Running Header: The Houston Questionnaire

Can Bilingual Children Self-Report their Bilingual Experience and Proficiency? The Houston Questionnaire

Anny Castilla-Earls
University of Houston
Department of Communication Disorders and Sciences
Houston, TX

Juliana Ronderos
University of Houston
Department of Communication Disorders and Sciences
Houston, TX

Lisa Fitton
University of South Carolina
Department of Communication Sciences and Disorders
Columbia, SC

Journal-formatted version of this manuscript: https://doi.org/10.1044/2022_JSLHR-21-00675
Conflict of Interest: There are no conflicts of interest.
Corresponding Author: Anny Castilla-Earls, University of Houston, Melcher Life Sciences 3871 Holman St. Room M242; Houston, Tx, 77204. Phone: (713) 743-0488. Email: annycastilla@uh.edu.
Funding: Research reported in this publication was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under Award Number K23DC015835 granted to Anny Castilla-Earls. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Abstract

Purpose: To develop a child self-report questionnaire measuring bilingual experience and self-perceptions of Spanish and English proficiency and establish preliminary evidence of validity and reliability for the questionnaire. **Method:** Participants included 113 Spanish-English bilingual children with and without developmental language disorders ranging in age from 4 to 8 years. All children completed the questionnaire in Spanish and participated in behavioral assessment of their language skills in both Spanish and English.

Results: Using confirmatory factor analysis, a model with three correlated factors (Self-Perception of Proficiency in Spanish, Self-Perception of Proficiency in English, and Bilingual Experience) emerged with the best global fit, reasonableness, consistency with theory, and model parsimony, suggesting that the questionnaire has good internal reliability. The scaled results of the questionnaires significantly correlated with behavioral measures of both Spanish and English, supporting the convergent validity of the measure.

Conclusion: The Houston Questionnaire is an assessment tool for the assessment of bilingual experience and self-perception of proficiency in Spanish and English bilingual children between the ages of 4 and 8 years. The results provide foundational evidence supporting the reliability and convergent validity of this tool.

Keywords: Bilingualism, assessment, self-report, child language

For the journal-formatted final version of this manuscript, see:
https://doi.org/10.1044/2022_JSLHR-21-00675

Can Bilingual Children Self-Report their Bilingual Experience and Proficiency? The Houston Questionnaire

Bilingual children represent a heterogeneous group of children who vary in their bilingual experiences and proficiency in each language (e.g., Bedore et al., 2010; Kapantzoglou et al., 2015). This variation poses a significant challenge for the identification of language disorders in bilingual children because speech-language pathologists must differentiate typical variations in bilingual experience (e.g., children with less exposure to a language resulting in lower proficiency in that language) from language ability limitations (e.g., language learning difficulties; Arias & Friberg, 2017; Bedore & Peña, 2008). Therefore, it is critical to gather information about the child's experiences in both languages during the bilingual assessment process to better understand the potential impact of exposure and use on bilingual language skills (Castilla-Earls et al., 2020; Kohnert, 2010).

Parents and teachers often serve as sources of information regarding the child's bilingual experiences (e.g., Restrepo, 1998; Rojas et al., 2016). However, parents might be better at estimating their child's abilities and experiences in the home language in comparison to the school language. Parents in immigrant families may not speak the school language (National Kids Count, 2020). Further, most parents do not have the opportunity to observe the child at school, making it difficult to rate their school language use appropriately (Bedore et al., 2011). Similarly, teachers might be limited in their ability to estimate children's language exposure and use outside the school environment (Vagh et al., 2009). From this perspective, children themselves might be better observers and reporters of their bilingual experience and knowledge of each language than either parents or teachers. We developed The Houston Questionnaire (Houston-Q) to gather information about bilingual experience and proficiency in Spanish and English from the child's perspective.

Self-Reporting of Bilingual Skills in Bilingual Children

To develop a self-report measure of bilingual experience and proficiency, it is crucial to first consider whether children have enough language awareness to express differences between Spanish and English proficiency and experiences in each language. Language awareness is a metalinguistic skill that requires the ability to reflect on one's own language (Svalberg, 2007). Specifically for bilingual children, language awareness includes the ability to reflect on both of their languages (Adesope et al., 2010). Language awareness in bilingual children develops as early as age two. For example, two-year-old bilingual children can name their languages and identify what language is being used by themselves and others (De Houwer, 2017).

Researchers examining language awareness in bilingual children have used various data collection tools, including drawing and coloring language activities (e.g., color a child silhouette following the languages spoken; Martin, 2012; Melo-Pfeifer, 2015; Rojo & Echols, 2017), interviews (open questions about their bilingual experience that allow elaboration in responses; Pérez-Leroux et al., 2011), and language questionnaires (Babino & Stewart, 2016; Rojo & Echols, 2017). For example, Babino and Stewart (2016) used a 4-point Likert scale and multiple-choice questions to examine cultural identity, language attitudes, and language use in and outside the school. Language questionnaires emerged as an appropriate instrument, having been used with bilingual children as young as four years of age (Rojo & Echols, 2017). In addition, questionnaires allow for a variety of question types to elicit theoretical and practice-driven information about language use, including yes/no questions (e.g., Do you use Spanish with your teacher?), short open questions (e.g., Tell me a family member who lives in your house. What language do you speak with him/her?), and quantifiable questions (e.g., How many friends do

you have who speak Spanish?). Questionnaires also can include visual aids to support more reliable responses to quantitative questions. Therefore, a questionnaire appeared to be an appropriate measurement tool for children to self-report their bilingual experience and proficiency in each language.

