Early Childhood Research Quarterly 54 (2021) 271–285

Contents lists available at ScienceDirect

Early Childhood Research Quarterly

Evidence of support for dual language learners in a study of bilingual staffing patterns using the Classroom Assessment of Supports for Emergent Bilingual Acquisition (CASEBA)

Alexandra Figueras-Daniel[a,*], Zijia Li[b]

[a] Straus Center for Young Children & Families, Bank Street College of Education
[b] National Institute for Early Education Research, Graduate School of Education, Rutgers, The State University of New Jersey

ARTICLE INFO

Article history:
Received 2 April 2019
Received in revised form 20 July 2020
Accepted 29 September 2020

Keywords:
Dual language learners
Classroom quality
Preschool
Lead and assistant teacher language

ABSTRACT

The mounting numbers of young Hispanic children in the United States (now about 25% of those under five) require unique considerations in efforts to meet their particular needs for preschool education. Among the most agreed upon practices to address these needs are the provision of strategic and intentional interactions in English with use of children's home languages (HLs) for instruction. However, challenges include the capacity among the workforce to deliver instruction in two languages and whether assistant teachers may be relied on to provide HL instruction. Further, to assess the quality of interactions deemed best for dual language learners (DLLs) use of an observation tool explicitly designed to understand these contexts for DLLs is warranted. The purpose of this study was to examine the relative effectiveness of different teacher and assistant teacher bilingualism combinations on teaching practices as assessed using the Classroom Assessment of Supports for Emergent Bilingual Acquisition (CASEBA; Freedson, Figueras, & Frede, 2009), a tool specifically designed for measuring language supports for preschool DLLs. Results indicate that classroom quality scores relative to staff language configurations differed with Spanish-speaking lead teacher pairs, earning significantly higher scores than the other staff language configuration groups. Associations between the CASEBA and children's English and Spanish receptive vocabulary scores were also examined, revealing a relationship between assistant teacher's home language use and English receptive vocabulary scores. These findings present a springboard for policy conversations regarding the supply and demand of bilingual teachers and teacher assistants, pre-service training and professional development, and the ways in which a specific classroom observation tool can inform all of these areas.

1. Introduction

The large population of young Hispanic dual language learners (DLLs) has stimulated academic interest in understanding critical elements of teaching and learning toward minimizing achievement opportunity gaps in preschool. While research has shown the benefits of preschool attendance for DLLs (Buysse, Peisner-Feinberg, Páez, Hammer, & Knowles, 2014; Gormley, 2008; Weiland & Yoshikawa, 2013;), it is also known that program quality varies for minority children (Valentino, 2018), and that specific practices that maximize outcomes for DLLs differ from those of monolingual children (Castro, Espinosa, & Paez, 2011). Specifically, research has shown that the most responsive environments by design should include use of intentional supports for developing English as well as, when possible, maintaining Spanish (Castro, Páez, Dickinson, & Frede, 2011).

The ability to provide home language (HL) instruction, however, presents other challenges. Given the demographics and preparation of the workforce, the presence of a speaker of a home language is typically not the lead teacher, but rather the assistant teacher (Whitebook, McLean, Austin, & Edwards, 2018). Even when research has documented the extent to which a HL is used (in most cases Spanish), they have found that teachers use it for managerial purposes as opposed to instruction (Jacoby & Lesaux, 2014). To begin to understand what practices matter and how to improve quality for DLLs, it is critical to have measures that can assist with effectively identifying those practices. Recent reviews of the literature, however, have cited the lack of widely available, published measures to assess quality of early education settings that include

[*] Corresponding author.
E-mail address: afiguerasdaniel@bankstreet.edu (A. Figueras-Daniel).

https://doi.org/10.1016/j.ecresq.2020.09.011
0885-2006/Published by Elsevier Inc.

language interactions between children and teachers and linguistic diversity (Castro et al., 2017; Shivers & Sanders, 2011).

To examine this question, the current study used a new, domain-specific tool, the Classroom Assessment for Support of Bilingual Emergent Acquisition (CASEBA; Freedson et al., 2009) to assess linguistically and culturally sensitive practices in preschool classrooms. More specifically, these practices included the use of a HL for promoting positive classroom experiences as well as for supporting English language acquisition. Further, given what is known about the challenges of delivering instruction in a HL, and that such instruction is often expected of assistant teachers (Whitebook et al., 2018), more data is needed to understand staffing arrangements and teacher bilingualism. To date, there has been little research on the quality of instruction between lead and assistant teachers based on their language dominance and its impact on children's language development. The findings of this study contribute to policy conversations not only regarding credentialing and certification, but also the potential importance of including assistant teachers in professional development activities such as coaching.

To begin, the CASEBA will be described and used to frame the study, along with a literature review that identifies research-based practices relative to the development of HLs in addition to the acquisition of English in preschool. Literature on what aspects of process quality matter for DLLs and how these depend on the workforce for implementation will also be reviewed. Further, to substantiate the need to explore staff language configurations, data on who is currently serving as lead teachers in preschool classrooms and what the pipeline of incoming teachers is like was included to underscore the importance of policy-level decisions and how more focused data collection tools can be helpful.

1.1. The Classroom Assessment of Supports for Emergent Bilingual Acquisition (CASEBA)

The CASEBA is a measure of both structural and process quality elements for conducting research in early childhood classrooms. Currently, the CASEBA is the only tool that includes items specifically designed to allow researchers to analyze teaching strategies that support both the HL and English language acquisition in early education classrooms for DLLs (Howes, Downer, & Pianta, 2011). In particular, the tool allows us to capture very specific elements of practice with each individual item based on research in order to potentially understand the supports that teachers uniquely provide to DLLs. By allowing us to pay attention to these supports, CASEBA analyses individual teacher efforts to establish an environment that embraces both language and culture, including things like how the teacher facilitates peer to peer groupings of DLLs with other children. In total, CASEBA includes 26 research-based items. Conceptually, these items capture five broad aspects of research-based supports for DLLs: (a) lead teacher HL supports, (b) assistant teacher HL supports, (c) English language supports, (d) environment, and (e) assessment.

Responding to the need for quality and quantity interactions, the underlying premise of the CASEBA is that use of high quality and meaningful interactions in the HL along with intentional and well-planned strategies for English language learning is the best approach to teaching preschool aged DLLs. As an example, two CASEBA items (see Table 2) were expressly designed to assess the way that teachers develop vocabulary, one in English (item 20) and the other in the HL (item 11). These parallel items were designed to assess how new words are introduced, defined, and contextualized throughout the day in both languages, thus highlighting the importance of learning new words in English and the HL. During observer training it was carefully explained that words that count for this item are not necessarily translations of each other, as words of varying difficulty would be appropriately considered "new" perhaps in the non-dominant language only. Further, the CASEBA includes four items that specifically evaluate the quality of language inputs (e.g., complexity of sentence structures, vocabulary used) for both the lead teacher (items 7 and 15) and the assistant teacher (items 8 and 16) separately in both the HL and English.

1.2. Elements of process quality for DLLs

Generally, the field of early childhood education has relied on two constructs for measuring quality in preschool classrooms that link to positive child outcomes. These include *structural* quality, which consists of visible classroom features (e.g., classroom size and materials) and *process* quality, which captures fewer tangible features like teacher-child interactions. While both have been shown to be predictive of later child outcomes, recent evidence has suggested that process quality is responsible for more direct relationships to academic outcomes for children when quality is high (Burchinal, Vandergrift, Pianta, & Mashburn, 2010; Mashburn et al., 2008). Research has also shown that process quality measures like the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008) are important for improving overall quality and increasing academic outcomes for children. However, quality in classrooms as measured by the CLASS has also shown that DLLs are disproportionately served in classrooms deemed to be low quality (Nores & Barnett, 2014; Valentino, 2018). Other research has further acknowledged that despite the success of more global observation tools (e.g., ECERS and CLASS) in helping to shape quality in the field of early childhood, they lack specific attention to how language and culture can influence student-teacher interactions in the classroom (Peisner-Feinberg et al., 2014; Vitiello, 2013). For these reasons, measures that specifically target processes deemed important for DLLs are needed not only to guide practice, but also to empirically understand what practices are most effective. Consequently, researchers have called for a need to develop and validate tools like CASEBA that can supplement general measures of quality and explicitly guide quality improvement (Peisner-Feinberg et al., 2014; Reilly, Johnson, Luk, & Partika, 2019; Valentino, 2018).

1.2.1. Supporting the home language in the classroom

One essential element of process quality for DLLs is the extent to which the HL is used for instruction. While we acknowledge that children need English inputs early, important research has also indicated that use of the HL for instruction along with the incorporation of home culture are significant features of effective practices yielding better child outcomes for DLLs than do English-only approaches (Barnett, Yarosz, Thomas, Jung, & Blanco, 2007; Burchinal, Field, Lopez, Howes, & Pianta, 2012; Farver, Lonigan, & Eppe, 2009; Gormley, 2008; Raikes et al., 2019; Spencer, Moran, Thompson, Petersen, & Restrepo, 2020). These findings suggest that not only can DLL children acquire English at rates equal to those of children instructed in English only, but that they also make greater progress in developing their HL skills when instruction is offered in both languages. In another study of Spanish speaking children in Head Start (n = 81), Davison, Hammer, and Lawrence (2011)) used a growth curve model to examine the factors predicting Spanish and English receptive vocabulary to later reading outcomes in first grade. Findings revealed that growth in children's Spanish receptive language positively predicted first grade English letter-word identification as well as passage comprehension.

Most research examining the use of Spanish in preschool classrooms even when a Spanish-speaker is present however, has found that Spanish is seldom used for instructional purposes, and most often focuses on directives (Hammer et al., 2020; Jacoby & Leseaux, 2014; Sawyer, et. al, 2018). In their study of Head Start teachers, Jacoby and Leseaux (2014) found that even lead teachers who identified as native Spanish speakers did not use Spanish for instruction and instead saw Spanish as a way to help children socially adapt to the classroom (Jacboy & Lesaux, 2014). In another study examining the amount of language inputs during the school day among both lead and assistant teachers, Sawyer et al. (2018) found that teachers (regardless of role) predominantly used English and that variation in language inputs did not occur based on children's DLL status or teachers' proficiency levels. Similarly, Head Start data from the Family and Child Experiences Survey showed that while 40% of lead teachers and 36% of assistant teachers in Head Start classrooms serving DLL children speak a language other than English, 92% of instruction is delivered in English only (Hulsey et al., 2011).

Seeking to specifically capture the quality of interactions between teachers and children in the HL, the CASEBA includes two subscales *Lead Teacher HL Supports* and *Assistant Teacher HL Supports* (see Table 2) with a total of eleven items to examine how this is carried out by teachers. All of these items distinctly focus on language interactions and supports used in the HL by teachers in the classroom. The CASEBA items are specifically designed to equalize use of the HL by providing children with opportunities to hear complex syntax, develop vocabulary and concepts, and engage in multi-turn conversations with teachers in the HL. The items also seek to capture whether these exchanges are evident across common groupings (individually, small group, large groups) throughout the day. Without evidence of HL instruction and/or conversations in classrooms, these items are assigned the lowest possible score.

1.2.2. Supporting english acquisition

In order for teachers to successfully manage English inputs for DLLs, they must create careful scaffolding to move children from one level of proficiency to the next via routine classroom interactions. To do this successfully, teachers must be intentional and understand the relationship between language ability and the demand of a task (Lucas, Villegas, & Freedson-Gonzalez, 2008). Fostering English acquisition therefore requires specific and well-planned interactions throughout the day to advance not only language, but concepts as well (Tabors, 2008). Research has indicated that these practices include things like providing explicit instruction in literacy components (Roberts, Vadasy, & Sanders, 2018), developing academic language that is embedded in content instruction (Pollard-Durodola et al., 2016), using visual cues to make content more comprehensible (Gámez, Neugebauer, Coyne, McCoach, & Ware, 2017), encouraging peer-assisted learning opportunities (Garcia, 2018; Reilly et al., 2019), using students' HL, knowledge, and cultural assets (Serafini, Rozell, & Winsler, 2020), and using small group supports (Landry et al., 2019). Other studies have pointed to the importance of language exposure, which suggests that language and literacy development depend on the *quantity* of their exposure to each language (Hammer et al., 2014). Further to be considered is the *quality* of language inputs, as research also shows that word diversity, grammatical structures, and contingent responsiveness matter as well (Farrow, Wasik, & Hindman, 2020; Gámez et al., 2017).

