were considered benchmarks for weak, moderate, strong, and very strong relationships, respectively (Evans, 1996).

There was little evidence of convergence for the composite accuracy measure and broad language skills across the two groups (see Table 3). For all analyses including the composite accuracy measure, children with fewer than three obligatory contexts for the five morphemes of interest were excluded, as accuracy could not be reliably calculated (Balason & Dollaghan, 2002). As a result, 53 BiTD and 10 BiLL participants were included in analyses involving the composite accuracy measure at Time 1; 72 BiTD and all 19 BiLL participants were included at Time 2. The only relationship to reach significance at Time 1 was between the composite accuracy score and MLUw for BiTD participants ($r = .281$, $p < .05$). At Time 2, accuracy was only significantly related to NDW for BiTD participants ($r = .419$, $p < .01$). For BiLL participants, accuracy was not significantly correlated with MLUw or NDW at either time point.

Analyses for the tense marker total and TAP score included all participants in each group. Both measures demonstrated consistent and significant convergence with broad language measures (see Table 3). At Time 1 for BiTD participants, the diversity and productivity measures demonstrated strong positive correlations with MLUw and NDW ($r$s = .658–.757, $p$s < .01). At Time 2 for this group, moderate to very strong positive relationships were demonstrated ($r$s = .500–.826, $p$s < .01). For BiLL participants, the diversity and productivity measures were moderately to strongly correlated with MLUw and NDW at Time 1 ($r$s = .531–.682, $p$s < .05) and Time 2 ($r$s = .537–.758, $p$s < .05).

Furthermore, the three measures of morpheme use were also related at both time points. Accuracy was related to the tense marker total and TAP score for BiTD participants at Time 1 and Time 2 ($r$s = .328–.520, $p$s < .001) and for BiLL participants at Time 2 ($r$s = .555–.637, $p$s < .05). The strength of these relationships was greater at the second time point.

*Do morpheme measures reflecting accuracy, diversity, and productivity capture differences across groups and over time?*

To address our second question, 2 × 2 analyses of variance that included participant group (BiTD vs. BiLL) as a between-subjects factor and time (Time 1 vs. Time 2) as a within-subject factor were conducted for each morpheme measure. Participants with fewer than three obligatory contexts for the T/A morphemes were again excluded from analyses for the composite accuracy measure (Balason & Dollaghan, 2002), but not for the tense marker total or TAP score.

For the accuracy-based measure of morpheme use, no significant main effect of group or time emerged, nor was the interaction significant ($p$s > .156). That is, accuracy rates were comparable across BiTD and BiLL participants, and scores were not indicative of growth over the course of the academic year (see Figure 1).

For the diversity and productivity measures, both main effects were significant (see Figures 2 and 3): Tense marker totals were higher for BiTD participants than for BiLL peers, $F(1, 91) = 4.621$, $p = .034$, $\eta_p^2 = .04$, and scores increased from Time 1 to Time 2, $F(1, 91) = 92.603$, $p < .001$, $\eta_p^2 = .408$ (see Figure 1). Similarly, TAP scores were higher for BiTD participants than for BiLL

**Table 2.** Performance on the three morpheme measures at Time 1 and Time 2 for BiTD and BiLL participants.

| Group | | Composite accuracy measure† | | Tense marker total | | TAP score | |
|---|---|---|---|---|---|---|---|
| | | Time 1 | Time 2 | Time 1 | Time 2 | Time 1 | Time 2 |
| BiTD | *M* | 63.52 | 66.62 | 2.43 | 4.95 | 4.41 | 9.31 |
| | *SD* | 20.30 | 18.65 | 2.20 | 2.53 | 4.63 | 5.77 |
| BiLL | *M* | 56.25 | 58.65 | 1.37 | 3.79 | 2.26 | 6.89 |
| | *SD* | 29.64 | 19.34 | 2.11 | 2.35 | 3.96 | 5.03 |

*Note.* BiTD $n = 74$; except †, $n = 53$. BiLL $n = 19$; except †, $n = 10$. BiTD = bilingual with typically developing language; BiLL = bilingual with low language skills; TAP score = tense and agreement productivity score.

**Table 3.** Pearson correlations between MLUw, NDW, tense marker total, TAP score, and composite accuracy measure at Time 1 (above the diagonal; in roman) and Time 2 (below each diagonal; in italics) for BiTD and BiLL participants.

| Group | Measure | MLUw | NDW | Tense marker total | TAP score | Composite accuracy measure† |
|---|---|---|---|---|---|---|
| BiTD | MLUw | — | .758** | .708** | .658** | .281* |
| | NDW | .498** | — | .754** | .757** | .055 |
| | Tense marker total | *.556*** | *.825*** | — | .923** | .328* |
| | TAP score | *.500*** | *.793*** | *.912*** | — | .390** |
| | Composite accuracy measure† | *.092* | *.419*** | *.520*** | *.517*** | — |
| BiLL | MLUw | — | .833** | .531* | .579** | .544 |
| | NDW | *.598*** | — | .643** | .682** | .449 |
| | Tense marker total | *.542** | *.794*** | — | .951** | .569 |
| | TAP Score | *.537** | *.758*** | *.929*** | — | .533 |
| | Composite accuracy measure† | *.292* | *.246* | *.555** | *.637*** | — |

*Note.* BiTD $n$ = 74; except †, $n$ = 53 for Time 1 and 72 for Time 2. BiLL participants = 19; except †, $n$ = 10 for Time 1. BiTD = bilingual with typically developing language; BiLL = bilingual with low language skills; MLUw = mean length of utterance in words; NDW = number of different words; TAP score = tense and agreement productivity score.

*$p$ < .05. **$p$ < .01.

peers, $F(1, 91) = 4.283$, $p = .041$, $\eta_p^2 = .045$, and scores increased from Time 1 to Time 2, $F(1, 91) = 44.870$, $p < .001$, $\eta_p^2 = .330$. The interactions were not significant for either measure.