Importantly for this study, the accuracy of children's judgments of their bilingual experience and proficiency has been largely unexplored. Previous studies investigating bilingual children's language awareness have primarily provided descriptive information about children's responses to the questionnaires (e.g., Babino & Stewart, 2016; Rojo & Echols, 2017). For a child self-report questionnaire of bilingual experience and proficiency to be practically useful, it is crucial to examine the descriptive information elicited by the tool and if children can respond in a reliable and valid manner to the questionnaire. In this study, we aim to examine evidence of the internal reliability and convergent validity of children's responses to the Houston-Q, a self-report questionnaire designed to quantify children's bilingual experience and proficiency in each language.

The Development of the Houston Questionnaire

The Houston-Q was designed to gather information about children's self-assessment of their language proficiency in both Spanish and English, and the child's perceptions of their bilingual experience. Other validated self-report measures exist for children to self-report similar constructs (e.g., health-related quality life, stress, and psychological dysfunction; Pagano et al., 2000; Solans et al., 2008; Osika et al., 2007). In bilingual adults, self-report studies show that self-report measures of proficiency can be valid measurement instruments (e.g., LEAP-Q; Marian et al., 2007). However, in some instances, mismatches between the classification of the adult's self-report of bilingual profile (Spanish dominant, English dominant, or Balanced) and

the adult's bilingual profile calculated from behavioral language measures have been reported (Gollan et al., 2021; Tomoschuk et al., 2019). In this study, we focus on Spanish-English bilingual children because they represent the largest bilingual population in the U.S., yet they continue to be disproportionally represented in special education programs (Artiles et al., 2002; Samson & Lesaux, 2009). Better understanding children's self-reported bilingual experience and proficiency in each language may complement clinical assessment practices by facilitating identification of children's baseline language experiences and strongest language prior to direct comprehensive language assessment. A reliable indication of the child's strongest language would be clinically meaningful in potentially reducing the time needed to problem-solve during the bilingual evaluation process, particularly in the context of screening for language disorders. Considering the child's abilities in their self-perceived strongest language may contribute to more accurate identification of language disorders.

Bilingual Experience

It is generally understood that exposure to a language is a prerequisite for language learning and proficiency (e.g., Bohman et al., 2010; Hoff & Core, 2013). That is, for children to learn a language, they need to be exposed to it. However, there is no agreement in the literature about the amount and quality of the input needed for language learning (for a detailed review of the methodological considerations regarding language input, see Carroll, 2017). For bilingual children, language experiences are partitioned between two languages, in contrast with monolingual children whose language input is completely in one language (Bridges & Hoff, 2012; Peña et al., 2018).

The amount of exposure a bilingual child has in each language robustly predicts their rate of growth and proficiency in each respective language (e.g., Hammer et al., 2012; Hoff et al., 2018; Peña et al., 2018). However, it is important to note that, in the U.S., English growth predominates even among children who have high exposure to Spanish since exposure to English tends to be greater outside the home, and Spanish exposure is likely to be limited to the home (Hoff, 2017). On the other hand, Spanish exposure is necessary, although not sufficient, for the development and maintenance of Spanish language skills of bilingual speakers, perhaps due to the lower social status of Spanish in the U.S. (Castilla-Earls et al., 2019; Duursma et al., 2007). Therefore, it is important to estimate how input is partitioned between languages to estimate current exposure and potential future growth in each language.

An important part of the bilingual experience for children is the language(s) spoken at home and its impact on language growth (De Houwer, 2004). For example, when both parents speak Spanish at home, children tend to have higher vocabulary in Spanish than in English, but when both parents speak English at home, English vocabulary tends to be higher than Spanish vocabulary (Place & Hoff, 2011). Siblings also play a role in the bilingual experience at home. For instance, homes with older school-age siblings tend to use more English than homes without an older sibling (Bridges & Hoff, 2012; Obied, 2009). Interestingly, when bilingual college students reflect on their experiences learning Spanish and English, they often attribute their parents and grandparent's encouragement to use Spanish as an important contributor to their current Spanish skills, while the use of English with siblings was considered a contributor to their English skills (Castilla-Earls & Fulcher-Rood, unpublished).

The language(s) used at school also predicts language growth for children. For many Spanish-English speaking children in the U.S., the start of formal education instigates a

significant shift in language proficiency from Spanish, the language spoken at home, to English, the language spoken in most schools (Lutz, 2008). Children who attend bilingual education schools tend to maintain Spanish language skills better than children who attend schools with English-only instruction (Castilla-Earls et al., 2019; Farver et al., 2009). However, by 5th grade, native Spanish-speaking children in bilingual education programs report that they prefer to use English for both social and academic purposes (Babino & Stewart, 2016).