Based on this research, the CASEBA includes one subscale, *English Language Supports* (see Table 2) with 10 items focused on the quality and quantity of language interactions in English. These items seek to understand not only the quality of English language interactions (e.g., use of complex syntax, multi-turn conversations, and appropriate questioning) but also to examine the extent to which specific strategies to support DLL children's comprehension (e.g., use of pictures and gestures when introducing new vocabulary words) are used. As with the *Lead* and *Assistant Teacher HL Supports* subscales, the items of this subscale also aim to assess the degree to which each of these kinds of interactions occur across learning formats and groupings.

1.3. Elements of structural quality for DLLs

Taken together, the implementation of high-quality processes heavily relies on the quality of the interactions nurtured by teachers, and this depends on teacher knowledge and behavior in the domains of English language development and language supports for the HL (Zepeda, Castro, & Cronin, 2011). To nurture responsive environments, it is critical that teachers are knowledgeable about children's linguistic and cultural backgrounds. Environmental features relative to this include language of displays, play materials, and learning topics (Gay, 2002; Tabors, 2008). To this end, the CASEBA *Responsive Environments* subscale includes five items (see Table 2) that address what teachers know about children's linguistic and cultural backgrounds (item 2) and how they are incorporated into the life of the classroom (items 3,12,13, and 14). This subscale includes the provision of structural elements such as play materials that reflect cultural diversity, books in the HL in equal quantities to books in English, and manipulatives that encourage literacy skill development in the HL.

1.4. Classroom quality observation tools to understand contexts for DLLs in preschool

Currently, there are several measures used in research to examine classroom quality and constructs relative to DLLs in preschool. However, other than CASEBA, none of these measures expressly assess the quality of language inputs in both English and Spanish for preschool classrooms. One of these is *The Language Interaction Snapshot* (LISn; Atkins-Burnett, Sprachman, López, Caspe, & Fallin, 2011), which utilizes a time-use method to examine various kinds of language interactions (as defined by the measure) between teachers and focal children through the course of the day through repeated 30 s observations over the course of the day. While this measure is used to assess the frequency and quality of language interactions between teachers and children, the language interactions captured are general and do not focus on the specific strategies recommended by research for DLLs. Another measure is the *Early Language and Literacy Classroom Observation-DLL* (ELLCO-DLL; Castro, 2005), which focuses somewhat on DLLs, but mostly assesses the literacy environment (presence of bilingual books, use of gestures, props during book reading) and whether the teacher reads to DLLs individually, in small groups, or in large groups in any language. While all these tools assess language and literacy supports for DLLs (in the case of the ELLCO-DLL), none capture specific research-based practices that have been demonstrated to support DLLs' language acquisition in both English and the HL.

1.5. Presence of bilingual preschool teachers

Due to increased awareness for the need to use the HL in the classroom, attention to the availability of bilingual teachers is critical, and quality processes are dependent on the extent to which teachers provide high-quality language interactions even when teachers are bilingual. In a study of five state-funded preschool programs, researchers found that 64% of sample teachers were White and 15% Latino, and that 32% of teachers reported that they or their assistant spoke Spanish in the classroom (Early et al., 2005). Also noted recently was that a pipeline of potential Spanish-speaking and certified teachers is also narrow (Bridges & Dagys, 2012; Buysse, Castro, West, & Skinner, 2005). These patterns are likely attributable to low rates of college enrollment in 4-year programs as well as completion of a BA by Hispanics (Fry & Taylor, 2013). One dissertation study notes the difficulty with completing even an AA by native Spanish-speaking students in a community college, despite a high interest in early childhood, due to language barriers (Eberly, 2015). A byproduct of this phenomenon then creates the possibility of Spanish-speaking adult students to instead pursue the assistant teacher position, which typically requires less education and credentialing (Whitebook et al., 2018). This option also enables a classroom to be staffed with more than one teaching staff at a lower cost (Ryan & Whitebook, 2012). The tradeoff is that assistant teachers are also less likely to be versed in practices specific to early childhood and DLL specific pedagogy due to fewer requirements for specific education and certifications in early childhood.

1.6. Current study

The goal of this study was to examine whether differing staff language configurations among teachers and assistant teachers yield different outcomes on different measures of observed classroom quality. To do this, the CASEBA — a domain specific tool — was designed and used to capture specific supports of classroom quality for young DLLs including contributions by each teacher role in the classroom. Specifically, the present study sought to address four major aims. First, we aimed to examine the latent factor structure of the CASEBA to ensure that it measures the latent factors that it was designed for. Second, we aimed to examine the relationship between the CASEBA and staff language arrangements to investigate how the tool is able to distinctly provide data that can later inform policy and planning. Third, we aimed to investigate the degree to which there is overlap between the features of classroom quality for linguistically diverse classroom settings as measured by the CASEBA, and those of a tool of global classroom quality tool as measured by the ECERS-R. We hypothesized that while classroom scores may not differ on the ECERS-R, differences in quality by staffing arrangement would be detected by the CASEBA. Finally, we examined whether staff language arrangements, the CASEBA, and the ECERS-R had any significant relationship with children's English and Spanish receptive vocabulary scores after controlling the child and teacher-level background variables. We hypothesized that, given the empirical research base on which the CASEBA was predicated, children's English and Spanish receptive vocabulary would be better associated with staff language arrangements in general, but also as well as how they are defined by the CASEBA, rather than by the ECERS-R.

2. Methods

2.1. Sample

The current study was based on a larger randomized trial study that utilized a pre/post-test design to examine the effects of the Professional Development (PD) intervention aimed at providing language and literacy strategies for DLLs. In that study, children were randomly assigned to classrooms with three different staffing structures that took into account teacher and teacher assistant languages. These teams included a Spanish speaking lead teacher with an English-speaking assistant (S—E), a Spanish speaking lead with a Spanish speaking assistant (S—S), and an English-speaking lead teacher with a Spanish speaking assistant (E—S). Teachers in the ES— group were randomly assigned to two different PD conditions. While the randomized portion of the study examining effects of PD only involved English speaking lead teachers, the Spanish speaking lead teacher groups were also included in the PD group. The purpose of this was to ensure that all teachers participating in the staff language configuration piece of the study had the same information and strategies from which to draw for teaching, and particularly to encourage the use of Spanish for instruction. As no effects of the PD were found on E—S treatment teachers (did not outperform the control group on CASEBA), both the experimental and control E—S groups were included in the analysis of the current study. This paper focuses on the relationships of staff language configurations on classroom supports and children's receptive vocabulary outcomes as outlined by the CASEBA. The goal of the study was to explore the ways in which CASEBA can strengthen information about supportive practices for DLLs in preschool.

For the CASEBA factor analysis only, we used a preliminary dataset from the pre-test period of the larger study, which occurred in spring 2009 prior to the start of the current study (Fall 2009-Spring 2010). This data set contained a larger sample (N = 91) from which the current study was a subset (N = 82). Further, this data set did not include staff language designations but was used because of its larger size.

2.1.1. Site context

The school district in which the study took place is located in a 71% Latino city on the east coast, in which all three- and four-year-olds are eligible for a free high-quality preschool education. All classrooms in the study were part of state-funded preschool and subject to district and state policies. Though some contracting of private providers helped to accommodate children, all classrooms in this study were housed in public school buildings. All teachers were early-childhood certificated (a four-year-degree with specialization), and all assistants held a CDA. Class sizes were limited to 15 children, and the district was required to choose from one of six state-vetted curricula. The curriculum for all sites in the study was High/Scope, which is a research-based, emergent curriculum (directed by children's interests) premised in the whole-child philosophy. In addition, the district was also required to monitor classroom quality on a yearly basis using an approved measure of early childhood classroom quality. The district enrolled almost 100% of its age-eligible preschool population, which was about 91% Hispanic.

2.1.2. Teachers and staff language configurations

—The ES group was the largest and was split randomly into a control and treatment group for a professional development intervention. Though the district did not previously use teachers' language to create teaching teams, they did adjust their process for the purpose of the study. Consequently, all language designations for teaching pairs were provided strategically by the district for the purpose of the study and were arranged to include as many types of each group as possible based on the population of teachers that were currently employed in public school-based classrooms. As often as possible, teaching teams that met any of the three group types designated by the research team were kept together so that only team types that were needed were newly created. No measures of proficiency in either English or Spanish were used to classify teachers as speakers of either language. Teachers reported that they felt they were assigned to staff language structures (e.g., Spanish vs. English) based on conversations and intake paperwork at the time of their interview and hire. Further, though teacher coaches also participated in PD, there was no consistent guidance over the course of regular coaching interactions about how or when to use Spanish for instruction throughout the school day. Due to teacher turnover and incomplete background data provided by teachers, numbers varied from the beginning of the study to the end. While the sample of teachers had high levels of education, it is interesting to note that only 10% of lead teachers in the S—S group had a Master's degree, as compared to 49% in the E—S group and 31.8% in the S—E group. Similarly, 100% of assistant teachers in the assistant teacher group held a Bachelor's degree in the S—S group, as compared to 41.7% of assistant teachers in the S—E group and 58.3% in the E—S group. Still, chi-square tests confirmed that lead teacher and assistant teacher's education status did not differ by staff language groups, as demonstrated by $F(6) = 12.198$, $p = .058$

Table 1
Teacher and child characteristics by staff language configurations.

		E–S	S-S	S-E
Classrooms		N = 49	N = 10	N = 23
Children		N = 209	N = 38	N = 129
		Mean (SD) or %		
Lead Teachers				
	Years Exp. in ECE	7.68 (5.0)	8.33 (3.44)	5.74 (3.02)
Degree	BA not finished	2.00		4.50
	BA	49.00	90.00	54.50
	MA	49.00	10.00	31.80
	Doc. degree			9.10
Certification	No Certificate			4.30
	Elementary Cert w/N-K			4.30
	P-3 Cert	95.70	100.00	91.30
	Elementary Cert w-N-8 with major in ECE	4.30		
Assistant Teachers				
	Years Exp. in ECE	6.16 (4.33)	4.06 (2.81)	5.21 (2.53)
Degree	BA Not Finished	30.60		41.70
	BA	58.30	100.00	41.70
	MA	5.60		16.70
	Doctoral degree	5.60		
Certification	Elementary Cert w/N-K	2.60		
	P-3 Cert	23.10	33.30	6.70
	No Cert	74.40	66.70	93.30
Children				
	PPVT Pre	64 (19)	67 (19)	67 (19)
	PPVT Post	73 (18)	70 (18)	74 (17)
	TVIP Pre	80 (12)	82 (11)	80 (11)
Age	Age	46 (7)	44 (5)	45 (6)
	3-year-old	75.60	84.20	72.90
	4-year-old	24.40	15.80	27.10
Gender	Female	52.60	65.80	47.30
		E–S	S-S	S-E
Classrooms		N = 49	N = 10	N = 23
Children		N = 209	N = 38	N = 129
		Mean (SD) or %		
	Male	47.40	34.20	52.70
Ethnicity	Asian	3.40	5.30	0.80
	Black	4.90		4.80
	Hispanic	89.70	92.10	93.50
	White		2.60	
	Other			0.80
Dominant Language	English	38.90	32.40	43.20
	Spanish	50.30	67.60	41.50
	Both English and Spanish	10.90		15.30

Note. E-S = English speaking lead teacher with Spanish speaking assistant, S-S = Spanish speaking lead and Spanish speaking assistant; S-E = Spanish speaking lead and English-speaking assistant; Exp. = experience; ECE = Early Child Education; AA = associate degree; BA = bachelor degree; MA = master degree; Cert = certification.

for lead teachers, and *F*(6) = 8.438), *p* = .208 for assistant teachers. Teacher characteristics for all groups based on the fall (N=82), are reported in Table 1. We would like to note that as teachers returned for spring observations, staff language configurations remained the same in each classroom.

2.1.3. Children

A total of 376 three- and four-year old children were included in the overall sample (see Table 1). The average age of our child sample was 46 months at baseline. Fifty-two percent of children were female and 89.7% were Hispanic. Fifty percent of children were reported as being from homes where Spanish is spoken, while only 10.9% reported using both Spanish and English at home.