As an additional test for group-level patterns, performance for each target morpheme was considered. As the TAP score is composed of scores ranging from 0 to 5 for -3s, -ed, cop *BE,* aux *BE*, and aux *DO,* it is possible to compare productive uses of each morpheme at the beginning and end of the year. Five 2 × 2 analyses of variance that included participant group (BiTD vs. BiLL) as a between-subjects factor and time (Time 1 vs. Time 2) as a within-subject factor were conducted for each morpheme. Productive use of each morpheme increased from the beginning to the end of the school year (see Figure 4), as evidenced by a main effect of time: -3s, $F(1, 91) = 7.561$, $p = .007$, $\eta_p^2 = .077$; -ed, $F(1, 91) = 22.225$, $p < .001$, $\eta_p^2 = .196$; aux *DO*, $F(1, 91) = 6.902$, $p = .010$, $\eta_p^2 = .070$; cop *BE,* $F(1, 91) = 23.511$, $p < .001$, $\eta_p^2 = .205$; and aux *BE,* $F(1, 91) = 38.034$, $p < .001$, $\eta_p^2 = .295$. In addition, BiTD children outperformed BiLL peers on productions of -3s, $F(1, 91) = 4.244$, $p = .042$, $\eta_p^2 = .045$, and aux *BE,* $F(1, 91) = 5.202$, $p = .025$, $\eta_p^2 = .054$. No significant interactions emerged for any morpheme.

The composite accuracy measure may also be separated into accuracy measures for each target morpheme. However, such analyses were not feasible for the present data set due to the limited obligatory contexts (e.g., the average number of obligatory contexts for -3s, -ed, and aux

**Figure 1.** Composite accuracy measure rates at Time 1 and Time 2 for both participant groups. BiTD = bilingual with typically developing language; BiLL = bilingual with low language skills.

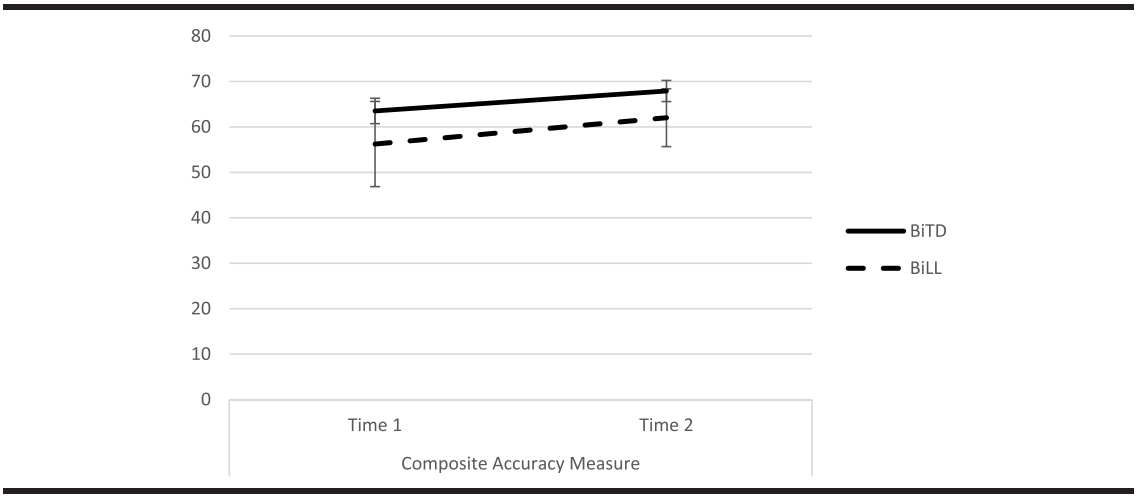
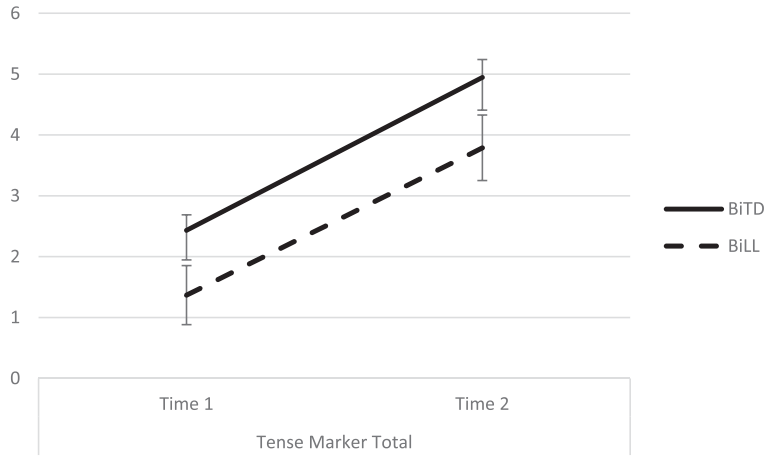


**Figure 2.** Tense marker totals at Time 1 and Time 2 for both participant groups. BiTD = bilingual with typically developing language; BiLL = bilingual with low language skills.




*DO* at Time 1 for BiTD participants was 2.38, 1.58, and 1.73, respectively; see Table 4).

*Do morpheme measures reflecting accuracy, diversity, and productivity successfully capture growth at the individual level (i.e., minimize backtracking)?*

Recall that backtracking is a phenomenon in which participants demonstrate lower scores at later time points. This pattern has been identified in the accuracy rates of young monolinguals acquiring the English T/A system (Fitzgerald et al., 2012; Rispoli et al., 2009). To evaluate backtracking in the present sample, each participant's performance on the three morpheme measures was compared at Time 1 and Time 2.

Results indicated that backtracking was common with the accuracy measure. Of the 53 BiTD participants whose samples included at least three obligatory contexts at Time 1, 42% had lower composite accuracy scores at Time 2. For the 10 BiLL children who met the criterion for calculating accuracy at Time 1, 50% demonstrated decreased accuracy rates at Time 2. Neither of these proportions significantly differed from chance ($ps > .27$), suggesting that, at the individual level, the accuracy measure did not reliably capture growth.

For the diversity and productivity measures, backtracking occurred less frequently. Of the 74 BiTD participants, 22% had lower scores on the tense marker total at Time 2, and the same percentage demonstrated backtracking on the TAP score. Of the 19 BiLL participants, only one child (5%) earned a lower tense marker total, and only two children (11%) earned lower TAP scores at Time 2. Each

**Figure 3.** TAP scores at Time 1 and Time 2 for both participant groups. BiTD = bilingual with typically developing language; BiLL = bilingual with low language skills; TAP score = tense and agreement productivity score.