Language Ability and Language Proficiency

During the development of the Houston-Q, we aimed to capture the child's self-assessment of language proficiency rather than language ability. In this study, language ability refers to the child's general language learning capacity that interacts with language input (Peña et al., 2018). Children with low language ability not explained by associated neurological disorders are identified as children with developmental language disorders (DLD; Leonard, 2014; Bishop et al., 2016). These children have low language ability even when input is present (Kan & Windsor, 2010; Peña et al., 2014). Language ability is traditionally measured with standardized language tests or spontaneous language measures (e.g., Peña et al., 2018; Restrepo, 1998). In bilingual children, language ability is determined using the child's strongest language to differentiate children whose language performance on a test or assessment task represent a lack of input in a language (i.e., second language learners) from children who show low performance in both languages (i.e., children with language disorders) (Kohnert, 2010; Peña et al. 2018).

Language proficiency refers to the specific knowledge of a language that is mediated by the child's language ability. Regardless of whether a bilingual child has typical language ability or low language ability, they will vary in their knowledge of each language. For example, a child with low language ability may have more knowledge of Spanish than English, more knowledge

of English than Spanish, or have about the same level of knowledge of both languages. In the same way, a child with typical language skills can vary in their bilingual profiles. However, how much knowledge children with low language ability have in each language will differ from the knowledge children with typical language ability have in their languages. That is, children with low language ability that have about the same level of knowledge in both languages would score lower on behavioral language assessments in comparison to children with typical language ability who also have similar levels of knowledge in both languages. Therefore, there are at least two levels of comparison[1]. At one level, there is a between-child comparison of how much language a child can learn provided input compared to their peers (language ability). At a second level, there is a within-child comparison of how much knowledge a child has in a given language compared to their other language (language proficiency). For the development of the Houston-Q, we focused on this second level of comparison. We suggest that bilingual children can self-report their proficiency in each language because they are aware of their two languages and can use their awareness to respond to questions that yield to a proficiency or experience measure. However, we do not suggest or expect that bilingual children would be able to self-report their language ability (i.e., if they have a language disorder or typical language skills) because this is a higher-level metalinguistic skill that requires a comparison between children.

Measurement Reliability and Validity

[1] There might be other levels of comparison, which are not the focus of this investigation. For example, a within child comparison of type of language skills, such as morphology and semantics (Bedore et al., 2012).

A core component of scale development is the evaluation of the scale's psychometric properties, such as reliability and validity. This evaluation is an inherently ongoing process that requires iterative examination of characteristics of the scale and how it functions for different individuals in different contexts (see Boateng et al., 2018). In the present work, we focus on the initial steps of psychometric evaluation, including examination of the developed measure's dimensionality, the overall scale and subscale internal consistency reliability, and preliminary convergent validity. These foundational properties directly influence the scoring structure and interpretation of individual responses to a measure (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) and correspondingly represent a first step in establishing the practical utility of the Houston-Q.

Dimensionality assessment encompasses the identification of any potential subscales or subtests within the overall measure. It is essential to establish the dimensionality of a measure prior to evaluating its reliability because each unique dimension must be scored separately. Scoring multiple dimensions together can lead to inaccurate estimates of item characteristics and, ultimately, individual performance (de Ayala, 2013; DeMars, 2012; McNeish & Wolf, 2020). Once subscales are identified, these can then be evaluated for evidence of internal consistency reliability, which is the consistency within the test items included within each subscale. For a subscale score to be meaningful, each test item included in that subscale should function in a relatively similar manner. Internal consistency reliability is generally evaluated by examining Cronbach's alpha or Coefficient omega in the case where some items contribute more strongly to the total subscale score than others (i.e., violations of the assumption of tau equivalence; McNeish, 2018).

Upon establishment of scale dimensionality and internal consistency reliability, evidence of validity may be examined. Although there are many forms of validity, we focus on the assessment of concurrent criterion validity, specifically convergent validity. Concurrent criterion validity refers to how closely the scale and/or subscales are associated with scores obtained from external measures administered to the same participants at approximately the same time (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Convergent validity may be evaluated empirically through the examination of correlations between participants' scores on the developed scale and their scores on other measures that are theoretically considered to be related. Examination of these psychometric properties goes beyond a descriptive approach to the responses provided by children (e.g., Martin, 2012; Melo-Pfeifer, 2015; Rojo & Echols, 2017) and instead targets the quality of the measurement.

This study

Previous research suggested that bilingual children as young as four years of age might have enough language awareness to self-report their bilingual experiences and proficiency in each language (Rojo & Echols, 2017). However, information about children's bilingual experience and proficiency is currently collected primarily through parents and teachers. In this study, we explore the possibility that children can provide a valid and reliable estimation of their bilingual experiences and proficiency using a questionnaire administered verbally in Spanish by an adult. We developed the Houston-Q as a tool to estimate variations in bilingual experience and proficiency during the bilingual assessment process. Our research questions are: (a) what is the dimensionality of the Houston-Q? (b)Is the Houston-Q a reliable tool for the self-report of

bilingual experience and proficiency in Spanish and English in bilingual children? and (c) Is there evidence of convergent validity between behavioral measures of language skills in Spanish and English and the Houston-Q?