2.2. Assessment procedures and measures

Observations in all staffing groups were conducted during one visit where both the CASEBA and ECERS-R were used simultaneously for a three to four-hour period during a morning of instruction. Observations were conducted by trained observers in the fall and again in the spring. For the purposes of this study, only observations from the spring data collection were used. All observers were also required to be bilingual as the instruments aimed to capture and count interactions in both English and Spanish. Training for both instruments was conducted over the course of one week in a classroom setting. On the first day the tool was explained item by item with video and picture examples to illustrate what interactions would count as evidence, followed by live reliability visits with the CASEBA's authors in actual preschool classrooms. Each observer was held to a standard of 80% within-one agreement with respective reliable observers. Observers were required to achieve these levels of agreement on both measures three times at the start of the study, with drift observations occurring after the tenth observation for each observer. This process was repeated in the spring data collection period in the same manner, along with a refresher training for all observers. All observers were expected to conduct the observations simultaneously and were therefore held to this expectation, even during reliability visits.

Child assessors were trained on each child assessment just prior to the start of data collection, and assessments were conducted in the fall and spring. Training consisted of one day of in-classroom learning time to learn the measures, and one day of one-on-one shadow scoring to ensure 100 percent accuracy. A one-day refresher training on all assessment measures took place in late spring, just prior to the final round of child assessments. All children were given both the PPVT and TVIP in the fall and again in the spring. Descriptive statistics for all variables used in the analyses of each measure are presented in Table 1.

2.3. Classroom observations

2.3.1. CASEBA

The CASEBA served as the main measure of the quality of language and literacy supports offered by the teachers with a specific focus on DLLs and HL maintenance. The CASEBA was a newly developed research tool designed to assess the degree to which preschool teachers and classrooms provide support for the social, cognitive, and linguistic development of DLLs, with a focus on language and literacy.

Each of the items on the CASEBA are rated on a 7-point Likert scale with scores of 1 (*no evidence*), 3 (*minimal evidence*), 5 (*good evidence*), and 7 (*strong evidence*) serving as anchors. Indicators under each of these anchor scores guide observers by posing questions about the evidence of specific practices and materials respective to the overall item to reach final item scores. Indicators were designed to guide observers from left (*no evidence*) to the right (*strong evidence*), with indicators organized by related strands building cumulatively upon each other so that evidence of similar but more ideal forms are described under its respective anchor score. Practices, interactions, and materials were quantified within each indicator and observers track the number of times a practice or interaction occurred to subsequently answer indicator level questions effectively to reach an overall score. Consequently, overall item scores are based on whether some or all of the indicators under the anchor scores were met or not. In-between scores (2,4,6) were created from the presence of some but not all of the indicators being met under the anchor scores.

Items 1, 2, and 26 (see Table 2) rely on teacher interviews which were conducted prior to the start of the observation. The purpose of interview questions for the first two items of the tool was to directly assess the degree to which teachers know the language and cultural background information about their students and how this data is collected. Other interview questions were embedded within items that examine family engagement (Item 14) relative to HL use and the other centering on curriculum (Item 25). For each interview question there were respective indicators that allowed responses to be reflected in the overall score. A total of four items (5, 6, 15, and 16) related respectively to lead and assistant teacher interactions in both English and the HL, with all other items considering the collective interactions observed across all teaching staff working with children. All other CASEBA items aimed to capture all interactions throughout the observation regardless of which teacher role is engaged. As with other measures of quality, all interactions pertinent to the observation were counted towards the score. Generally, observers gathered information within corresponding items on a score sheet over the course of the whole observation period but did not assign overall scores until the end of the observation immediately after they have left the classroom. The full list of CASEBA items can be found in Table 2.

2.3.2. ECERS-R

The *Early Childhood Environmental Rating Scale - Revised* (ECERS-R; Harms, Clifford, & Cryer, 2005) was used to provide a global measure of preschool classroom quality. With 43 items that cover a broad range of quality considerations from safety to teacher-child interaction to parent involvement, the study utilized items 1–37, which covered all of the aspects of the instrument with the exception of the teacher self-report items at the end. The ECERS-R is widely used (National Center on Early Childhood Quality Assurance, 2016) and has been studied extensively (Gordon et al., 2015; La Paro, Thomason, Lower, Kintner-Duffy, & Cassidy, 2012). Prior research on reliability and validity has found a 70% agreement at the indicator, item, and total score levels and that Cronbach's alpha for the ECERS-R is above 0.90 (Harms et al., 2005).

2.4. Child assessments

2.4.1. Receptive vocabulary

The Peabody Picture Vocabulary Test – Third Edition (PPVT – III; Dunn & Dunn, 1997) and Test de Vocabulario en Imágenes Peabody (TVIP; Dunn, Lugo, Padilla, & Dunn, 1986*)* were used to measure children's English and Spanish receptive vocabulary. The PPVT–III is a 204-item test of receptive vocabulary in standard English. The Spanish language counterpart to this measure, the TVIP, uses 25 translated items from the PPVT to assess receptive vocabulary acquisition of Spanish-speaking and bilingual students. These instruments were normed on separate populations of native language speakers and have established split/half and test/retest reliability as well as concurrent and predictive validity (Dunn et al., 1986).

2.5. Analysis

2.5.1. CASEBA factors analysis

There have been two previous studies which have examined CASEBA's factor structure. Both utilized data from 100 publicly-funded preschool classrooms in districts with large Spanish bilingual populations, In the first study, a confirmatory factor analysis (CFA) was used to find a five-factor solution with 25 items (item 26 was excluded) including: (1) supports for English acquisition, (2) supports for English print literacy, (3) supports for HL, (4) culturally responsive environment, and (5) knowledge of child background (Freedson et al., 2011). Freedson et al. (2011) also included a concurrent validity analysis by estimating correlations between CASEBA and ECERS-R scores. Findings suggested a weak correlation between the CASEBA's supports for HL factor and language interactions measured by the ECERS-R, suggesting that the CASEBA captures language supports for DLLs in different ways from the ECERS-R.

Valentino (2015) revisited the same dataset used by Freedson et al. (2011) and found support for a four-factor structure with 20 items (excluding items 2, 3, 13, 16, 21, and 22) based on both exploratory and confirmatory factor analysis which capitalized on the latent dimensions of lead and assistant teachers. Valentino's factors included: (1) assistant teacher HL (2) lead teacher HL and (3) English language development and (4) classroom, structure/environment. In this study, Valentino (2015) also found evidence of the predictive validity, with the CASEBA predicting growth in DLLs' vocabulary and executive functioning skills on classrooms scoring using a threshold analysis.

To address the first aim, we conducted an exploratory factor analysis (EFA) to identify the most appropriate model solution to fit our analytic sample, the initial data collection from the previous spring. After the EFA, we confirmed the identified model again using CFA to address any cross-loading items and to finalize the model structure with each item only loaded on one factor. Finally, we compared the model fit of our final model against the five-factor model by Freedson et al. (2011) and the four-factor model by Valentino (2015).

Both CFA and EFA were conducted in M*plus* 8.2 (Muthén & Muthén, 1998-2019Muthén & Muthén, 1998-2019) based on a mean and variance adjusted robust weighted least squares (WLSMV) to account for the ordinal nature of item responses (Muthén & Muthén, 1998-2019Muthén & Muthén, 1998-2019). To evaluate the model fit, we adopted the following criteria: (1) the chi-square statistic (Jöreskog, 1969; Muthén & Muthén, 1998-2015; Wang & Wang, 2012), (2) comparative fit index (CFI; Bentler, 1990) and the Tucker-Lewis index (TLI; Tucker & Lewis, 1973) > = 0.95

Table 2
Final Five-Factor Model Configured through current Confirmatory Factor Analysis and Item Descriptive Statistics.

	Item Description	Standardized Factor Loading	
		Estimate	SE
Lead Teacher HL Supports	4. The lead teacher uses a home language of the ELL children for instructional purposes (Barnett, et al., 2007; Durán, et al., 2010; Farver, et al., 2009; Restrepo et al., 2010; Ryan, 2007).	0.818	0.026
	7. The lead teacher uses high quality talk in the students' home language (Cummins, 1979, 2000; Vitiello, Downer, & Williford, 2011).	0.862	0.027
	10. Teaching staff interact with individual ELL children in ways that support on-going development of the home language (Gonzalez et al., 2014).	0.879	0.035
	11. Teaching staff intentionally expand children's repertoire of concepts and vocabulary in the home language (Méndez, Crais, Castro, & Kainz, 2015; Castro, et al., 2006)	0.906	0.033
Assistant Teacher HL Supports	5. The paraprofessional or assistant teacher uses a home language of the ELL children for instructional purposes (Barnett, et. al, 2007; Durán, et al., 2010; Farver, et al., 2009; Restrepo et al., 2010; Ryan, 2007).	0.914	0.020
	8. The assistant teacher uses high quality talk in the students' home language (Cummins, 1979, 2000; Vitiello et al., 2011).	0.943	0.021
	9. Teaching staff use effective strategies during group instruction to support on-going development of the home language (Barnett, et al, 2007; Farver, et al., 2009; Jacoby & Leseaux, 2014).	0.819	0.036
English Language Supports	6. The teacher attempts to learn and use the home language/s spoken by the ELL children in the classroom, although she/he lacks proficiency in the language (Espinosa, 2010; Gillanders, 2007; Lee & Oxelson, 2006).	0.418	0.051
	15. The lead teacher uses high quality talk in English (Dickinson & Porsche, 2011; Hindman & Wasik, 2012; Lipsky, 2013).	0.863	0.031
	16. The assistant teacher uses high quality talk in English (Dickinson & Porsche, 2011; Hindman & Wasik, 2012; Lipsky, 2013).	0.596	0.037
	17. Teaching staff use effective strategies to scaffold children's comprehension of instructional content in English (Galloway & Lesaux, 2017; Garcia & Frede, 2010).	0.867	0.026
	18. Teaching staff use effective strategies during group instruction to build children's communicative skills in English (Jacboy & Leseaux, (2014).	0.853	0.028
	19. Teaching staff interact with individual ELL children in ways that support the acquisition of English (Collins, 2010; Garcia & Kleifgan, 2010; Gersten & Geva, 2003; Gillanders & Castro, 2011).	0.863	0.026
	20. Teaching staff intentionally expand children's repertoire of concepts and vocabulary in English Dickinson & Porsche, 2011; Jacoby & Leseaux, 2014).	0.820	0.030
	21. Books, print and literacy props are available in English (August, et al., 2005; Davison et al., 2011; Garcia & Kleifgan, 2010; Gillanders, Castro, & Franco, 2014; Tabors, 2008).	0.868	0.031
	22. Teaching staff support the learning of word-level early literacy skills in English (Morrow & Schickedanz, 2006; Gersten & Geva, 2003).	0.819	0.028
	24. Teaching staff foster a calm and respectful learning environment in which ELL children are able to hear adult talk (Gillanders, 2007; Piker & Rex, 2008; Tabors, 2008).	0.715	0.036
	25. Teacher staff create a content-rich curriculum that offers meaningful opportunities to acquire and use new language skills (Gillanders et al., 2014).	0.656	0.041
Responsive Environments for DLLs	2. The teacher knows the language and cultural background of each child in the classroom (Garcia, 2012).	0.642	0.041
	3. The cultural backgrounds and life experiences of the ELL children are incorporated into the life of the classroom (Gay, 2002; Tabors, 2008).	0.768	0.036
	12. Books, print, and literacy props are available in the ELL children's home language/s (Castro, Ayankoya, & Kasprzak, 2011).	0.799	0.035
	13. Teaching staff support the learning of word-level early literacy skills in the ELL children's home language/s (Davison et al., 2011; Jackson, Schatschneider, & Leacox, 2014).	0.882	0.032
	14. Teaching staff encourage ELL parents to maintain children's home language (Barrueco, Smith, & Stephens, 2015; Galloway & Lesaux, 2017).	0.711	0.038
Assessment for DLLs	1. The teacher and/or center collect systematic information on the language and cultural background of each child in the classroom (Garcia, 2012; Garcia, Arias, Murri, & Serna, 2010; Tabors, 2008).	0.890	0.040
	23. Teaching staff provide a warm, emotionally supportive and low-anxiety classroom environment for English language learners (Galloway & Lesaux, 2017; Goodrich, Lonigan, Kleuver, & Farver, 2015).	0.766	0.046
	26. Teaching staff use appropriate assessment practices to identify children's language strengths and needs in their home language and in English (Barrueco, et al., 2010).	0.912	0.024

for good model fit and > = 0.90 for acceptable model fit; (3) the root-mean-square error of approximation (RMSEA) $\leq$ 0.08, and the standardized root-mean-square residual (SRMR; Asparouhov & Muthén, 2018) <.08 (Brown, 2006; Yu, 2002). The CFA approach was used to address the cross-loading issues we encountered in the EFA models and to finalize the factor structure for our CASEBA tool. To compare and select better fit model, we adopted a set of stricter cutoff criteria: ΔCFI > 0.002 and ΔRMSEA $\geq$ .007 (Meade, Johnson, & Braddy, 2008).