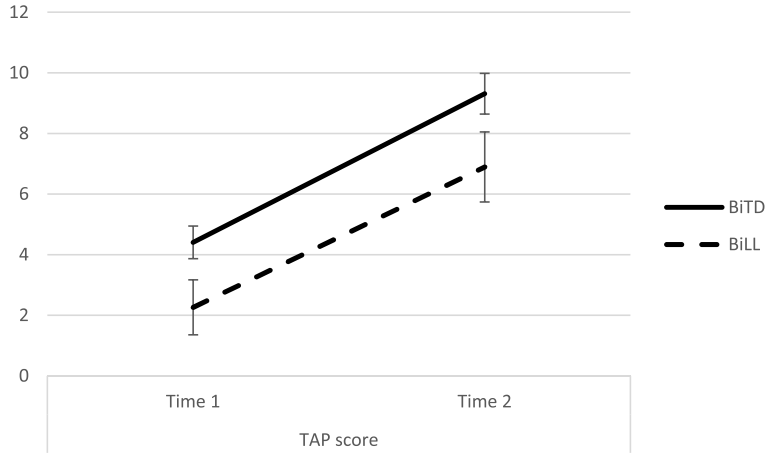
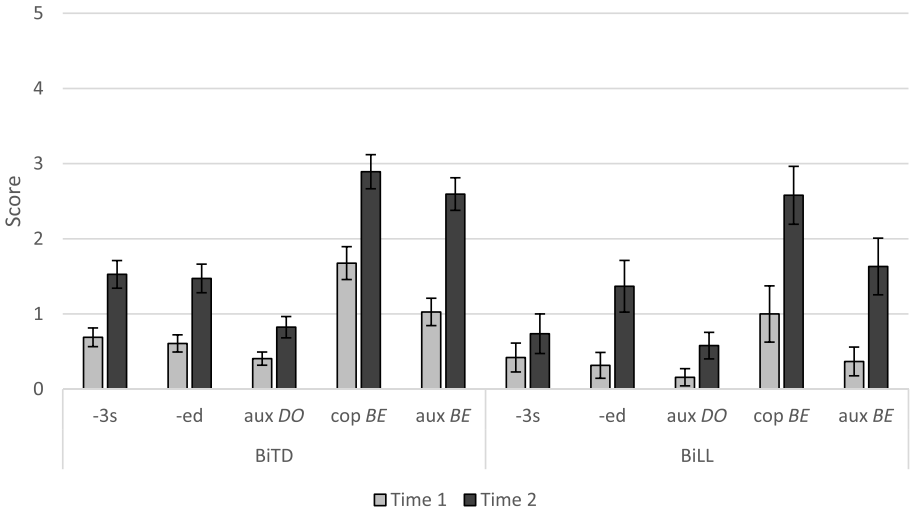


**Figure 4.** TAP score subscores for individual morphemes at Time 1 and Time 2 for each participant group. BiTD = bilingual with typically developing language; BiLL = bilingual with low language skills; TAP score = tense and agreement productivity score; -3s = third-person singular; -ed = past tense; cop *BE* = copula *BE*; aux *BE* = auxiliary *BE*; aux *DO* = auxiliary *DO*.




of these proportions was below chance ($p$s < .001), indicating limited backtracking. Indeed, fewer children demonstrated backtracking for these two measures than for accuracy ($p$s < .05, $V$s = .0083–.0592).[2]

## Discussion

The present work sought to identify appropriate measures of English T/A morpheme use in preschool-age Spanish–English developing bilinguals with varying language skills. We considered three morpheme measures that can be derived from spontaneous language samples. One measure, a composite score capturing T/A accuracy, is frequently used for assessing language in clinical and research settings for monolingual and bilingual children (Balason & Dollaghan, 2002; Bedore & Leonard, 1998; Gladfelter & Leonard, 2013; Gutiérrez-Clellen et al., 2008; Rice et al., 1998). However, evidence suggests that there are drawbacks to accuracy measures when working with children whose English T/A systems are emerging (Fitzgerald et al., 2012; Rispoli et al., 2009). Composite accuracy measures do not safeguard against repetitions (e.g., multiple instances of -3s and -ed with the same lexical verb) or potentially formulaic constructions (e.g., *he's, it's, there's*). As a result, this measure may overestimate abilities at early stages of T/A development. Furthermore, limited obligatory contexts may make for an unreliable measure at certain stages of development (Balason & Dollaghan, 2002). The second two measures, the tense marker total and TAP score, were designed to compensate for those weaknesses and have demonstrated clinical utility for monolingual children with emerging T/A systems (Gladfelter & Leonard, 2013; Guo & Eisenberg, 2014; Hadley & Holt, 2006; Hadley & Short, 2005; Hadley et al., 2014; Rispoli et al., 2009; Rispoli, Hadley, & Holt, 2012). This study tested whether these measures are appropriate for young developing bilinguals, as they, too, are likely to be in emerging stages of English T/A development.

**Table 4.** Number of obligatory contexts per morpheme for BiTD and BiLL participants at Time 1.

| Group | | -3s | -ed | cop *BE* | aux *BE* | aux *DO* |
|---|---|---|---|---|---|---|
| BiTD | *M* | 2.38 | 1.58 | 10.11 | 6.47 | 1.73 |
| | *SD* | 3.28 | 2.26 | 11.23 | 8.03 | 2.37 |
| | Range | 0–14 | 0–9 | 0–44 | 0–36 | 0–9 |
| | *Mdn* | 1 | 0 | 6 | 3 | 1 |
| BiLL | *M* | 2.05 | 1.42 | 6.63 | 2.53 | 0.47 |
| | *SD* | 3.78 | 3.10 | 8.88 | 5.09 | 0.91 |
| | Range | 0–11 | 0–11 | 0–29 | 0–20 | 0–3 |
| | *Mdn* | 0 | 0 | 3 | 0 | 0 |

*Note.* BiTD = bilingual with typically developing language; BiLL = bilingual with low language skills; -3s = third-person singular; -ed = past tense; cop *BE* = copula *BE*; aux *BE* = auxiliary *BE*; aux *DO* = auxiliary *DO*.

[2]An additional set of analyses considered individual stability for the three measures. Stability is characterized by consistent relative rankings over time, such that children with strong performance relative to peers at one point demonstrate a similarly high ranking at a second test point. This pattern is identified by significant correlations between performance on the same measure at multiple time points (Bornstein, Brown, & Slater, 1996). For BiTD participants, the diversity and productivity measures at Time 1 and Time 2 were found to be moderately positively correlated: tense marker total, $r$ = .454, $p$ < .001; TAP score: $r$ = .403, $p$ < .001. Results were comparable for BiLL participants: tense marker total, $r$ = .543, $p$ = .016; TAP score: $r$ = .518, $p$ = .023. Conversely, accuracy scores at pre- and posttesting were significantly associated only for BiTD participants: $r$ = .361, $p$ = .008.