Method

Validation Participant Sample

The Institutional Review Board at the University of Houston approved this study. Parents provided written informed consent, and children provided verbal assent to participate in the sessions. Participants were recruited from school districts and speech-language clinics in the Greater Houston area as part of a broader longitudinal study of bilingual language development. To be eligible for the study, children spoke and understood both Spanish and English, passed an otoacoustic emission hearing screening, and obtained a score greater than 70 on the Matrices subtest of the *Kaufman Brief Intelligence Test—Second Edition* (KBIT-2; Kaufman & Kaufman, 2004) as a measure of non-verbal IQ[2].

The validation sample for the current study included 113 Spanish-English bilingual children ranging in age from 3 years, 11 months to 8 years, 2 months (M = 71.05, SD = 12.46 in months). The sample was 43% girls (n = 49). Approximately 54% of the children came from families where the mother had not attended college, and 70% of the children qualified for free or reduced-price lunch as reported via parental questionnaire. Parents also reported that their families spoke either Spanish only (49%) or both Spanish and English at home (39%). The remaining 12% of the parents reported their children spoke English only at home and Spanish at

[2] There was one instance of a child with a score below 70 on the KBIT-2 but with all scores on language assessments within normal limits. It appears that the KBIT-2 score was not indicative of the child's actual abilities. For this reason, we ran all analysis twice: a) excluding this child, and b) including this child. We found no differences in the results of this study. Therefore, we included this child in the reported sample.

school. Regarding the language of instruction at school, 90% of the children in our sample attended bilingual Spanish-English or Spanish language immersion education programs. Further information about the children's language skills will be presented as descriptive information in the results section.

Measures

The Houston Questionnaire

The Houston-Q was developed to provide a self-assessment measure for children regarding their language proficiency and experiences in each language. The questionnaire was constructed to be completed in approximately 10 minutes with children as young as four years old. For this reason, we designed questions with simple wording and vocabulary and used visual support when needed. In addition, all questions were designed to be verbally presented in Spanish by an examiner who recorded the child's responses. Questions included yes/no questions, short open questions, and questions with Likert-scale options to obtain quantifiable information. Some questions required a combination of yes/no responses followed by a 5-point Likert-scale question (e.g., 1- *a little* to 5- *a lot; 1- few* to *5- many*). To support children, we used pictures with different amounts of candy to indicate a little or few (one piece of candy) to a lot or many (five pieces of candy). Other questions asked about home and school activities and the language in which they occurred (Spanish, English, both, or not performed at all).

The questionnaire was designed to target three main areas of children's language: self-assessment of Spanish proficiency, self-assessment of English proficiency, and bilingual experience in Spanish and English. It consists of 25 questions in total. The section for self-assessment of proficiency in the languages in the Houston-Q includes questions regarding how good children are at speaking a language, how easy they perceive the language to be, and how

many friends they have who speak the languages. These questions are a combination of a yes/no question (e.g., Are you good at speaking Spanish? Do you think Spanish is easy? Do you have friends who speak only Spanish?) followed by a 5-point Likert-scale question (e.g., If you are good, how good? If it's easy, how easy? If you have friends who speak that language, how many?). On the 5-point Likert scale follow-up questions, lower values indicated lower proficiency, and higher values indicated higher proficiency. Other items in the proficiency section of the Houston-Q included questions regarding how much Spanish and English children heard during the day, which were also 5-point Likert scale questions with lower values indicating lower frequency and higher values indicating higher frequency. To estimate bilingual experience, questions listed a variety of activities (e.g., read books, watch TV, play at the park, etc.) and children were provided four options regarding the language they used during these activities (e.g., I do this in Spanish, I do this in English, I do this in both Spanish and English, and I don't do this). A final set of questions prompted children to name three people from their family and identify what language they used with each of them. Children were provided with three options to respond: Spanish, English, or both Spanish and English.

Behavioral language measures

Receptive Vocabulary. We used the standard scores from the *Peabody Picture Vocabulary Test—Fourth Edition* (PPVT-4; Dunn & Dunn, 2007) and the *Test de Vocabulario en Imágenes Peabody* (TVIP; Dunn et al., 1986) as measures of receptive vocabulary in English and Spanish, respectively. The PPVT-4 is a standardized measure of receptive vocabulary for use with individuals ages 2-90 years old. This assessment has been normed with English monolinguals from across the United States and has been frequently used in research studies with

children as a measure of vocabulary. The TVIP is a parallel measure to the PPVT and assesses receptive vocabulary in Spanish in individuals ages 2-18. The TVIP has been normed with Spanish monolingual speakers in Mexico and Puerto Rico. In both assessments, children are presented with stimulus pages consisting of four pictures. The examiner provides a vocabulary word to the child, and the child responds by either pointing or stating the number for the picture they believe best represents the word. It is important to note that both of these tools were normed with monolingual children and therefore are not ideal for measuring receptive vocabulary abilities in bilingual children (Wood et al., 2018).