2.5.2. *Associations between staff language groups and classroom quality*

After the subscale structure was determined for the CASEBA, we addressed the second and third aims, and calculated the average of the observed item scores within each factor to index the corresponding factor to be consistent with the scoring scheme by other researchers on CASEBA and the ECERS-R. To define the degree of correlation, we adopted Rather's (2009) suggestion whereby: (1) values between 0 and 0.3 (0 and −0.3) indicate a weak positive (negative) linear relationship; (2) values between 0.3 and 0.7 (0.3 and −0.7) indicate a moderate positive (negative) linear relationship; and (3) values between 0.7 and 1.0 (−0.7 and −1.0) indicate a strong positive (negative) linear relationship.

We then examined how each of the CASEBA subscales were associated with the staffing group: E—S, SS, and SE——. Multivariate analysis of variance (MANOVA) was used to analyze the CASEBA subscales (i.e., average of items identified for the corresponding factor) and ECERS-R subscale scores. Tukey post-hoc tests are further reported for each subscale of the CASEBA and ECERS-R, if any significant difference was detected in MANOVA. Correlations among CASEBA and ECERS-R subscale scores were also examined. Bonferroni correction was used when making multiple comparisons for factors and items to reduce the chance of Type I error (Dunn, 1961). Effect sizes were also reported indexing partial eta squared. The cutoff points we adopted were partial eta-squared greater than 0.01 for small effect size, greater than 0.06 for medium effect size, and greater than 0.14 for large effect size (Cohen, 1988; Miles & Shevlin, 2001).

2.5.3. *Association between classroom characteristics and child receptive vocabulary*

To address the final aim, linear mixed models were implemented to analyze how child outcomes as defined by the PPVT and TVIP at the end of the school year were associated with the CASEBA subscales, the ECERS-R subscales, and staff language groups accounting for the multilevel structure of our data. We conducted two series of models for PPVT and TVIP separately. In both series, the PPVT/TVIP standardized scores at the end of the school year were used as outcome variables. At the child level, child gender, ethnicity, HL, PPVT/TVIP standardized scores at the beginning of the school year were used as control variables. At the classroom level, lead and assistant teacher years of experience, education, and certification status were used as control variables. In the first steps of the model series, we included only child and classroom level control variables. In the second steps, we added the staffing configuration variable, and kept it in the model as it is the focus variable of the current study. In a following step, we subsequently tested CASEBA and ECERS subscales one at a time by adding each to the second-step model one by one. A total of 13 models were run for PPVT and TVIP separately. SPSS 26.0 Mixed Models function was used for our analysis. To address any missing data issue in the mixed models, full information maximum likelihood estimation that is available in SPSS 26.0 was applied, which makes use of any available data in estimation of a parameter that is most likely to have resulted in the observed data (Collins, Schafer, & Kam, 2001; Hox, 1999).

3. Results

3.1. *CASEBA factor analysis*

In efforts to meet the study's first aim, we discovered a five-factor model based on our analytical sample through EFA (see Table A1). We found that five out of the 26 CASEBA items had cross-loading issues. To define cross-loading items, we combined two criteria: (1) the item-factor loadings of an item are greater than 0.40 on more than one latent factor (Maskey, Fei, & Nguyen, 2018); (2) the difference between the two item-factor loadings is less than 0.20. To test models with cross-loading item issues, we first relied on professional judgment by specifically examining the cross-loaded items and their focus. Most specifically, our theorized models for testing took into account the content of each item, and whether the focus was explicitly about lead or assistant teacher interactions and when items focused on English language supports. We subsequently used these judgements alongside the magnitude of the item-factor loadings then tested the model with CFA approach. The final five-factor model we proposed and confirmed is shown in Table 2. The goodness-of-fit test of the presented 5-factor model is chi-square = 648.908 (df = 289, p < .000), which is lower than the other two competing models; RMSEA = 0.117 > .08 (which did not pass the cutoff, but was lower than the other two competing models), CFI = 0.967 and TLI = 0.963 (both passed the 0.95 cutoff, and were greater than the other two competing models); SRMR = 0.075 (passed the 0.08 cutoff and outperformed the other two competing models).

We then tested the Freedson et al. (2011) and the Four-Factor model by Valentino (2015) through CFA and compared the model fit indices with our final five-factor mode did not outperform the five-factor model we discovered through EFA in terms of model fit. As shown in Table 3, the difference between our model CFI and Freedson's model is 0.005 > 0.002 and the RMSEA difference is 0.013 > 0.007. According to the model selection criteria stated in the Method section above, we determined that the currently proposed CFA model better fit our data sample. In addition, Freedson et al. (2011)'s five-factor model excluded item 16 (the assistant teacher uses high quality talk in English), and item 26 (teaching staff assess strengths and needs in both languages), which we considered important assets to have for the CASEBA, as well as critical in seeking to meet our aim about staffing arrangement and two of their factors only had two items loaded on, which lead to instability of the factor structure. Similarly, Valentino's (2015) model excluded six items, and our data did not converge on this model. We therefore assert that the final five-factor model we discovered through EFA and confirmed with CFA had good and better fit than other two competing models. Table 2 presents the standardized item factor loadings for each of the 26 items in CASEBA.

Based on the five-factor model, we define five CASEBA subscales. Within the first factor, *Lead Teacher HL Supports*, both items addressing the lead teacher's use of the home language for instructional purposes were included (items 4 and 7). Additionally, two items about interactions with DLLs in ways that support ongoing development of the HL and vocabulary development (items 10 and 11) were confirmed, thus leading this factor to be associated with how the lead teacher nurtures the HL. In the second factor, *Assistant Teacher HL Supports,* the same pattern emerged with both items related to the assistant teacher's use of the home language for instruction (items 5 and 8), as well as one other about use of strategies during group instruction to support ongoing development of the HL (item 9). In the third factor, 11 items, all covering English language supports to promote comprehension and build vocabulary, were confirmed (items 6–22 and 24–25), making it natural to denote this factor as that of English Language Supports. In the fourth factor, *Responsive Environments*, items dealing with the presence of materials, books, and displays in the classroom environment were confirmed (items 3 and 12). In addition, item 13 loaded, which fits, given that this item seeks to capture the ways that the classroom incorporates the HL for literacy skill building including signs, labels, and manipulatives related to letters. Item 14 loaded on this factor as well. While it relates to the encouragement of HL maintenance with families, it partially covers the presence of lending libraries and other classroom features that seek to support families. Item 2 also loaded on the fourth factor regarding whether teachers know the HL and cultural background of children, which made sense: without this knowledge, teachers were not able to set up responsive environments. Finally, the fifth factor, *Assessment*, had three items confirmed, which were all related to assessment, though item 26 is most specifically about this. Item 1 — about whether teachers collect data on children's language — seemed related, as without this information teachers are unable to know about where children stand relative to language proficiency in each of their languages.

Table 3
Factor Structure Model Fit through Confirmatory Factor Analysis.

Model Configured by	Number of Items in the model	Chi-Square	df	P-Value	RMSEA	CFI	TLI	SRMR
Five Factor Model by Freedson, et al. (2011)	25	671.285	265	0	.130	.961	.955	.080
Four-Factor Model by Valentino (2015) with 20 items	20	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Current study CFA	26	648.908	289	0	.117	.967	.963	.075

Note. Freedson et al. (2011) N = 100; Valentino N = 91, Current Study CFA N = 91; df = degree of freedom; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis Index; SRMR = standardized root mean squared residual; n/a = model did not converge due to positive definite issue.

Table 4
MANOVA Test Results for CASEBA and ECERS-R subscale scores by staff language configuration.

	E–S (N = 49)		S-S (N = 10)		S-E (N = 23)		Test of Between Subjects Effects			Post-Hoc Tukey HSD Test
	M	SD	M	SD	M	SD	F	*p*-value	Partial Eta Squared	Mean (E–S) – (S-S)
CASEBA										
LT HL Support	2.99	.70	3.95	.81	3.95	.90	15.367*	<.01	.275	E–S < S-S*E –S < S-E*
AT HL Support	4.40	.85	4.70	.81	2.22	.84	58.772*	<.01	.592	S-E < E–S* S-E < S-S*
English Lang, Support	4.29	.55	4.28	.39	4.26	.53	.028	.973	.001	
Environment	3.29	.58	3.92	.78	3.63	.68	5.246*	.007	.115	E–S < S-S*
Assessment	5.49	.97	5.77	.82	5.45	.89	.439	.646	.011	
ECERS										
Space/Furn.	5.33	.79	5.31	.59	5.62	.63	1.442	.243	.035	
Personal Care	4.95	1.12	5.05	1.06	5.25	1.23	.582	.561	.014	
Lang/Reasoning	4.78	1.04	4.85	1.22	4.92	.96	.159	.853	.004	
Activities	4.72	.68	4.68	.68	4.89	.62	.641	.529	.016	
Interactions	5.91	.83	6.14	.40	5.88	1.06	.343	.710	.009	
Program Structure	5.96	.91	5.91	1.16	6.24	.67	.940	.395	.023	

Note. M = Mean; SD = Standard Deviation; LT = Lead Teacher; AT = Assistant Teacher; SD = standard deviation; E-S = English speaking lead teacher with Spanish speaking assistant, S-S = Spanish speaking lead and Spanish speaking assistant; S-E = Spanish speaking lead and English-speaking assistant; HL=Home Language; * Bonferroni correction was adopted at the univariate comparison and post-hoc Tukey HSD Test.

Item 23, about the provision of warm classroom environments in this factor, seems unexpected, but given that it includes the ways in which teachers group children and help them to feel comfortable also has reasonable fit here given that teachers must have some assessment of children in order to be able to do this successfully.

Associations Between Staff Language Groups and the CASEBA. To meet the second aim of the study, and to understand staff language configurations and their relationships to CASEBA, we examined CASEBA item and subscale means across groups. All of the individual CASEBA subscale score means and standard deviations by staff language group are presented in Table 4.

The MANOVA revealed a statistically significant multivariate difference in the CASEBA subscale scores across the three staff language groups, $F(10, 154) = 18.947$, $p < .001$; Wilk's $\Lambda = .201$, partial $\eta 2 = 0.552$, indicating a large effect size. We further examined each CASEBA subscale and found that the *Lead Teacher HL Supports*, and *Assistant Teacher HL Supports*, and *Responsive Environments* differed across staff language groups after we made the Bonferroni correction to set alpha level at 0.01 (= .05/5), as shown in Table 4. The Tukey HSD post-hoc test was then performed, and we found that the E—S group scored significantly lower than the S—S group after setting alpha level at 0.017 (= .05/3). Results can be found in Table 4.