Identifying meaningful measures of morpheme use for bilingual children has important clinical implications, as T/A morpheme marking is an area of particular weakness for English-speaking monolingual and bilingual children with language impairment (Bedore & Leonard, 1998; Gutiérrez-Clellen et al., 2008; Rice & Wexler, 1996).

### *T/A Morpheme Measures and Broad Language Sample Measures*

In order to determine whether the three measures of morpheme use were appropriate for preschool-age developing bilinguals, we first established whether they were associated with culturally and developmentally sensitive measures of language development. Both the tense marker total and TAP score were positively associated with MLUw and NDW for both groups at both testing points. These relationships were generally strong, indicating that the tense marker total and TAP score convey information that is relevant to language development (Ebert & Pham, 2017).

The accuracy composite scores were not found to be consistently correlated with MLUw or NDW. Each may be considered a broad measure, as accuracy takes into account all utterances with obligatory contexts for the target morphemes and MLUw and NDW are calculated with reference to all complete and intelligible utterances. And yet, it was the streamlined diversity and productivity measures that correlated with MLUw and NDW. These results support the use of the tense marker total and TAP score for children with emerging morphological skills. Conversely, the composite accuracy measure is associated with morpheme mastery and may thus be better suited for capturing later stages of morphosyntactic development (Fitzgerald et al., 2012; Rispoli et al., 2009).

The three morpheme measures were also significantly related to one another. This is to be expected, as the three measures all seek to capture the same expressive language skill. Notably, however, these relationships were stronger at the end of the school year for both groups. The tense marker total and TAP score appear to be appropriate at both testing points, as evidenced by their consistent correlations with MLUw and NDW. The increased correspondence between the tense marker total and TAP score with accuracy at Time 2 may indicate that accuracy has become an increasingly reliable measure as the young bilinguals develop their English T/A system.

Overall, results point to the relevance of the tense marker total and the TAP score for measuring morpheme use in developing bilinguals acquiring English. Support for a composite measure of T/A accuracy was less consistent—particularly at earlier stages of T/A morpheme acquisition (in the case of this study, at Time 1). These findings parallel those found for younger monolingual children, who, like the participants in this study, are acquiring the English T/A system (Fitzgerald et al., 2012; Hadley & Short, 2005; Rispoli et al., 2009).

### *T/A Morpheme Use Across Groups*

This study included two groups of preschool-age Spanish–English developing bilinguals: those with typical development and those with low language skills. The groups were comparable in language exposure, age, and maternal education. As expected, the typically developing group outperformed their peers with reported low language skills on numerous language sample measures, including the developmentally sensitive MLUw and NDW. The diversity and productivity measures—but not the composite accuracy measure—reflected these group differences. BiLL participants used fewer surface forms, as evidenced by lower tense marker totals, and they used the target T/A morphemes less contrastively, as evidenced by lower TAP scores. These observed differences indicate that the tense marker total and TAP score may produce information that is relevant to language assessment in preschool-age children acquiring English.

Conversely, accuracy rates were comparable across children in the typically developing and low language groups. This finding diverged from prior research that measured English T/A accuracy in bilingual children. Gutiérrez-Clellen et al. (2008) found that accuracy did differentiate between typical and atypical language development in young Spanish–English bilinguals. Important differences in participant characteristics may explain this discrepancy. Participants in the Gutiérrez-Clellen et al. study included preschoolers, kindergarteners, and first graders with relatively strong English skills (i.e., received minimum parent ratings of 3 for English use on a scale of 0–4, with "substantial difficulty" speaking Spanish, Gutiérrez-Clellen et al., 2008, p. 8). Participants in this study were generally younger and had greater exposure to Spanish. Overall, the participants in Gutiérrez-Clellen et al. likely had more advanced English language skills, making accuracy a more reliable measure. In fact, their typically developing bilingual participants were 84% accurate on English T/A morphemes, performing well above our groups. The present pattern of results for accuracy may thus be related to our participants' relatively early stage of English morphological development. Variable morpheme marking in bilinguals (e.g., Paradis et al., 2008) and unreliable measures of accuracy due to limited obligatory contexts (e.g., Balason & Dollaghan, 2002) are also relevant. Language assessment measures must be appropriate for a child's background and stage of development. The present results highlight the appropriateness of the tense marker total and TAP score measures for preschool-age bilinguals who are learning the English T/A system.

### *T/A Morpheme Use Over Time*

Both tense marker totals and TAP scores increased from the beginning to the end of the school year for both participant groups. Furthermore, significant increases in the productivity of -3s, -ed, cop *BE,* aux *BE,* and aux *DO* were captured for both BiTD and BiLL participants. This is consistent with significant improvements in both MLUw

and NDW from Time 1 to Time 2. The ability to monitor T/A acquisition in terms of both overarching measures (i.e., tense marker totals and TAP scores) and specific morphemes provides versatility that is valuable in clinical contexts.

Conversely, group composite accuracy rates did not improve across the two testing points for BiTD or BiLL participants. We argue that these results are likely not indicative of the children's true abilities and that the children's current morphological development must be considered. The limited number of obligatory contexts in these samples is consistent with our understanding that these Spanish–English bilinguals have emerging morphological skills in English. Likewise, they have yet to meet standard criteria for morpheme mastery (i.e., 80%–95% accuracy; e.g., Brown, 1973). Potentially, at this stage, measuring morpheme use with accuracy—as opposed to the more appropriate diversity and productivity measures—underestimates these children's gains. Under unfortunate circumstances, this could contribute to known problems with overidentifying language disorders in bilingual children (Artiles, Rueda, Salazar, & Higareda, 2005; Bedore & Peña, 2008).

### *Individual Patterns of Growth in T/A Morpheme Use*

For both the tense marker total and the TAP score, individual-level findings mirrored group patterns: Most BiTD and BiLL children earned higher tense marker totals and TAP scores by the end of the year than at Time 1, with relatively little evidence for backtracking. For the composite accuracy measure, however, backtracking was common for both groups: Time 2 accuracy rates were lower for 50% of BiLL children and 42% of BiTD children who met the criterion for three obligatory T/A contexts at Time 1. Given the profile of growth demonstrated by the tense marker total, TAP score, MLUw, and NDW—and considering comparable backtracking in accuracy demonstrated by monolinguals acquiring the English T/A system (Rispoli et al., 2009)—the backtracking in accuracy rates observed here is likely due to the relatively poor fit of this measure for this population at this time.