Morphosyntax. We used the morphosyntax subtests of the *Bilingual English-Spanish Assessment* (BESA; Peña et al., 2018) and the *Bilingual English-Spanish Assessment—Middle Extension* (BESA-ME; Peña et al., 2008, 2016). The BESA is a standardized test designed to evaluate the language abilities of Spanish-English bilingual children ages 4;0-6;11 (years; months) in the U.S. The BESA-ME is an experimental measure, similar to the BESA, to assess language skills of Spanish-English bilingual children ages 7-9;11 (years; months). The BESA (and BESA-ME) was used in this study to estimate language ability because it is currently the gold standard normed-reference measure for identification of Spanish-English bilingual children with developmental language disorders in the United States. The morphosyntax subtest of both tests consists of a cloze item section and a sentence repetition section targeting complex grammatical structures in each language. Standard scores (M = 100, SD = 15) are calculated for each language. The BESA and BESA-ME morphosyntax subtests can be administered as stand-alone subtests with good diagnostic accuracy to identify bilingual children with developmental language disorders (Peña et al., 2008, 2016, 2018). In order to combine BESA and BESA-ME morphosyntax, we used standard scores. The BESA morphosyntax subtests standard scores range

from 52-145. However, the BESA-ME experimental version standard scores did not have a specific range at this time. For purposes of the analyses in this study, we mirrored the range on the BESA-ME to the one used for the BESA so that the lowest possible score on the BESA-ME was also 52[3]. We used the best language score as a measure of language ability, as suggested in the BESA and BESA-ME testing manuals, following current best practices for the assessment of bilingual children (Kohnert, 2010; Peña et al., 2018).

Sentence Repetition. We also used the scaled scores of the Recalling Sentences subtest (Recordando Oraciones in the Spanish version) in the latest versions of the *Clinical Evaluation of Language Fundamentals* in English and Spanish (CELF-5 for English, Wiig et al., 2013; and CELF-4 for Spanish, Semel et al., 2006). In these subtests, children are asked to repeat the sentence after the evaluator. The subtest is designed to evaluate the child's knowledge of the language structure and vocabulary in addition to cognitive processing skills such as verbal working memory (Pratt et al., 2020). Because this task assesses the knowledge of the language (i.e., to be able to repeat a sentence, one needs to have the language structure and vocabulary in that language), sentence repetition tasks might be considered biased for the assessment of language ability in bilingual children if only one language is used (Armon-Lotem & Meir, 2016). Sentence repetition tasks have been found to have high sensitivity and specificity for the diagnosis of developmental language disorder (Archibald & Joanisse, 2009; Rujas et al., 2021).

Procedures

Parents provided consent for their children's participation in the study and completed a questionnaire about demographics and the use of Spanish and English. Children provided assent to participate. Children completed the behavioral language tasks and the Houston-Q as part of a

[3] Sensitivity analyses were performed to assess the potential impact of the truncated scores on correlational results. No substantial differences were noted, so only the results from the truncated scores are reported.

larger battery of assessments. The Spanish language tasks were part of the Spanish language skills session, and the English language tasks were part of the English language skills session. Each of these sessions was approximately 50 minutes long. Task order in each session varied across participants. All the tasks were administered in person and scored by a trained research assistant who was a native speaker of the target language.

The Houston-Q was administered in Spanish as part of the Spanish language skills session. Children were first shown pictures with different amounts of candy to indicate a little (one piece of candy) to a lot (five pieces of candy). The examiner said in Spanish, “I know you speak both Spanish and English; I am going to ask you some questions about Spanish and English. For some questions, you can answer a little, like one piece of candy; for others, you can answer a lot, like five pieces of candy. For some questions, you may want to answer something in between, like two, three, or four pieces of candy.” The examiner gauged the child’s understanding of the task by asking questions to ensure that the child understood what was expected (e.g., Do you have any questions? Do you understand what we are doing?). Once the examiner felt that the child understood the task, they would start asking the questions in the Houston Questionnaire. The examiner monitored whether the child answered each question in a manner aligned with the intended content of the question to ensure understanding of the task. Repetition of the instructions was allowed. The examiner wrote down all answers from the child in the questionnaire response form. Although the questionnaire was administered in Spanish only, responses were allowed in Spanish or English. All children in this study were able to complete this task using this procedure. There were no reports of no compliance or difficulties understanding the task.

Analytic Approach

Children's responses were first examined for frequencies of each response (see Supplementary Figures 1-3). Item responses were evaluated for evidence of restriction of range (i.e., floor or ceiling effects), which would limit information extractable from any given item, based on a criterion of 95% for any specific response. No items met this criterion. Correspondingly, all items were included in subsequent analyses.

Dimensionality and Reliability

We used confirmatory item-level factor analysis to assess the dimensionality, or underlying factor structure, of the scale. An inherent strength of this analytic approach is that it allows for the evaluation of the characteristics of individual questionnaire items by partitioning out different sources of variability in children's responses. Item-based confirmatory factor analysis yields separate estimates for individual item characteristics (e.g., difficulty, discrimination) and individual participant characteristics (e.g., self-perception of Spanish proficiency). This analysis is useful for supporting the development of a generalizable scale. However, the robustness of the specific item parameters is limited by the representativeness of the participant sample compared to the local population.

We based all model testing on a priori hypotheses of possible constructs underlying the items. The most complex model assessed included six possible underlying factors (see Figure 1, Model A), and the most parsimonious included three underlying factors (Figure 1, Model B), in alignment with the construction of the scale. All factors were correlated, consistent with the theoretical framing that general language learning abilities contribute to the development of proficiency in both languages. Models were estimated using unweighted least squares means and variance (ULSMV) in Mplus Version 8.4 (Muthén & Muthén, 2019). Item intercepts, factor

loadings, and residual variances were freely estimated, with latent factor means fixed at 0 and latent factor variances fixed at 1 for model identification.