As a supplement to the MANOVA tests on the CASEBA subscale scores, we conducted a series of MANOVA analyses on the individual CASEBA items within each of the CASEBA subscales separately (see Table 5). The MANOVA tests were considered more appropriate, as the items within their corresponding subscales are associated with each other. According to the series of MANOVA tests, some statistically significant differences were found. These included: (1) a MANOVA test for all four items in the *Lead Teacher HL Supports* across staff language groups, $F(8, 156) = 17.656$, $p < .001$; Wilk's $\Lambda = 0.275$, partial $\eta 2 = .475$ and the follow-up post-hoc test demonstrated that items 4 and 7 were significant; (2) a second MANOVA test for all three items (5, 8, and 9) in *Assistant Teacher HL Supports* across staff language groups, $F(6, 158) = 27.675$, $p < .001$; Wilk's $\Lambda = .238$, partial $\eta 2 = .512$ and the follow-up post-hoc test demonstrated that all three items were significant; and (3) a third MANOVA test for ten items in *English Language Supports*, $F(20, 144) = 1.570$, $p < .001$; Wilk's $\Lambda = .499$, partial $\eta 2 = .294$ and the follow-up post-hoc test demonstrated that items 15, 16, and 19 were significant. All three partial η2 are greater than .14, indicating large effect sizes. No significant difference was detected for items in the *Responsive Environment* or *Assessment* subscales. Detailed Tukey HSD post-hoc tests were reported for the first three subscales accordingly (see Table 5). Generally findings were consistent with what we found at the subscale level, except for *English Supports* where we found that the S—E group scored lower than the E—S group on item 15 and 19, but scored higher than the E—S group on item 17 (see Table 5) for the means of each item across staff language groups). The opposite difference may eventually cancel out the effect of the *English Language Support* subscale as a whole.

In summary, for the *Assistant teacher HL* and the *Teacher HL Support* subscales, individual items remained consistent with the results of the individual item averages, suggesting this thread of differences across staff language configurations is evident through all or most of the items in these subscales. In contrast, whereas the item averages differed for the *Responsive Environment* subscale, none of the individual items in this subscale differed by staff language configuration, suggesting weaker evidence for this thread. Finally, new differences emerged for the *English Language Supports* subscale at the individual item level that were not apparent in the subscale-level item averages.

3.2. Associations of staff language groups, CASEBA and the ECERS-R

In an effort to further address the third aim of the study and understand how the two tools comparatively capture elements of classroom quality as well as differences among staff language configurations, we compared CASEBA to the ECERS-R. To begin with, correlations were run between ECERS-R and CASEBA. The correlations between the CASEBA and ECERS subscale scores ranged from 0.508 to .940, indicating moderate to strong positive correlation (Rather, 2009) and the correlations between these two tools were

Table 5
Items MANOVA Tests for Each of the CASEBA Subscales.

CASEBA	E–S (N = 49)		S-S (N = 10)		S-E (N = 23)		Test of Between Subjects Effects			Post-Hoc Tukey HSD Test
	M	SD	M	SD	M	SD	F	*p*-value	Partial Eta Squared	
LT HL Support; alpha = .0125 (= .05/4)										
Item 4	1.96	.94	5.30	1.25	4.74	1.39	71.099*	<.0125	.637	E–S < S-S*; E–S < S-E*
Item 7	1.55	1.39	4.30	2.11	4.83	1.64	4.632*	<.0125	.501	E–S < S-S*; E–S < S-E*
Item 10	3.45	.9	4.30	.95	3.39	1.03	3.751	.0280	.085	
Item 11	2.43	1.49	2.60	1.58	2.83	1.47	.559	.5740	.014	
AT HL Support; alpha = .0167 (= .05/3)										
Item 5	5.31	1.33	5.30	1.16	1.57	1.34	67.459*	<.0167	.625	S-E < E–S*; S-E < S-S*
Item 8	4.61	1.77	4.60	1.51	1.22	.74	41.227*	<.0167	.504	S-E < E–S*; S-E < S-S*
Item 9	3.27	.92	4.20	.79	3.87	1.14	5.539*	.0060	.120	E–S < S-S*; E–S < S-E*
English Lang, Support; alpha = .005 (= .05/10)										
Item 15	5.08	1.66	3.40	1.84	3.91	1.93	5.962*	.0040	.128	S-S < E–S*; S-E < E–S*
Item 16	2.90	1.79	2.60	2.17	4.35	1.34	6.375*	.0030	.136	E–S < S-E*; S-S < S-E*
Item 17	5.63	1.18	5.50	1.18	5.3	1.06	.627	.5370	.015	
Item 18	4.49	1.33	4.80	1.32	3.7	1.22	3.76	.0270	.085	
Item 19	5.10	1.3	5.10	1.20	3.83	1.27	8.25*	.0010	.169	S-E < E–S*; S-E < S-S*
Item 20	2.90	1.5	2.90	.88	2.91	1.44	.001	.9990		
Item 21	5.88	.91	5.90	.74	6	.74	.155	.8570	.004	
Item 22	3.61	.7	4.10	.74	3.87	.76	2.501	.0880	.058	
Item 24	6.43	.92	6.60	.97	6.35	1.19	.219	.8040	.005	
Item 25	2.35	1.52	2.20	1.55	2.39	1.62	.055	.9470	.001	
Responsive Environment; alpha = .01 (= .05/5)										
Item 2	5.16	1.74	6.30	.95	5.91	1.31	3.309	.0420	.076	
Item 3	3.35	.82	3.70	.48	3.78	.85	2.315	.1050	.055	
Item 12	4.08	1.26	4.30	1.34	4.52	1.34	1.021	.3650	.025	
Item 13	1.53	.83	1.70	1.06	1.52	.79	.194	.8240	.005	
Item 14	2.38	1.71	3.60	2.22	2.39	1.88	1.96	.1480	.047	
Assessment; alpha = .0167 (= .05/3)										
Item 1	4.53	1.16	4.60	.97	4.74	1.18	.268	.7660	.007	
Item 23	6.02	1.33	6.30	1.25	5.74	1.45	.660	.5200	.016	
Item 26	5.92	1.29	6.40	.84	5.87	.97	.804	.4500	.019	

Note. M = Mean; SD = Standard Deviation; LT = Lead Teacher; AT = Assistant Teacher; SD = standard deviation; E—S = English speaking lead teacher with Spanish speaking assistant, S—S = Spanish speaking lead and Spanish speaking assistant; S—E = Spanish speaking lead and English-speaking assistant; *** Bonferroni correction was adopted at the univariate comparison and post-hoc Tukey HSD Test; HL=Home Language; Item 6 is excluded due to missing data on both S—S and SE— groups. Because N/A is used here when teachers speak the HL.

Table 6
Subscale Score Correlation Matrix within CASEBA and with ECERS based on Observed Item Averages.

	CASEBA					ECERS					
	F1	F2	F3	F4	F5	LR	SP	ACT	INT	PS	PCR
S1 (Lead Teacher HL Support)	1					.741	.644	.690	.647	.650	.544
S2 (Assistant Teacher HL Support)	.581	1				.687	.631	.668	.603	.612	.508
S3 (English Language Development)	.613	.651	1			.940	.876	.940	.906	.911	.84
S4 (Responsive Environment)	.605	.537	.789	1		.848	.875	.918	.867	.895	.788
S5 (Assessment)	.461	.519	.703	.669	1	.741	.783	.750	.751	.816	.650

Note. HL=Home Language; LR = Language Reasoning; SP = Space and Furnishings; ACT = Activities; INT = Interaction; PS = Program Structure; PCR = Personal Care Routines.

greater than the correlations within CASEBA, indicating an overlap on the latent dimensions for these measures. As shown in Table 6, each CASEBA subscale is consistently correlated with all ECERS subscales at moderate to high levels, indicating that the two scales were measuring latent structures with overlap. Specifically, CASEBA subscales 1,3,4, and 5 were strongly correlated with ECERS Language Reasoning subscale, ranging from 0.741 to .940; three CASEBA (i.e., S3, S4, and S5) subscales were strongly correlated with three ECERS subscales (Space and Furnishings, Activities, Interactions, and Program), ranging from .750–940; two CASEBA subscale (i.e. S3 and S4) is highly correlated with the ECERS Personal Care Routines subscale, ranging from 0.788 to 0.840. All other elements on the correlation matrix between CASEBA and ECERS subscales were moderate, ranging from 0.544 to .690. The correlations between the two instruments ranged from moderate to strong, and we therefore did not find discriminant evidence for these two instruments (i.e. weak or low correlation); instead, we found indications that the latent structure measured by both instruments may have had a moderate to strong overlap. Given the moderate-to-high correlation among the CASEBA factors, we also tested for multicollinearity, and found that the VIF values of all CASEBA subscales and ECERS subscales alike ranged from 1.163 to 2.133, falling between the range of 1–10 suggesting no violation of multicollinearity.

To understand whether staff language configurations related to ECERS-R scores, the MANOVA test results showed that there was not a significant multivariate main effect for the staff language group on the ECERS-R subscales scores, F(12, 150) = 0.399, p = .962, Wilks' λ = 0.939, partial η2 = .031, indicating a small effect size (Table 4).

3.3. *Associations between classroom DLL characteristics and child outcomes*

Finally, to address the fourth aim of the study we examined relationships between staff language configurations, the CASEBA, and the ECERS-R with PPVT and TVIP scores. After running through all five CASEBA subscales and six ECERS subscales, we found only the second CASEBA subscale, *Assistant Teacher Supports the HL*, was significantly associated with the PPVT at the post test. Neither staff language groups nor ECERS-R had significant associations with either PPVT or TVIP after controlling the same set of child and teacher level background variables. In the pool of control vari-

Table 7
Fixed Effects Estimation through Mixed Modeling.

	PPVT			TVIP		
Parameter	Estimate	Std. Error	Sig.	Estimate	Std. Error	Sig.
Intercept	80.76	12.77	0.00	24.68	12.12	0.04
[age_pre = 3 vs. 4]	−1.65	2.07	0.43	2.94	1.84	0.11
[gender = female vs. male]	−1.59	1.77	0.37	−2.42	1.51	0.11
[Ethnicity = Asian vs. White]	−9.53	8.67	0.27	−0.16	7.34	0.98
[Ethnicity = Black vs. White]	−7.97	8.18	0.33	−2.63	7.18	0.71
[Ethnicity=Hispanic vs. White]	−1.13	7.46	0.88	2.88	6.49	0.66
[Ethnicity = Other vs. White]	3.88	14.55	0.79	2.77	12.55	0.83
[Dominant Language = Spanish vs. English]	−11.52	2.29	0.00***	4.44	1.87	0.02*
[Dominant Language = Both vs. English]	−2.88	3.01	0.34	8.30	2.59	0.00***
[Lead Teacher BA vs MA]	4.81	2.16	0.03*	0.23	1.84	0.90
[Lead Teacher ECE Cert: P-3 vs N-8]	−11.14	8.73	0.20	9.76	7.04	0.17
[Assistant Teacher Education Less than BA vs. MA or higher]	−3.35	4.92	0.50	1.21	4.28	0.78
[Assistant Teacher Education Less than AA vs. MA or higher]	−7.64	4.78	0.11	−1.47	4.12	0.72
[Assistant Teacher Education Less than BA vs. MA or higher]	−6.97	4.06	0.09	1.45	3.59	0.69
[Assistant Teacher ECE Cert: Element N-K vs No]	5.21	6.73	0.44	−1.36	5.72	0.81
[Assistant Teacher ECE Cert: P-3 vs No]	1.54	2.77	0.58	−0.68	2.37	0.78
[staffing = E–S vs. S-E]	5.70	3.81	0.14	−2.22	1.94	0.26
[staffing = S-S vs. S-E]	3.65	6.22	0.56	5.76	3.97	0.15
PRE	0.45	0.05	0.00***	0.46	0.07	0.00***
Lead Teacher Year of Experience	−0.22	0.20	0.28	0.11	0.17	0.51
Assistant Teacher Year of Experience	−0.76	0.41	0.07	0.11	0.35	0.75
CASEBA AT HL Support_Fa09	−2.85	1.32	0.03*	NA	NA	NA

Note. SE = Standard Error; LT = Lead Teacher; AT = Assistant Teacher; E-S = English speaking lead teacher with Spanish speaking assistant, S-S = Spanish speaking lead and Spanish speaking assistant; S-E = Spanish speaking lead and English-speaking assistant; HL=Home Language; *** p-value<.001.

Table A1
Model Fit Comparison through Exploratory Factor Analysis.

	Against	Chi-Square Difference	df	P-Value	RMSEA	CFI	TLI	SRMR
1-factor	2-factor	187.064	25	0	0.146	0.723	0.699	0.101
2-factor	3-factor	164.977	24	0	0.130	0.800	0.763	0.079
3-factor	4-factor	147.326	23	0	0.111	0.867	0.827	0.054
4-factor	5-factor	89.903	22	0	0.087	0.926	0.894	0.038
5-factor	6-factor	51.207	21	0.0002	0.069	0.958	0.934	0.028
6-factor	7-factor	38.929	20	0.0068	0.059	0.973	0.952	0.024

Note. df = degree of freedom; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis Index; SRMR = standardized root mean squared residual.

ables, child dominant language was significantly associated with both PPVT and TVIP at the post test; while the lead teacher education level was significantly associated with only the TVIP at the post test. Table 7 reports the final model for the child outcomes via PPVT and TVIP.