### *Case Examples and Clinical Implications*

Adding to the findings discussed above, the value of the tense marker total and TAP score is illustrated with two case examples. Isabella and Ruby were two participants matched on age, Spanish exposure, and MLUw in English at Time 1. Critically, Ruby's parent report indicated concern with her language development, but Isabella's did not (see Appendix B).

Parent concern, or lack thereof, was reflected in the children's tense marker totals and TAP scores. At Time 1, Isabella used the T/A morphemes contrastively nine times, whereas Ruby did not produce any. By Time 2, Isabella demonstrated considerable productivity, earning a TAP score of 16. Ruby also made improvements by the second testing point—but still used fewer surface forms in fewer contexts than her typically developing peer did at Time 1.

The diversity and productivity measures captured growth for both children while complementing other language sample measures. Based on MLUw, it might appear that the children had comparable language skills at Time 1. Alternately, referring to composite accuracy scores might lead to concern regarding Isabella's language development. Isabella's Time 2 composite accuracy rate (57.83%) demonstrated backtracking from her Time 1 accuracy rate (75.61%) and was lower than her peer's Time 2 accuracy rate (72.41%). However, these observations are inconsistent with Isabella's relatively high TAP score at Time 2, her notable improvement in productivity from Time 1, and her parents' lack of concern regarding language development. A closer look at the two language samples continued to reveal differences between Isabella and Ruby. Isabella's Time 2 transcript included 83 obligatory contexts, whereas Ruby's relatively high Time 2 accuracy rate was based on only 29 obligatory contexts and was paired with a low tense marker total and TAP score. In this case, calculating a percentage- or proportion-based measure like composite accuracy masked absolute counts and could be misleading if taken on its own, particularly when obligatory contexts are limited. Designed to capture emerging morphological skills, the tense marker total and TAP score help characterize Isabella's and Ruby's productive language.

Valuable information is clearly presented in the scoring tables for the diversity and productivity measures (see Appendix B). In reviewing the tense marker total table, one quickly sees which surface forms are missing from Isabella's language sample. Similarly, a review of her TAP score table allows us to ascertain the depth of her knowledge of each morpheme. This criterion-based approach may be useful for identifying areas of strength and weakness (Stockman, 1996) and for tracking progress in specific areas. Comparing Time 1 and Time 2 scoring tables makes clear which new forms have been demonstrated and whether gains in productivity had been made. In contrast, traditional measures do not provide this type of detailed information. For example, Isabella's composite accuracy measures do not indicate which morphemes or surface forms were used or with what degree of success. The tense marker total and TAP score allow clinicians to quickly access meaningful and specific information about a child's morphological development that may be relevant to assessment, treatment, and progress monitoring.

The diversity and productivity measures complement one another. For example, by noting that Isabella's relatively high Time 1 TAP score (9) is paired with a lower tense marker total (3), a scorer is able to deduce that she used few surface forms, but she used them contrastively and in a variety of contexts. Although Isabella's tense marker total reveals that the only surface form of cop *BE* she produced was *is,* her TAP score indicates that she used that form highly productively. Indeed, her language

sample included five different subjects: "*There is a frog.*" "*Where's up?*" "*What story is this?*" "*Is it off?*" and "*That is cars.*" By Time 2, we see relatively high scores for both novel measures, suggesting that we should see contrastive uses across multiple morphemes. Reviewing a child's tense marker total and TAP score provides a clinician with concrete information regarding the child's morpheme use over and above what may be captured with other measures.

This valuable information is acquired through streamlined scoring procedures, making the tense marker total and TAP score a practical means of assessing language skills. These measures focus only on productive morpheme uses (i.e., correct productions and overregularizations) of five morphemes and impose a clear ceiling rule. Furthermore, a number of forms are exempt from scoring, including verbs produced with no subject, repeated subject/surface form combinations, and, for copula and auxiliary verbs, contractions onto pronouns (see Hadley & Short, 2005). Therefore, the number of utterances that a clinician must review is substantially reduced relative to other measures. And yet, this focused approach does not appear to detract from the measures' meaningfulness: The tense marker total and TAP score provided substantial and relevant information about T/A morpheme use in bilingual children.

### *Future Research and Limitations*

We encourage other researchers and clinicians to investigate the tense marker total and TAP score with other young bilinguals. Of note, the participants in this study were preschool-age Spanish–English developing bilinguals from low socioeconomic backgrounds in Southern California. Other bilingual groups, including children being raised in additive bilingual communities or those from higher socioeconomic backgrounds, may demonstrate differing tense marker totals and TAP scores. Relatedly, it would be important to consider the diversity and productivity of T/A forms in developing bilinguals as a function of factors relevant to bilingualism, including relative exposure to the two languages. The present sample was characterized by greater exposure to Spanish (e.g., in each group, the modal reported Spanish exposure at home was 100), limiting our ability to investigate the impact of this important variable. The specifics of our sampled population of English learners notwithstanding, the tense marker total and TAP score were designed to capture initial stages of morpheme emergence and productivity; as such, we anticipate that these measures would similarly track morpheme development in children acquiring English under different conditions, though absolute scores may differ. In addition, the lack of ceiling effects suggests that these measures may be explored in older bilingual children. Much could yet be gained from work, extending use of these measures to bilingual children with different profiles.

In the present study, we identified group differences across typically developing and low language groups, which is indicative of diagnostic potential and is consistent with related research in monolinguals with typical and atypical language development (e.g., Gladfelter & Leonard, 2013; Guo & Eisenberg, 2014). We did not compare children with and without confirmed language impairment. However, parent concern—the criterion used to determine group status in this study—is a valuable indicator of language status (Gutiérrez-Clellen & Kreiter, 2003). Nevertheless, future research should explore the ability of these measures to differentiate bilingual groups with and without confirmed language impairment. Similarly, comparisons with other culturally sensitive measures, such as grammaticality (Bedore et al., 2010; Ebert & Pham, 2017), may serve to bolster the relevance of these measures.