Model fit was assessed through (a) evaluation of parameter estimates and residuals, with models examined for evidence of misfit through indicators such as negative residual variances and unexpectedly large or small estimates; (b) consideration of global fit indices, including the chi-square test of model fit, root mean square error of approximation (RMSEA), comparative fit index (CFI), Tucker Lewis index (TLI), and standardized root mean square residual (SRMR) following guidance summarized by Lomax (2013); and (c) chi-square difference testing of nested models using the DIFFTEST option for ULSMV in Mplus (Muthén & Muthén, 2019). More parsimonious models were favored when no significant difference in global fit was observed.

There were two items that, from a theoretical perspective, could contribute to more than one underlying construct. These items were #10 "¿Tienes amigos que hablen inglés y español? / Do you have friends who speak English and Spanish?" and the follow-up question #11 "¿Cuántos? / How many?" We hypothesized that these two items might reflect Spanish exposure and English exposure, or they might only reflect to Spanish exposure (because English is the majority language in the U.S.). To assess this, we compared models including these items cross-loaded onto both factors to models with the items only loaded onto the Spanish exposure factor.

Upon identification of the underlying structure with the best balance of model fit, parsimony, and alignment with theoretical construction, we computed reliability indices for each subscale identified. Coefficient omega hierarchical was used to accommodate potential violations of tau equivalence (McNeish, 2017).

Practical Scoring Approaches

We considered several scoring approaches for practical use of the scale, drawing on related discussion from DiStefano et al. (2009) and Logan et al. (2019). Ease of administration and interpretation is essential to the practical, day-to-day useability of assessments. Consequently, we examined a restriction on the factor loadings, which required each item to contribute equally to its corresponding subscale. This analysis is similar to comparing a 2-parameter item response theory (2-PL IRT) model to a 1-PL IRT model. We compared global fit for the restricted model to a model without restriction. We also obtained metrics of parameter bias to determine the practical difference between equal weighting of items compared to differential item weighting within each subscale. Based on the results, we constructed a preliminary useable system for scoring the measure.

Convergent Validity

After identifying the underlying structure with the best fit to the data, we examined indicators of convergent validity for the Houston-Q. To do this, we used the developed measure to compute scores for each scale construct for each child. We then examined correlations among the obtained scale scores and concurrent measures of Spanish and English language. The concurrent measures of language in Spanish were BESA/BESA-ME Morphosyntax, CELF-4 Recordando Oraciones, and TVIP. In English, the three language measures were BESA/BESA-ME Morphosyntax, CELF-5 Sentence Recall, and PPVT-4. We expected the subscales of self-reported Spanish proficiency to be positively associated with the Spanish language measures and the subscales of self-reported English proficiency to be positively associated with the English language measures. Similarly, we hypothesized that the subscales of bilingual experience would correlate with the Spanish and English measures, such that greater Spanish experience would

correspond with higher Spanish language scores and greater English experience would correspond with higher English language scores. Finally, we examined correlations between children's subscale scores on the Houston-Q and age.

Results

Descriptive information

Children in the sample varied widely in terms of their language proficiency profiles. To illustrate this variability, we descriptively examined participants' standard scores on the language measures used in this study separately for each language. These included standard scores in Spanish and English for the BESA/BESA-ME, sentence repetition subtest of the CELF-4 in Spanish and CELF-5 in English, and receptive vocabulary using the PPVT and TVIP. For 46% of the children in this sample, the difference between their scores in Spanish and English for the BESA/BESA-ME were within 10 standard points of each other, suggesting that about half of the children had relatively balanced morphosyntactic skills in both languages. For the remaining children, 31% had stronger English morphosyntactic skills (more than a 10-point difference in standard scores), while 23% had stronger Spanish skills. For vocabulary, 41% of the children had scores in Spanish and English within ten standard points of each other. In comparison, 24% of the children had stronger receptive vocabulary in English, and 35% had stronger Spanish receptive vocabulary.

Children in this sample also varied in language ability. The average score in the best language for the BESA/BESA-ME was 92.19 (*SD* = 17.06), for PPVT was 82.63 (*SD*=25.00), and for TVIP was 84.53 (*SD*=25.50). Forty-two percent of the children were receiving speech-language services in their schools. These aspects of language ability, proficiency, and use

indicate that our participants represent a heterogeneous group of bilingual children. Detailed information for the children in our sample is included in Table 1.

Sample Characteristics

Response frequencies for each item in the questionnaire are depicted in the Supplementary Material (Figures S1-S3). Generally, children in the present sample rated themselves as speaking both Spanish and English well (Spanish, n = 101, and English, n = 103, out of 113). However, the degree of how well children rated themselves as speaking each language varied. Children were slightly more likely to report Spanish as being easy (n = 97 out of 112) than English being easy (n = 87 out of 112), with more variability present in the reported degrees of English easiness compared to Spanish (Figure S1).