4. Discussion

Through use of a new tool specifically designed to observationally measure quality (CASEBA), this study aimed to understand of the impacts of staff language configurations and whether different configurations yielded higher quality learning environments for young DLLs. Teacher and assistant teacher pairs were defined as English speaking lead teacher with Spanish speaking assistant (E-S), Spanish speaking lead and Spanish speaking assistant (S-S) and Spanish speaking lead and English-speaking assistant (S-E). Classroom quality was assessed using the CASEBA, a tool expressly designed to measure quality of supports offered in English and Spanish under the premise that HL development buttresses English acquisition. In short, this study sought to capture associations of features of classroom quality for DLLs that existing classroom observation tools do not directly assess. Consequently, the study provides insight into the extent to which such tools may be necessary for creating a lens through which to view practices. Also, with findings relative to staff language roles and their association with receptive child language outcomes, the CASEBA shows promising patterns relative to classroom quality for DLLs. Each will be discussed below in the context of our other main study findings.

The present study contributes several key findings relative to staff language groupings and the use of a tool like CASEBA for DLLs. First, when classrooms were looked at through a domain-specific lens, supports for DLLs were more pronounced in classrooms where a lead teacher spoke Spanish. Second, in classrooms where an assistant teacher also spoke Spanish, the classroom quality outcomes as measured by CASEBA exceeded those where only one staff member spoke Spanish. Third, none of the groups scored better than another on the English language support items, indicating a need for specific attention to nurturing English language development in very specific ways. Fourth, through use of the CASEBA's subscale relative to assistant teacher's HL use, we found positive associations with children's English receptive vocabulary scores. Finally, the availability of a tool designed to measure language supports for DLLs that are sensitive enough to capture the kinds of interactions needed to maintain HL and increase English acquisition in linguistically responsive ways is necessary; measures like the ECERS-R do not capture them.

4.1. *Associations between staff language groups relative to CASEBA*

Given the well-documented shortage of Spanish speaking lead teachers, and the potential importance of staff language configuration (E-S, S-S, or S-E), it is noteworthy that there have been no prior preschool studies that have sought to investigate the relationship of staff language configuration to classroom quality and learning for young DLLs. Though studies have found that use of Spanish in classrooms with DLLs has had positive impacts on children, particularly with respect to Spanish language acquisition, none of these studies have explored the quality of language based on who the deliverer of Spanish was in environments that use a lead and assistant teacher for instruction (Barnett et al., 2007; Burchinal et al., 2012; Farver et al., 2009; Gormley, 2008).

In the present study, classrooms with a lead Spanish speaking teacher (S—S, SE—) scored higher on the CASEBA than those in which an English-speaking lead partnered with a Spanish speaking assistant. The largest difference was found in the *Lead Teacher HL Supports* and *Assistant Teacher HL Supports* subscales of the CASEBA. The S—S and SE— groups scored higher on *Lead Teacher HL Supports* subscale, especially for item 4, *The lead teacher uses a home language of the ELL children for instructional purposes* and 7, *The lead teacher uses high quality talk in the students' home language* with a large effect size. The S—E group scored lower than the E—S and SS— group on *Assistant Teacher HL Supports* as reflected by the items of assistant teacher's use of the home language generally (items 5 and 8) with a large effect size. Item 9, *Teaching staff use effective strategies during group instruction to support on-going development of the home language from Assistant Teacher HL Supports* showed that E—S group scored lower with a small effect size, indicating that item 9 is less important than item 5 and 8 in terms of differentiating between staff language groups. This may be due to the way in which the CASEBA is scored given that when no instruction is offered in Spanish, a total of 10 items cannot be scored more than a "1" While assistant teacher interactions or instruction are not discounted, scores maintain low HL instruction is offered. In the largest E—S group, scores were not found to be significantly higher on CASEBA items focused on English acquisition strategies, further asserting the need for supporting teachers on these critical strategies in the classroom about the needs of DLLs to be intentionally supported during English language activities.

In sum, the results imply that the presence of a Spanish speaker, even as assistant teachers, can have modest associations on observed quality. The results further begin to indicate that, with more focused efforts on developing teachers' frequency and sophisticated use of Spanish, that both classroom quality scores and child outcome scores could see an impact. The findings also suggest that while teacher language and cultural match can be critical, they are not a substitute for well-planned and intentional interactions backed by knowledge about how DLLs learn a second language and practices that promote both languages well. In one study, for example, researchers looked at the extent to which linguistically responsive practices were used with DLLs, and how they were associated with teacher-level factors including their own bilingualism. As in the current study, mean scores on items rating linguistically responsive practices were low for all teachers, and while scores were higher for Spanish speaking teachers, the authors described all the scores as "deficient" on the ELLCO-DLL (Sawyer et al., 2016). Their conclusions connected a lack of adequate knowledge focused on teaching DLLs and little administrative support for using Spanish for instruction to the low observation scores. The implications of this are significant, considering what we know about the lack of pre-service teacher preparation with respect to DLLs, and that there are not currently sufficient opportunities for relevant coursework (Maxwell, Lim, & Early, 2006). Further, it validates the need to utilize policy levers that are able to address certification and specialized training requirements for teachers to work with young DLLs so that we can ensure teachers have this specialized knowledge to bolster children's language learning.

While the goal of the current study was not to assess the predictive validity of the CASEBA with respect to child outcomes, it is clear that latent dimensions of the tool are closely linked to staff language inputs. This is particularly clear through our finding of associations between the *Assistant Teacher HL Supports* subscale and PPVT scores. This finding strengthens arguments relative to maintenance of the HL in instruction and lends insights for further investigation of the basis for CASEBA's predictive validity. This finding also adds to the need for the consideration of assistant teachers as valuable assets, particularly when they are the bilingual staff in a classroom of DLLs. Further attention to this subsection of the workforce is warranted and use of the CASEBA is promising to highlight strengths and opportunities for the professional development of this group.

4.2. Associations of staff language groups relative to ECERS-R

No differences were found on the ECERS-R by staff language configurations for any ECERS-R subscale or overall scores. This indicates that the ECERS-R did not discern differences in classroom practice as captured by the CASEBA, and that these specific aspects of classroom practice would go unnoticed in otherwise commonly used measures of "global" quality. This evidence underscores the need for domain specific tools that are sensitive to the practices deemed critical for young DLLs.

4.3. Limitations

There are several limitations. First, the sample sizes are modest and limited to a single district. A larger more diverse sample would provide greater power to test for an association that might have generated a greater range of scores with more variability in practices and, in particular, more scores at the high end of the distribution. While the S—S group did score higher on the CASEBA, it is important to acknowledge that this group included only 10 teachers and that as a group they did not score especially well on the CASEBA compared to what the authors considered "good" to "excellent" practice, again bringing to the forefront the problem of threshold and quality.

Second, the staff language configurations were determined by the school district and were not indicative of actual Spanish proficiency in standardized ways. Anecdotally, teachers expressed feeling uneasy about being asked to instruct in Spanish despite identifying themselves as "bilingual" on paperwork at their time of hire. Teachers felt that while they had marked that they were bilingual, that they were not asked about their level of proficiency, comfort, or willingness to conduct instruction in Spanish. In addition, the district did not adhere to any proficiency testing policies of teachers or assistants at time of hire, nor was this done for the purpose of the study. Implications of this speak to the need for programs to have established and well understood goals if there are expectations for instruction to be carried out in the HL. Further, these goals are critical in the recruitment and hiring of staff.

One other important limitation is that, as described, the lack of ongoing support or guidance from an administrative perspective relative to Spanish use for instruction may have hindered findings relative to the frequency of use of Spanish language for instruction intentionally. While the frequency of instruction in Spanish was not a key question of this study, and some items in the CASEBA do assess the overall amount of time that Spanish was used in an observation period, teacher decisions to use Spanish over English and when cannot be known through a measure of observed quality without direct conversation. Given that that there was no clear policy that designated particular times of the day/activities to be delivered in Spanish, we assume that teachers used their own judgements about when and how much Spanish to use. With more direct guidance, it is possible teachers would have done so more frequently. In future research, qualitative additions that systematically pursue teachers' views of challenges or anxieties regarding the implementation of Spanish for instruction and supports for English acquisition could reveal useful information about what teachers need or want and potentially eliminating stress relative to this issue, particularly as related to staff roles. Further, data relative to teacher planning and rationale for use of Spanish could also be insightful.

This study is unique in its use of the CASEBA, which equally prioritizes the languages used in the classroom. However, the presence of at least one Spanish speaking teaching staff member in the classroom should increase the likelihood that classrooms score higher. We acknowledge that this study provides no insights into what scores would look like in configurations with no Spanish speaker. Further, while the CASEBA shows promise for understanding classroom contexts for DLLs, we acknowledge that additional research with larger samples is needed to establish a more robust validation effort, and also to conduct item response theory (IRT) analyses. In addition, we are aware that the scoring structure of the tool mimics that of the ECERS-R, and that since the CASEBA was designed, this structure has been found to be problematic for the ECERS-R suggesting that there could also be implications for the CASEBA (Fujimoto, Gordon, Peng, & Hofer, 2018; Gordon, Fujimoto, Kaestner, Korenman, & Abner, 2013).

Also important to consider is discussion about the ECERS-3 (Harms, Clifford, & Cryer, 2015), which includes various interactional-based indicators within the original items, which were more exclusively structural in nature, accounting for materials and activities that were available and accessible to children (Harms et al., 2015). In total, two items under the *Language and Literacy* subscale mention accommodations/supports for diverse learners within one indicator each. These include item 12, *Helping children expand vocabulary,* which states, "special accommodations are observed for children to suit their diagnosed disabilities or family language needs" (Harms et al., 2015; p. 37); and number 14, *Staff use of books with children,* where indicator 5.3 states that "accommodations are made for children who require additional support during book time (e.g., Children not fluent in classroom language, with developmental delays, or who do not do well in large groups have special provision, such as smaller group)" (Harms et al., 2015; p. 41). While updates to this version of the tool underscore more process-based quality, more research to investigate whether the revised version of the tool would capture the nuances that the CASEBA did in relation to ECERS-R.

Finally, one very important limitation relative to child outcomes is that the PPVT and TVIP have been suggested by some to be problematic for Spanish speakers as they have been normed on monolingual populations in English and Spanish respectively, which largely differ from those of DLLs in the U.S, and the sample of the present study (Bandel, Atkins-Burnett, Castro, Wulsin, & Putman, 2012; Barrueco, Lopez, Ong, & Lozano, 2012). Some contend that a problem with this lies in the fact that monolingual norming groups may set standards that are too high for DLLs suggesting that their language abilities are underestimated on these tests (Bedore et al., 2012). Though the PPVT and TVIP are widely used in research with DLLs to date, it is fair to acknowledge that they may not be the best tool to use to capture the language learning trajectory as it occurs for young DLLs within the context of the U.S. Still, we acknowledge that while others have found consistent positive associations of the PPVT and TVIP (Brunsek et al., 2017; Burchinal et al., 2016), one caution to keep in mind is that the PPVT and the TVIP used in this study may not be sensitive enough measures (especially for bilinguals) to maximally capture aspects of language development that might be predicted by variations in CASEBA scores. We also hypothesize that, like many other studies investigating the use of classroom quality measures to predict child outcomes, lack of associations between staff configurations and child outcomes could also be that even the highest performing group did not reach a threshold of quality that would be associated with increased child outcomes. As with other measures of classroom quality, this study raises such questions for future work with the CASEBA, though admittedly needing more appropriate child outcome measures (Zaslow et al., 2016).

4.4. Implications

Taken together, the results for each question in this study provides important implications for future research directions and policy conversations. These include the replication of studies already conducted with global measures of quality using a tool like CASEBA to begin to understand the intricacies of teaching that might best increase the achievement of young DLLs in preschool. Studies that explicitly explore the relationship between tools like the CASEBA *and* other dimensions are also important. These include factors like the use of DLL strategies within various curriculum types and intervention models of teacher development with varied objectives. In addition, it would be useful for such studies to examine and inform the content provided in pre-service programs for teachers of DLLs.