This study would also be strengthened with a larger sample of children with low language skills, particularly considering that the composite accuracy measure could only be calculated for subsets of each group. However, a number of findings indicate that the differential findings for the accuracy measure and the diversity and productivity measures are not a result of reductions in sample size for accuracy analyses. As one example, at Time 2, when all BiLL participants were eligible for accuracy analyses, significant correlations emerged between broad language sample measures and the diversity and productivity measures, but not the composite accuracy measure. In addition, group-level effects (e.g., growth over time) persisted when analyses for the tense marker total and the TAP score were repeated using the reduced participant groups imposed by the accuracy measure. That accuracy could not be reliably calculated for a number of our participants may be interpreted as an indicator that this measure is less appropriate for bilingual children at this stage of English T/A development.

Yet another exciting direction for future research is to consider more closely the use of each individual T/A morpheme. In work completed by Hadley, Rispoli, and colleagues, a robust onset pattern has emerged for young monolinguals: cop *BE* increases in productivity most rapidly, followed by -3s, -ed, and aux *DO,* with aux *BE* demonstrating growth in productivity most slowly (e.g., Rispoli et al., 2012). In the present data, we see relatively high productivity of aux *BE,* a trajectory that diverges from the monolingual data but is consistent with previous work that demonstrated "precocious" use of this morpheme in bilingual children in terms of accuracy (Paradis & Blom, 2016). These findings point to potential qualitative differences in the development of the English T/A system in bilingual children relative to monolinguals, though there may be broad similarities in how the two groups develop morphosyntactic skills gradually (e.g., Rice, 2010). Identifying areas of similarity and contrast across bilingual and monolingual trajectories in acquiring the English T/A system is important for establishing appropriate reference points for clinical settings. Present results also revealed that, in addition to lower tense marker totals and TAP scores, the BiLL group showed lower productivity for -3s and aux *BE* relative to their typically developing peers. Future research can investigate whether these morphemes are particularly sensitive to varying language skills in young bilinguals.

### *Summary*

Results indicate that English morphological development in Spanish–English developing bilinguals can be meaningfully measured with reference to morpheme diversity and productivity. Specifically, the tense marker total and TAP score (Hadley & Short, 2005) converged with traditional language measures, were sensitive to varying language skills, and demonstrated growth over time. Several weaknesses of accuracy measures were identified, suggesting that the diversity and productivity measures may be an important complement to language assessment in bilingual children, particularly for children whose English language skills are emerging. When used appropriately, English language measures may have a valuable place in bilingual language assessment (Bedore et al., 2018; Gillam et al., 2013; Gutiérrez-Clellen et al., 2008), particularly when combined with a parent report and assessment of the native language (e.g., Bedore & Peña, 2008; Gillam, Peña, & Miller, 1999).

## Acknowledgments

The first and third authors were supported by the Lipinsky Fellowship at San Diego State University. In addition, funding for the project was provided by a National Institute on Deafness and Other Communication Disorders RO3 grant (Grant DC012141) and a Price Family Philanthropies research grant awarded to the final author. We would like to thank the participants, their families, and the teachers at Rosa Parks Elementary School. We would also like to thank the members of the Child Language Development, Disorders and Disparities Lab for research assistance and helpful discussions of this work. We extend sincere appreciation to Pamela Hadley for graciously sharing training materials for the tense marker total and the TAP score and for her thoughtful feedback on the current research.

## References

**Armon-Lotem, S.** (2014). Between L2 and SLI: Inflections and prepositions in the Hebrew of bilingual children with TLD and monolingual children with SLI. *Journal of Child Language, 41*(1), 3–33.

**Artiles, A. J., Rueda, R., Salazar, J. J., & Higareda, I.** (2005). Within-group diversity in minority disproportionate representation: English language learners in urban school districts. *Exceptional Children, 71*(3), 283–300.

**Balason, D. V., & Dollaghan, C. A.** (2002). Grammatical morpheme production in 4-year-old children. *Journal of Speech, Language, and Hearing Research, 45*(5), 961–969.

**Bedore, L. M., & Leonard, L. B.** (1998). Specific language impairment and grammatical morphology: A discriminant function analysis. *Journal of Speech, Language, and Hearing Research, 41*(5), 1185–1192.

**Bedore, L. M., & Peña, E. D.** (2008). Assessment of bilingual children for identification of language impairment: Current findings and implications for practice. *International Journal of Bilingual Education and Bilingualism, 11*(1), 1–29.

**Bedore, L. M., Peña, E. D., Anaya, J. B., Nieto, R., Lugo-Neris, M. J., & Baron, A.** (2018). Understanding disorder within variation: Production of English grammatical forms by English language learners. *Language, Speech, and Hearing Services in Schools, 49,* 277–291.

**Bedore, L. M., Peña, E. D., Gillam, R. B., & Ho, T. H.** (2010). Language sample measures and language ability in Spanish–English bilingual kindergarteners. *Journal of Communication Disorders, 43*(6), 498–510.

**Blom, E., Paradis, J., & Duncan, T. S.** (2012). Effects of input properties, vocabulary size, and L1 on the development of third person singular *-s* in child L2 English. *Language Learning, 62*(3), 965–994.

**Bornstein, M. H., Brown, E., & Slater, A.** (1996). Patterns of stability and continuity in attention across early infancy. *Journal of Reproductive and Infant Psychology, 14*(3), 195–206.

**Brown, R.** (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.

**Caesar, L. G., & Kohler, P. D.** (2007). The state of school-based bilingual assessment: Actual practice versus recommended guidelines. *Language, Speech, and Hearing Services in Schools, 38*(3), 190–200.

**Ebert, K. D., & Pham, G.** (2017). Synthesizing information from language samples and standardized tests in school-age bilingual assessment. *Language, Speech, and Hearing Services in Schools, 48*(1), 42–55.

**Eisenberg, S. L., Guo, L. Y., & Germezia, M.** (2012). How grammatical are 3-year-olds? *Language, Speech, and Hearing Services in Schools, 43*(1), 36–52.

**Evans, J. D.** (1996). *Straightforward statistics for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole Publishing.

**Fitzgerald, C., Rispoli, M., Hadley, P., & McKenna, M.** (2012). Productivity scoring as a metric of early finiteness marking. *Poster presentation at the Symposium on Research in Child Language Disorders.*

**Gillam, R. B., Peña, E. D., Bedore, L. M., Bohman, T. M., & Mendez-Perez, A.** (2013). Identification of specific language impairment in bilingual children: I. Assessment in English. *Journal of Speech, Language, and Hearing Research, 56*(6), 1813–1823.

**Gillam, R. B., Peña, E. D., & Miller, L.** (1999). Dynamic assessment of narrative and expository discourse. *Topics in Language Disorders, 20*(1), 33–47.