On items focused on bilingual experience, children were asked about different activities and whether these were done using both Spanish and English, only Spanish, or only English. There was also an option to indicate that they did not do the activity. Of these options, children most often reported using both languages during the activities. With their classroom teacher, 60% of children indicated that they used both Spanish and English, 21% used only Spanish, and 19% used only English. With respect to reading books, 59% of children reported reading in both languages, 28% read only in Spanish, and 13% read only in English. Similarly, 66% reported learning to write in both languages, 22% reported learning to write only in Spanish, and 13% reported learning to write only in English. When watching TV, 54% of children watched in both languages, 16% watched only in Spanish, and 30% watched in only English. At the park, 41% played using both Spanish and English, 29% used only Spanish, and 31% used only English. In family reunions, 41% of children used both languages, 34% used only Spanish, and 25% used only English. These findings are provided in Figure S2.

Overall, children reported having both family members and friends who spoke Spanish, English, and a combination of Spanish and English. When asked how much Spanish and English they heard each day, 27% of the children reported hearing a lot of both Spanish and English. Sixty-three children out of 112 reported hearing a lot of Spanish per day. Finally, 54 children out of 112 reported hearing a lot of English per day (Figure S3).

No patterns were observed in missing data. Children elected not to respond to questions randomly, with 32 instances of missing responses. Given that 3,051 total responses were possible (27 items and 113 total participants), and no patterns were observed, data were considered missing at random. We also examined patterns in children's responses for evidence of contradictory patterns or illogical response combinations. The questionnaire items were written to allow for all possible combinations of responses, but one noteworthy pattern occurred among 12 participants. Six children indicated that they were not good at speaking English but thought English was easy. Another six children stated that they were not good at speaking Spanish but thought Spanish was easy. Although this combination of perceptions seems unlikely, individuals can have the belief that learning a language is easy, even though they do not consider themselves to be good at speaking that language. Consequently, we did not interpret these response combinations as problematic.

Dimensionality and Reliability

Confirmatory factor analyses indicated that a model with three correlated factors yielded the best balance of global fit, reasonableness, consistency with theory, and model parsimony (see Figure 2). The model included a single factor underlying the items designed to measure children's self-perceptions of their proficiency in Spanish (i.e., "Self-Perception of Spanish"), a single factor underlying items designed to measure children's self-perceptions of their

proficiency in English (i.e., “Self-Perception of English”), and a single factor underlying self-reported bilingual experience (i.e., “Bilingual Experience”). This model, specified with item loadings and thresholds freely estimated, provided a good fit to the data: $\chi^2(296) = 325.90$ and $p = .1118$, RMSEA = 0.030 (90% CI [0.001, 0.048]), CFI = 0.936, TLI = 0.930, SRMR = 0.114. Coefficient omega hierarchical was computed to be .910 for Self-Perception of Spanish, .753 for Self-Perception of English, and .893 for Bilingual Experience indicating that the three factors showed good internal consistency reliability.

The two items that were hypothesized to contribute to more than one underlying factor (#10 “¿Tienes amigos que hablen inglés y español? / Do you have friends who speak English and Spanish?” and follow-up question #11 “¿Cuántos? / How many?”) were examined as indicators of Self-Perception of Spanish and of Self-Perception of English. Item loadings and model comparisons suggested that item #10 did not fit well on either factor, whereas #11 contributed reasonably to children’s Self-Perception of Spanish. Chi-square testing of model B (see Figure 1) with item #10 freely loaded onto Self-Perception of Spanish compared to being fixed at zero resulted in no significant difference in fit: $\Delta\chi^2$ (1) = 0.10, $p = .751$. Item #10 was removed from subsequent modeling, and #11 was loaded onto only the Self-Perception of Spanish proficiency factor. Global model fit statistics and chi-square comparisons of nested models are provided in Table 2. Standardized item loadings and thresholds are provided by item in Table 3.

Scoring System for the Houston-Q

When item loadings were restricted to be equivalent (analogous to a 1-PL IRT model), global model fit comparisons revealed a significantly worse fit to the data compared to the model with freely estimated loadings $\Delta\chi^2(23) = 59.34$, $p < .001$. Additionally, this restriction resulted in a total parameters bias of 35% across the subscales, with the least bias observed for the

Bilingual Experience factor (28%) compared to the Self-Perception of English (41%) or Self-Perception of Spanish (40%) factors. Consequently, the free estimation of item loadings was retained for the preliminary scoring system of the measure, which was constructed based on the standardized weighted contributions of each item (see Houston-Q Español, Houston-Q English, and Houston-Q Research spreadsheets). Given the random missing data patterns observed in the data used for the present study, the scoring system is designed to allow for the computation of scores with missing individual item responses.

The measure was scaled from 0-10 for the Self-Perception Scores of Spanish and English proficiency, where 0 = no proficiency and 10 = full proficiency. For Bilingual Experience, we scaled responses from 0-20, with 0 = all experiences in Spanish, 10 = equal experiences in Spanish and English, and 20 = all experiences in English. We elected to scale the values differently to reflect the differences in the underlying constructs.