Finally, the CASEBA can better inform studies examining how instructional coaches support teachers and assistant teachers of DLLs specifically. In a descriptive analysis of the 130 grantees involved in the Head Start early learning mentor coach initiative, researchers found that only 19.8% of grantees identified improving services for (DLLs) and 6.9% indicated that improving cultural responsiveness were goals for coaching interactions with teachers (Howard et al., 2013). Interestingly, 72% answered to improve CLASS ratings with only 1% of coaches (n = 350) identifying cultural competency as a goal for working with teachers (Howard et al., 2013). Again, our findings show the critical value of what a specific tool can show and how it can potentially inform a continuous improvement cycle between teachers and coaches.

5. Conclusion

Considerations of the CASEBA an assessment of teaching quality for DLLs are timely given increased use of global measures like the ECERS and CLASS for high stakes purposes and accountability in QRIS and Head Start monitoring (National Center on Early Childhood Quality Assurance, 2016). While it is important to acknowledge the impact that global tools have had on the field in raising awareness and regulation of quality in early childhood, it should not be overlooked that these measures have also persistently shown a lack of strong predictive validity on child outcomes (Burchinal, 2018; Hestenes et al., 2015; Setodji, Schaack, & Le, 2018). Consequently, it will be critical to the field to heighten the need for more equitable measures of quality to be considered in QRIS systems as it impacts not only the quality of environments for DLL children, but potentially also to the workforce serving DLLs (Sugarman & Park, 2017). Though a larger sample is needed to fully understand its psychometric properties, this study shows the potential of the CASEBA and the need for the field to shift its perspective of global measures as all-encompassing tools, acknowledging the dangers that can come from relying on them for too many purposes, including high-stakes accountability.

Acknowledgements

The data used in this work were part of a larger research study funded by the Pew Charitable Trust, Washington, DC. The authors of this article are solely responsible for the content thereof, and shall not be deemed to constitute the position of any of the funding sources.

References

Castro, D. C., Ayankoya, B., & Kasprzak, C. (2011). *The new voices: Nuevas voces guide tocultural and linguistic diversity in early childhood*. Paul H. Brookes Publishing Company.

Asparouhov, T., & Muthén, B. (2018). *SRMR in Mplus* Retreived from: http://www.statmodel.com/download/SRMR2.pdf..

Atkins-Burnett, S., Sprachman, S., López, M., Caspe, M., & Fallin, K. (2011). The LanguageInteraction snapshot (LISn): A new observational measure for assessing languageinteractions in linguistically diverse early childhood programs. In C. Howes, J. T. Downer, & R. C. Pianta (Eds.), *Dual language learners in the early childhood classroom* (pp. 117–146). Baltimore, MD: Brookes Publishing.

Castro, D., Espinosa, L., & Paez, M. (2011). Defining and measuring quality in early childhoodpractices that promote dual language learners' development and learning. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle (Eds.), *Quality measurement in early childhood settings* (pp. 257–280). Brooks Publishing.

Bandel, E., Atkins-Burnett, S., Castro, D. C., Wulsin, C. S., & Putman, M. (2012). *Examining the use of language and literacy assessments with young dual languagelearners. Research report #1. Center for Early Care and Education Research-DualLanguage Learners (CECER-DLL). Chapel hill: The University of North Carolina, FrankPorter Graham Child Development Institute*.

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

Barrueco, S., Lopez, M., Ong, C., & Lozano, P. (2012). *Assessing spanish-english bilingual preschoolers: A guide to best approaches and measures*. Baltimore, MD: Brookes Publishing Company.

Barrueco, S., Smith, S., & Stephens, S. A. (2015). *Supporting parent engagement inlinguistically diverse families to promote young children's learning: Implications forearly care and education policy*.

Bedore, L., Pena, E., Summers, C., Boerger, K., Resendiz, M., Greene, K., . . . & Gillam, R. (2012). The measure matters: Language dominance profiles across measuresin Spanish–English bilingual children. *Bilingualism Language and Cognition*, *15*(03), 616–629.

Bentler, P. M. (1990). Comparative fit indexes in structuralmodels. *Psychological Bulletin*, *107*, 238–246, doi: 10.1037/0033-2909.107.2.238.

Bridges, M., & Dagys, N. (2012). *Who will teach our children? Building a qualified early childhood workforce to teach English-language learners*. Berkley's Institute for Human Development: New Journalism on Latino Children.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Burchinal, M. (2018). Measuring early care and education quality. *Child DevelopmentPerspectives*, *12*(1), 3–9.

Burchinal, M., Vandergrift, N., Pianta, R., & Mashburn, A. (2010). Threshold analysis ofassociation between child care quality and child outcomes for low-income children inpre-kindergarten programs. *Early Childhood Research Quarterly*, *25*(2), 166–176.

Burchinal, M., Field, S., Lopez, M., Howes, C., & Pianta, R. (2012). Instruction in Spanish inpre-kindergarten classrooms and child outcomes for English language learners. *EarlyChildhood Research Quarterly*, *27*, 188–197.

Buysse, V., Castro, D. C., West, T., & Skinner, M. (2005). Addressing the need of Latino children: A national survey of state administrators of early childhood programs. *Early Childhood Research Quarterly*, *20*, 146–163.

Buysse, V., Peisner-Feinberg, E., Páez, M., Hammer, C. S., & Knowles, M. (2014). Effects ofearly education programs and practices on the development and learning of dual languagelearners: A review of the literature. *Early Childhood Research Quarterly*, *29*(4), 765–785.

Castro, D. C., Páez, M. M., Dickinson, D. K., & Frede, E. (2011). Promoting language andliteracy in young dual language learners: Research, practice, and policy. *Childdevelopment perspectives*, *5*(1), 15–21.

Castro, D. C. (2005). *Early language and literacy classroom observation- addendum forEnglish language learners*. Chapel Hill, NC: University of North Carolina, FPG ChildDevelopment Institute.

Castro, D. C., Gillanders, C., Franco, X., Bryant, D. M., Zepeda, M., Willoughby, M. T., & Méndez, L. I. (2017). Early education of dual language learners: An efficacy study of theNuestros Niños School Readiness professional development program. *Early ChildhoodResearch Quarterly*, *40*, 188–203.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330.

Cummins, J. (2000). Language, power and pedagogy: Bilingual children in the crossfire. In C. Baker, & Hornberger (Eds.), *Bilingual education and bilingualism series (23)*. Clevedon: Multilingual Matters.

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. In *Working papers on bilingualism, No. 19*. pp. 121–129.

Davison, M. D., Hammer, C., & Lawrence, F. R. (2011). Associations between preschoollanguage and first grade reading outcomes in bilingual children. *Journal ofcommunication disorders*, *44*(4), 444–458.

Dickinson, D. K., & Porsche, M. V. (2011). Relation between language experiences in preschoolclassrooms and children's kindergarten and fourth-grade language and reading abilities. *Child Development*, *82*, 870–886.

Dunn, O. J. (1961). Multiple comparison among means. *Journal of the American Statistical Association*, *56*, 52–64.

Dunn, L., & Dunn, L. (1997). *Peabody picture vocabulary test, third edition (PPVT-111)*. Circle Pines, MN: American Guidance Service.

Dunn, L., Lugo, D., Padilla, E., & Dunn, L. (1986). *Test de vocabulario en imágenes peabody*. Circle Pines, MN: American Guidance Service.

Early, D., Barbarin, O., Bryant, D., Burchinal, M., Chang, F., Clifford, R., . . . & Barnett, W. S. (2005). *Pre-kindergarten in eleven states: NCEDL's multi-state study of pre-kindergarten & study of state-wide early education programs (SWEEP): Preliminary descriptive report* Retrieved at: http://fpg.unc.edu/resources/pre-kindergarten-eleven-states-ncedls-multi-state-study-pre-kindergarten-study-state-wide-.

Espinosa, L. (2010). Assessment of young English language learners. In E. Garcia, & E. Frede (Eds.), *Young English language learners current research and emerging directions for practice and policy* (pp. 119–142). New York, NY: Teachers College Press.

Farrow, J., Wasik, B. A., & Hindman, A. H. (2020). Exploring the unique contributions ofteachers' syntax to preschoolers' and kindergarteners' vocabulary learning. *EarlyChildhood Research Quarterly*, *51*, 178–190.

Farver, J. A. M., Lonigan, C. J., & Eppe, S. (2009). Effective early literacy skill development foryoung Spanish-speaking English language learners: An experimental study of twomethods. *Child Development*, *80*(3), 703–719.

Freedson, M., Figueras, A., & Frede, E. (2009). *Classroom assessment of supports for EmergentBilingual acquisition (CASEBA)*. New Brunswick, NJ: NIEER.

Fry, R., & Taylor, P. (2013). *Hispanic high school graduates pass Whites in rate of college enrollment*. Washington, D.C: Pew Hispanic Center, May. http://www.pewhispanic.org/files/2013/05/PHC_college_enrollment_2013-05.pdf

Fujimoto, K. A., Gordon, R. A., Peng, F., & Hofer, K. G. (2018). Examining the categoryfunctioning of the ECERS-R across eight data sets. *AERA Open*, *4*(1).

Galloway, E. P., & Lesaux, N. (2017). A matter of opportunity: Language and readingdevelopment during early childhood for dual-language learners. In *In the routledge international handbook of early literacy education*. pp. 26–39. Routledge.

Gámez, P. B., Neugebauer, S. R., Coyne, M. D., McCoach, D. B., & Ware, S. (2017). Linguisticand social cues for vocabulary learning in Dual Language Learners and their Englishonly peers. *Early Childhood Research Quarterly*, *40*, 25–37.

Garcia, E. (2018). The classroom language context and English and Spanish vocabularydevelopment among dual language learners attending Head Start. *Early ChildhoodResearch Quarterly*, *43*, 148–157.

Garcia, E. (2012). Language, culture, and early education in the United States. In R. Pianta, W. S. Barnett, L. Justice, & S. Sheridan (Eds.), *Handbook of early childhood education* (pp. 137–157). Guilford Publications.

Garcia, E., & Frede, E. (2010). A policy and research agenda for teaching young English language learners. In E. Garcia, & E. Frede (Eds.), *Young English language learners current research and emerging directions for practice and policy* (pp. 1–9). New York, NY: Teachers College Press.

Garcia, O., & Kleifgan, J. (2010). *Educating emergent bilinguals. Policies, programs andpractices for English language learners*. New York, NY: Teachers College Press.

Garcia, E., Arias, B., Murri, N., & Serna, C. (2010). Developing responsive teachers: A challenge for a demographic reality. *Journal of Teacher Education*, *61*(1-2), 123–142.

Gay, G. (2002). Preparing for culturally responsive teaching. *Journal of Teacher Education*, *52*(2), 106–116.

Gersten, R., & Geva, E. (2003). Teaching reading to early language learners. *EducationalLeadership*, *60*(7), 44–49.

Gillanders, C. (2007). An English-speaking prekindergarten teacher for young Latino children: Implications of the teacher-child relationship on second language learning. *Early Childhood Education Journal*, *33*(1), 47–54.

Gillanders, C., Castro, D. C., & Franco, X. (2014). Learning words for life: Promoting vocabulary in dual language learners. *The Reading Teacher*, *68*(1), 213–221.

Gonzalez, J. E., Pollard-Durodola, S., Simmons, D. C., Taylor, A. B., Davis, M. J., Fogarty, M., . . . & Simmons, L. (2014). Enhancing preschool children's vocabulary: Effects of teachertalk before, during and after shared reading. *Early Childhood Research Quarterly*, *29*(2), 214–226.

Goodrich, J., Lonigan, C., Kleuver, C., & Farver, J. (2015). Development and transfer ofvocabulary knowledge in Spanish-speaking language minority preschool children. *Journal of Child Language*, 1–24.

Gordon, R. A., Fujimoto, K., Kaestner, R., Korenman, S., & Abner, K. (2013). An assessment ofthe validity of the ECERS-R with implications for measures of child care quality andrelations to child development. *Developmental Psychology*, *49*(1), 146.

Gordon, R. A., Hofer, K. G., Fujimoto, K. A., Risk, N., Kaestner, R., & Korenman, S. (2015). Identifying high quality preschool programs: New evidence on the validity of theECERS-R in relation to school readiness goals. *Grantee Submission*, *26*(8), 10861110.