**Gladfelter, A., & Leonard, L. B.** (2013). Alternative tense and agreement morpheme measures for assessing grammatical deficits during the preschool period. *Journal of Speech, Language, and Hearing Research, 56*(2), 542–552.

**Golberg, H., Paradis, J., & Crago, M.** (2008). Lexical acquisition over time in minority first language children learning English as a second language. *Applied Psycholinguistics, 29*(1), 41–65.

**Guo, L. Y., & Eisenberg, S.** (2014). The diagnostic accuracy of two tense measures for identifying 3-year-olds with language impairment. *American Journal of Speech-Language Pathology, 23*(2), 203–212.

**Guo, L. Y., Spencer, L. J., & Tomblin, J. B.** (2013). Acquisition of tense marking in English-speaking children with cochlear implants: A longitudinal study. *Journal of Deaf Studies and Deaf Education, 18*(2), 187–205.

**Gutiérrez-Clellen, V. F., & Kreiter, J.** (2003). Understanding child bilingual acquisition using parent and teacher reports. *Applied Psycholinguistics, 24*(2), 267–288.

**Gutiérrez-Clellen, V. F., Restrepo, M. A., Bedore, L., Peña, E., & Anderson, R.** (2000). Language sample analysis in Spanish-speaking children: Methodological considerations. *Language, Speech, and Hearing Services in Schools, 31*(1), 88–98.

**Gutiérrez-Clellen, V. F., Simon-Cereijido, G., & Wagner, C.** (2008). Bilingual children with language impairment: A comparison with monolinguals and second language learners. *Applied Psycholinguistics, 29*(1), 3–19.

**Hadley, P. A., & Holt, J. K.** (2006). Individual differences in the onset of tense marking: A growth-curve analysis. *Journal of Speech, Language, and Hearing Research, 49*(5), 984–1000.

**Hadley, P. A., Rispoli, M., Holt, J. K., Fitzgerald, C., & Bahnsen, A.** (2014). Growth of finiteness in the third year of life: Replication and predictive validity. *Journal of Speech, Language, and Hearing Research, 57*(3), 887–900.

**Hadley, P. A., & Short, H.** (2005). The onset of tense marking in children at risk for specific language impairment. *Journal of Speech, Language, and Hearing Research, 48*(6), 1344–1362.

**Heilmann, J. J.** (2010). Myths and realities of language sample analysis. *SIG 1 Perspectives on Language Learning and Education, 17*(1), 4–8.

**Heilmann, J. J., Miller, J. F., & Nockerts, A.** (2010). Using language sample databases. *Language, Speech, and Hearing Services in Schools, 41*(1), 84–95.

**Hewitt, L. E., Hammer, C. S., Yont, K. M., & Tomblin, J. B.** (2005). Language sampling for kindergarten children with and without SLI: Mean length of utterance, IPSYN, and NDW. *Journal of Communication Disorders, 38*(3), 197–213.

**Leonard, L. B.** (2014). *Children with specific language impairment*. Cambridge, MA: MIT Press.

**Miller, J., & Iglesias, A.** (2012). *SALT: Systematic Analysis of Language Transcripts. Software for the analysis of oral language*. Middleton, WI: SALT Software.

**Nicoladis, E., Song, J., & Marentette, P.** (2012). Do young bilinguals acquire past tense morphology like monolinguals, only later? Evidence from French–English and Chinese–English bilinguals. *Applied Psycholinguistics, 33*(3), 457–479.

**Oetting, J. B., Newkirk, B. L., Hartfield, L. R., Wynn, C. G., Pruitt, S. L., & Garrity, A. W.** (2010). Index of Productive Syntax for children who speak African American English. *Language, Speech, and Hearing Services in Schools, 41*(3), 328–339.

**Paradis, J.** (2005). Grammatical morphology in children learning English as a second language: Implications of similarities with specific language impairment. *Language, Speech, and Hearing Services in Schools, 36*(3), 172–187.

**Paradis, J., & Blom, E.** (2016). Do early successive bilinguals show the English L2 pattern of precocious BE acquisition? *Bilingualism: Language and Cognition, 19*(3), 630–635.

**Paradis, J., & Crago, M.** (2000). Tense and temporality: A comparison between children learning a second language and children with SLI. *Journal of Speech, Language, and Hearing Research, 43*(4), 834–847.

**Paradis, J., & Kirova, A.** (2014). English second-language learners in preschool: Profile effects in their English abilities and the role of home language environment. *International Journal of Behavioral Development, 38*(4), 342–349.

**Paradis, J., Nicoladis, E., Crago, M., & Genesee, F.** (2011). Bilingual children's acquisition of the past tense: A usage-based approach. *Journal of Child Language, 38*(3), 554–578.

**Paradis, J., Rice, M. L., Crago, M., & Marquis, J.** (2008). The acquisition of tense in English: Distinguishing child second language from first language and specific language impairment. *Applied Psycholinguistics, 29*(4), 689–722.

**Pearson, B. Z., Fernandez, S. C., Lewedeg, V., & Oller, D. K.** (1997). The relation of input factors to lexical learning by bilingual infants. *Applied Psycholinguistics, 18*(1), 41–58.

**Pruitt, S., & Oetting, J.** (2009). Past tense marking by African American English–speaking children reared in poverty. *Journal of Speech, Language, and Hearing Research, 52*(1), 2–15.

**Restrepo, M. A.** (1998). Identifiers of predominantly Spanish-speaking children with language impairment. *Journal of Speech, Language, and Hearing Research, 41*(6), 1398–1411.

**Rice, M. L.** (2010). Evaluating maturational parallels in second language children and children with specific language impairment. *Applied Psycholinguistics, 31*(2), 320–327.

**Rice, M. L., & Wexler, K.** (1996). Toward tense as a clinical marker of specific language impairment in English-speaking children. *Journal of Speech and Hearing Research, 39*(6), 1239–1257.

**Rice, M. L., & Wexler, K.** (2001). *Rice/Wexler Test of Early Grammatical Impairment*. San Antonio, TX: Psychological Corporation.

**Rice, M. L., Wexler, K., & Cleave, P. L.** (1995). Specific language impairment as a period of extended optional infinitive. *Journal of Speech and Hearing Research, 38*(4), 850–863.

**Rice, M. L., Wexler, K., & Hershberger, S.** (1998). Tense over time: The longitudinal course of tense acquisition in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 41*(6), 1412–1431.

**Rispoli, M., & Hadley, P.** (2011). Toward a theory of gradual morphosyntactic learning. In I. Aron & E. V. Clark (Eds.), *Experience, variation, and generalization: Learning a first language* (pp. 15–34). Amsterdam, the Netherlands: John Benjamins.

**Rispoli, M., Hadley, P. A., & Holt, J. K.** (2009). The growth of tense productivity. *Journal of Speech, Language, and Hearing Research, 52*(4), 930–944.

**Rispoli, M., Hadley, P. A., & Holt, J. K.** (2012). Sequence and system in the acquisition of tense and agreement. *Journal of Speech, Language, and Hearing Research, 55*(4), 1007–1021.

**Roid, G. H., & Miller, L.** (1997). *Leiter International Test of Intelligence–Revised.* Chicago, IL: Stoelting.

**Rojas, R., & Iglesias, A.** (2009). Making a case for language sampling: Assessment and intervention with (Spanish-English) second language learners. *The ASHA Leader, 14*(3), 10–13.

**Simon-Cereijido, G., & Gutiérrez-Clellen, V. F.** (2009). A cross-linguistic and bilingual evaluation of the interdependence between lexical and grammatical domains. *Applied Psycholinguistics, 30*(2), 315–337.

**Stockman, I. J.** (1996). The promises and pitfalls of language sample analysis as an assessment tool for linguistic minority children. *Language, Speech, and Hearing Services in Schools, 27*(4), 355–366.

## Appendix A

Abbreviated Language Sample

| Utterance no. | Utterance | Morpheme | Correct/error |
|---|---|---|---|
| 1 | This is a girl. | cop *BE* | Correct |
| 2 | And now we're gonna sit down | aux *BE* | Correct |
| 3 | The mama's going up at the car. | aux *BE* | Correct |
| 4 | The baby going under. | aux *BE* | Omission |
| 5 | Daddy's driving down. | aux *BE* | Correct |
| 6 | Daddy's driving! | aux *BE* | Correct |
| 7 | That looks like french fries. | -3s | Correct |
| 8 | The boy eat french fries. | -3s | Omission |
| 9 | They is hungry. | cop *BE* | Agreement error |
| 10 | Yesterday he cry. | -ed | Omission |
| 11 | But I am tired now. | cop *BE* | Correct |
| 12 | He doesn't know. | aux *DO* | Correct |
| 13 | I played basketball. | -ed | Correct |
| 14 | He played basketball | -ed | Correct |
| 15 | It breaked! | -ed | Overregularization |

Tense Marker Total: Diversity of the Tense and Agreement System

| | | aux *DO* | | | cop *BE* | | | | | aux *BE* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **-3s** | **-ed** | **do** | **does** | **did** | **is** | **am** | **are** | **was** | **were** | **is** | **am** | **are** | **was** | **were** | **Total** |
| Utterance 7 | Utterance 13 | | Utterance 12 | | Utterance 1 | Utterance 11 | | | | Utterance 3 | | | | | 6 |

Tense and Agreement Productivity Score: Productivity of the Tense and Agreement System

| | -3s | -ed | aux *DO* | cop *BE* | aux *BE* | Total |
|---|---|---|---|---|---|---|
| Instance 1 | Utterance 7 | Utterance 13 | Utterance 12 | Utterance 1 | Utterance 3 | 8 |
| Instance 2 | | Utterance 15 | | Utterance 11 | Utterance 5 | |
| Instance 3 | | | | | | |
| Instance 4 | | | | | | |
| Instance 5 | | | | | | |

*Note.* -3s = third-person singular; -ed = past tense; cop *BE* = copula *BE*; aux *BE* = auxiliary *BE*; aux *DO* = auxiliary *DO*.

## Appendix B

Case Examples: Isabella and Ruby

| Participant | Age (years; months) | Language exposure at home | MLUw | | Tense marker total | | TAP score | | Composite accuracy measure | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Time 1 | Time 2 | Time 1 | Time 2 | Time 1 | Time 2 | Time 1 | Time 2 |
| Isabella (BiTD) | 3;10 | 100% Spanish | 2.36 | 3.68 | 3 | 7 | 9 | 16 | 75.61% | 57.83% |
| Ruby (BiLL) | 3;10 | 100% Spanish | 2.19 | 3.14 | 0 | 2 | 0 | 4 | 42.86% | 72.41% |

*Note.* MLUw = mean length of utterance in words; TAP score = tense and agreement productivity score; BiTD = developing bilingual children with typical language development; BiLL = developing bilingual children with low language skills.

Isabella's Time 1 Tense Marker Total

| | | aux *DO* | | | cop *BE* | | | | | aux *BE* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -3s | -ed | do | does | did | is | am | are | was | were | is | am | are | was | were | Total |
| ✓ | | | | | ✓ | | | | | ✓ | | | | | 3 |

Isabella's Time 1 Tense and Agreement Productivity Score

| | -3s | -ed | aux *DO* | cop *BE* | aux *BE* | Total |
|---|---|---|---|---|---|---|
| Instance 1 | ✓ | | | ✓ | ✓ | 9 |
| Instance 2 | ✓ | | | ✓ | | |
| Instance 3 | ✓ | | | ✓ | | |
| Instance 4 | | | | ✓ | | |
| Instance 5 | | | | ✓ | | |

Isabella's Time 2 Tense Marker Total

| | | aux *DO* | | | cop *BE* | | | | | aux *BE* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -3s | -ed | do | does | did | is | am | are | was | were | is | am | are | was | were | Total |
| ✓ | ✓ | | ✓ | ✓ | ✓ | | | ✓ | | ✓ | | | | | 7 |

Isabella's Time 2 Tense and Agreement Productivity Score

| | -3s | -ed | aux *DO* | cop *BE* | aux *BE* | Total |
|---|---|---|---|---|---|---|
| Instance 1 | ✓ | ✓ | ✓ | ✓ | ✓ | 16 |
| Instance 2 | ✓ | | ✓ | ✓ | ✓ | |
| Instance 3 | | | ✓ | ✓ | ✓ | |
| Instance 4 | | | ✓ | ✓ | ✓ | |
| Instance 5 | | | | ✓ | | |

*Note.* -3s = third-person singular; -ed = past tense; cop *BE* = copula *BE*; aux *BE* = auxiliary *BE*; aux *DO* = auxiliary *DO*.