Gormley, W. T. (2008). The effects of Oklahoma's pre-K program on Hispanic children. *SocialScience Quarterly*, *89*(4), 916–936.

Hammer, C. S., Burchinal, M., Hong, S. S., LaForett, D. R., Páez, M., Buysse, V., . . . & López, L. M. (2020). Change in language and literacy knowledge for Spanish–English duallanguage learners at school-entry: Analyses from three studies. *Early ChildhoodResearch Quarterly*, *51*, 81–92.

Hammer, C. S., Hoff, E., Uchikoshi, Y., Gillanders, C., Castro, D. C., & Sandilos, L. E. (2014). The language and literacy development of young dual language learners: A criticalreview. *Early Childhood Research Quarterly*, *29*(4), 715–733.

Harms, T., Clifford, R., & Cryer, D. (2005). *Early childhood environmental rating scale-revised*. New York, NY: Teacher's College Press.

Harms, T., Clifford, R., & Cryer, D. (2015). *Early childhood environmental rating Scale-3*. New York, NY: Teacher's College Press.

Hestenes, L. L., Kintner-Duffy, V., Wang, Y. C., La Paro, K., Mims, S. U., Crosby, D., . . . & Cassidy, D. J. (2015). Comparisons among quality measures in childcare settings:Understanding the use of multiple measures in North Carolina's QRIS and their links tosocial-emotional development in preschool children. *Early Childhood ResearchQuarterly*, *30*, 199–214.

Hindman, A. H., & Wasik, B. A. (2012). Unpacking an effective language and literacy coachingintervention in head start: Following teachers' learning over two years of training. *Theelementary school journal*, *113*(1), 131–154.

Howes, C., Downer, J. T., & Pianta, R. C. (2011). *Dual language learners in the early childhoodclassroom*. Baltimore, MD: Brookes Publishing Company.

Hox, J. J. (1999). A review of current software for handling missing data. *Kwantitatieve Methoden*, *62*, 123–138.

Hulsey, L. K., Aikens, N., Kopack, A., West, J., Moiduddin, E., & Tarullo, L. (2011). *Head start children, families, and programs: Present and past data from FACES. OPRE report 2011-33a*. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health andHuman Services.

Jackson, C. W., Schatschneider, C., & Leacox, L. (2014). Longitudinal analysis of receptivevocabulary growth in young Spanish English–speaking children from migrantfamilies. *Language, Speech, and Hearing Services in Schools*, *45*(1), 40–51.

Jacoby, J. W., & Lesaux, N. K. (2014). Support for extended discourse in teacher talk with linguistically diverse preschoolers. *Early Education and Development*, *25*, 1162–1179.

Jöreskog, K. G. (1969). Efficient estimation in image factor analysis. *Psychometrika*, *34*, 51–75, doi:10.1007/BF02290173.

La Paro, K. M., Thomason, A. C., Lower, J. K., Kintner-Duffy, V. L., & Cassidy, D. J. (2012). Examining the definition and measurement of quality in early childhood education: AReview of studies using the ECERS-R from 2003 to 2010. *Early Childhood Research &Practice*, *14*(1), n1.

Landry, S. H., Assel, M. A., Carlo, M. S., Williams, J. M., Wu, W., & Montroy, J. J. (2019). Theeffect of the Preparing Pequeños small-group cognitive instruction program on academicand concurrent social and behavioral outcomes in young Spanish-speaking dual-languagelearners. *Journal of School Psychology*, *73*, 1–20.

Lipsky, M. G. (2013). Head Start teachers' vocabulary instruction and language complexityduring storybook reading: Predicting vocabulary outcomes of students in linguisticallydiverse classrooms. *Early Education and Development*, *24*(5), 640–667.

Lucas, T., Villegas, A., & Freedson-Gonzalez, M. (2008). Linguistically responsive teachereducation: Preparing classroom teachers to teach English language learners. *Journal ofTeacher Education*, *59*(4), 361–373.

Mashburn, A., Pianta, R., Hamre, C., Downer, J., Barbarin, O., Bryant, D., Burchinal, M., & Early, D. (2008). Measures of classroom quality in prekindergarten and children'sdevelopment of academic, language and social skills. *Child Development*, *79*(3), 732749.

Maskey, R., Fei, J., & Nguyen, H. O. (2018). Use of exploratory factor analysis in maritime research. *The Asian Journal of Shipping and Logistics*, *34*(2), 91–111.

Maxwell, K., Lim, C., & Early, D. (2006). *Early childhood teacher preparation programs in the United States: National report*. Chapel Hill: NC: The University of North Carolina, FPG Child Development Institute.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *The Journal of Applied Psychology*, *93*(3), 568.

Méndez, L. I., Crais, E. R., Castro, D. C., & Kainz, K. (2015). A culturally and linguistically responsive vocabulary approach for young Latino dual language learners. *Journal of Speech Language and Hearing Research*, *58*(1), 93–106.

Miles, J., & Shevlin, M. (2001). *Applying regression and correlation: A guide for students andResearchers*. London: Sage.

Morrow, L. M., & Schickedanz, J. (2006). The relationships between sociodramatic play andliteracy development. *Handbook of early literacy research*, *2*, 269–280.

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus user's guide* (seventh edition). LosAngeles, CA: Author.

Muthén, L. K., & Muthén, B. O. (1998-2019). *Mplus (version 8.2) [Computer software]*. LosAngeles, CA: Author.

National Center on Early Childhood Quality Assurance. (2016). *QRIS compendium 2016 FactSheets: Use of observational tools in QRIS* Retrieved from:https://childcareta.acf.hhs.gov/sites/default/files/public/qris_observational_tools_2016.pf.

Nores, M., & Barnett, W. S. (2014). *Access to high quality early care and education: Readinessand opportunity gaps in America. National Institute for Early Education and Center onEnhancing early learning policy report. New Brunswick, NJ: Center on enhancing EarlyLearning outcomes. Accessed February, 13, 2015*.

Peisner-Feinberg, E., Buysse, V., Fuligni, A., Burchinal, M., Espinosa, L., Halle, T., . . . & Castro, D. (2014). Using early care and education quality measures with dual language learners:A review of the research. *Early Childhood Research Quarterly*, *29*, 786–803.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system:Pre-K (CLASS Pre-K)*. Baltimore, MD: Paul H. Brookes.

Piker, R. A., & Rex, L. A. (2008). Influences of teacher–child social interactions on Englishlanguage development in a Head Start classroom. *Early Childhood EducationJournal*, *36*(2), 187–193.

Pollard-Durodola, S. D., Gonzalez, J. E., Saenz, L., Soares, D., Resendez, N., Kwok, O., . . . & Zhu, L. (2016). The effects of content-related shared book reading on the languagedevelopment of preschool dual language learners. *Early Childhood ResearchQuarterly*, *36*, 106–121.

Raikes, H. H., White, L., Green, S., Burchinal, M., Kainz, K., Horm, D., . . . & Esteraich, J. (2019). Use of the home languagein preschool classrooms and first-and second-language development among duallanguage learners. *Early Childhood Research Quarterly*, *47*, 145–158.

Reilly, S. E., Johnson, A. D., Luk, G., & Partika, A. (2019). Head Start Classroom Features andLanguage and Literacy Growth among Children with Diverse Language Backgrounds. *Early Education and Development*, 1–22.

Restrepo, M., Castilla, A., Schwanenflugel, P., Neuharth-Pritchett, S., Hamilton, C., & Arboleda, A. (2010). Effects of supplemental Spanish oral language program on sentence length, complexity, and grammaticality in Spanish-speaking children attending English-only preschools. *Language, Speech, and Hearing Services in Schools*, *41*, 3–13.

Roberts, T. A., Vadasy, P. F., & Sanders, E. A. (2018). Preschoolers' alphabet learning: Lettername and sound instruction, cognitive processes, and English proficiency. *Early Childhood Research Quarterly*, *44*, 257–274.

Ryan, A. (2007). Two tests of the effectiveness of bilingual education in preschool. *Journal of Research in Childhood Education*, *21*(4), 352–363.

Ryan, S., & Whitebook, M. (2012). More than teachers: The early care and educationworkforce. In B. Pianta (Ed.), *Handbook of early education* (pp. 92–110). New York: Guilford Press.

Sawyer, B., Atkins-Burnett, S., Sandilos, L., Scheffner Hammer, C., Lopez, L., & Blair, C. (2018). Variations in classroom language environments of preschool children who AreLow income and linguistically diverse. *Early Education and Development*, *29*(3), 398416.

Sawyer, B. E., Hammer, C. S., Cycyk, L. M., López, L., Blair, C., Sandilos, L., . . . & Komaroff, E. (2016). Preschool teachers' language and literacy practices with dual language learners. *Bilingual Research Journal*, *39*(1), 35–49.

Serafini, E. J., Rozell, N., & Winsler, A. (2020). Academic and English language outcomes forDLLs as a function of school bilingual education model: The role of two-way immersionand home language support. *International Journal of Bilingual Education and Bilingualism*, 1–19.

Setodji, C. M., Schaack, D., & Le, V. N. (2018). Using the early childhood environment ratingscale-Revised in high stakes contexts: Does evidence warrant the practice? *EarlyChildhood Research Quarterly*, *42*, 158–169.

Shivers, E. M., & Sanders, K. (2011). Measuring culturally responsive early care and education. In M. J. Zaslow (Ed.), *Quality measurement in early childhood settings* (pp. 191–225). Baltimore, MD: Paul H Brookes.

Spencer, T. D., Moran, M., Thompson, M. S., Petersen, D. B., & Restrepo, M. A. (2020). EarlyEfficacy of Multitiered Dual-Language Instruction: Promoting Preschoolers' Spanish andEnglish Oral Language. *AERA Open*, *6*(1), Article 2332858419897886.

Sugarman, J., & Park, M. (2017). *Quality for whom? Supporting diverse children and workersin early childhood quality rating and improvement systems*. Washington, D.C: Migration Policy Institute.

Tabors, P. (2008). *One child, two languages* (2nd ed). Baltimore, MD: Brooks Publishing.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1–10. http://dx.doi.org/10.1007/BF02291170

Valentino, R. (2015). *High quality and effective instruction for young children: Variation bysocioeconomic status, race, and language status. (Unpublished doctoraldissertation). Stanford University, Palo Alto, CA*.

Valentino, R. (2018). Will public pre-K really close achievement gaps? Gaps in prekindergartenquality between students and across states. *American Educational Research Journal*, *55*(1), 79–116.

Vitiello, V. (2013). *Rejoinder to Teachstone's "Dual language learners and the CLASS measure." campaign for quality early education (CQEE) coalition* Retrieved from: http://www.afabc.org/getattachment/85a2c0a9-31d7-444e-9ba25034f07f84a0/CQEE_Rejoinder.aspx.

Vitiello, V., Downer, J., & Williford, A. (2011). Preschool classroom experiences of dual language learners: Summary of findings from publicly funded programs in 11 states. Dual language learners in the early childhood classroom. In C. Howes, J. Downer, & R. C. Pianta (Eds.), *Dual language learners in the early childhood classroom* (pp. 69–91). Baltimore, MD: Paul H. Brookes.

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using mplus*. Chichester, WS: John Wiley & Sons.

Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children'smathematics, language, literacy, executive function, and emotionalskills. *ChildDevelopment*, *84*(6), 2112–2130.

Whitebook, M., McLean, C., Austin, L. J., & Edwards, B. (2018). *Early childhood WorkforceIndex 2018. Center for the study of child care employment, university of California at berkeley*.

Yu, C. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models withbinary and continuous outcomes (Doctoral dissertation). Retrieved from ProQuestDissertations & Theses Full text database*.

Zaslow, M., Anderson, R., Redd, Z., Wessel, J., Daneri, P., Green, K., Cavadel, E., Tarullo, L., Burchinal, & Martinez-Beck, I. (2016). I. Quality thresholds, features, and dosage in early care and education: Introduction and literature review. *Monographs of the Society for Research in Child Development*, *81*(2), 7–26.

Zepeda, M., Castro, D., & Cronin, S. (2011). Preparing early childhood teachers to work withyoung English language learners. *Child Development Perspectives*, *5*(1), 10–14.