Convergent Validity

Within the present participant sample, children scored an average of 7.73 (*SD* = 2.15) for Self-Perception of Spanish proficiency, suggesting relatively high levels of self-perceived proficiency in Spanish. Self-Perception of English was similarly high, with an average of 7.69 (*SD* = 2.09). The children's self-perception scores for proficiency in each language were significantly and positively associated with the behavioral measures of language with small to moderate correlations. The self-perception scores were negatively associated across languages (*r* = -.24, 95% CI [-.40, -.05], *p* = .013), indicating that children who reported high proficiency in Spanish tended to report lower proficiency in English and vice versa. Self-Perception of Spanish correlated with the Spanish measures CELF-4 Recordando Oraciones, TVIP, and BESA Morphosyntax at $r = .36$ (95% CI [.19, .51], $p < .001$), $r = .23$ (95% CI [.04, .40], $p = .017$), and

r = .42 (95% CI [.25 .56], p < .001), respectively. Self-Perception of English similarly correlated with the English measures CELF-5 Sentence Repetition, PPVT-4, and BESA Morphosyntax at r = .32 (95% CI [.14, .48], p < .001), r = .24 (95% CI [.05, .40], p = .013), and r = .23 (95% CI [.04, .40] p = .017), respectively.

On average, children indicated generally balanced bilingual experience, with slightly higher experience in Spanish than English, evidenced by the average Bilingual Experience at 8.94 *(SD* = 4.53). Appropriately, increased experience in Spanish was associated with a higher self-perception of Spanish proficiency: r = -.61 (95% CI [-.72, -.48] p < .001), and increased experience in English was associated with a higher self-perception of English proficiency: r = .42 (95% CI [.25, .56] p < .001). Age correlated weakly with self-perception of Spanish (r = -.19, 95% CI [-.36, -.01], p = .050), but not with the other two subscales. See Table 4 for full correlations.

Discussion

This study aimed to examine the reliability and convergent validity of the Houston Questionnaire in a sample of young bilingual children. In this study, we included children with varying levels of bilingual proficiency and language ability to capture variability in bilingual experiences and proficiency. Our results provide initial evidence supporting the internal consistency reliability and preliminary criterion validity of the Houston Questionnaire as a child self-report assessment tool.

Dimensionality and Reliability

Our findings indicate that three correlated factors underlie children's responses to the Houston-Q: Self-Perception of Spanish Proficiency, Self-Perception of English Proficiency, and

Bilingual Experience. These three factors were moderately correlated, which suggests that participants' responses reflected distinct but related constructs. Each subscale had overall good internal consistency reliability, which indicates that the questionnaire items were generally cohesive within each factor (Revelle & Condon, 2019). These results suggest that Houston-Q can elicit reliable responses from young bilingual children. In other words, the questions of the Houston-Q elicit responses that are generally consistent in terms of bilingual experience and self-ratings of Spanish and English proficiency. For example, a child is likely to respond that they are good at speaking Spanish and that Spanish is easy. This appropriate internal consistency reliability is crucial for a self-report measure since the questions must reliably measure the same construct (Dunn et al., 2014; McNeish, 2017). Failing to do so would suggest that the measure is not designed appropriately (e.g., not worded properly) or that different constructs are being measured (e.g., constructs other than bilingual experience).

The questionnaire items aligned well with the hypothesized underlying factors. For example, the questions that we expected to reflect Self-Perception of Spanish Proficiency were reliably associated with one another. The same was found for Self-Perception of English Proficiency and Bilingual Experience. There was no evidence of misfit in the final model, which suggests that obtaining these three subscale scores from the Houston-Q is appropriate.

We hypothesized that two questionnaire items could contribute to Self-Perception of Spanish proficiency and/or to Self-Perception of English proficiency. We directly tested the fit of question #10, "Do you have friends who speak Spanish and English?" and follow-up question #11, "How many?" as indicators of these underlying factors. The results indicated that question #10 did not directly align with either self-perception of Spanish or self-perception of English, but question #11 did align with self-perception of Spanish. We interpret these findings as primarily

reflective of the sampling context in Houston. In the present sample of participants, most children reported having at least some friends who speak Spanish and English, which resulted in relatively limited variability (i.e., restriction of range) for question #10. This limited variability restricted the item's potential to contribute to any factor. Question #11, however, did result in sufficient response variation to serve as an indicator of self-perception of Spanish. Because English is the predominant language used in the U.S. and especially in schools in the U.S., it is reasonable that children who report having more friends who speak both English and Spanish would similarly have a greater self-perception of their Spanish proficiency.

For the present study, we purposefully included children with diverse ranges of exposure and from a relatively broad age range to reflect the variability typically seen among bilingual children in the U.S. However, these bilingual children are speakers of Spanish in a city where Spanish is frequently heard and used by the broader community, and where opportunities for formal education in Spanish exist. Therefore, these results provide initial evidence supporting the utility of the Houston-Q across these characteristics. If there were substantial differences in the validity or reliability of the measure between the subgroups, we would expect evidence of lack of fit such as poor global model fit and spurious parameter estimates. Instead, we found that the global fit of the model was good, especially given the relatively small sample size, and the parameter estimates were generally stable. Although replication is certainly necessary to further explore the validity, reliability, and overall functioning of the scale across subpopulations of bilingual learners, the current findings provide preliminary evidence of the utility of the scale across diverse Spanish-English speaking learners.

Scoring System for the Houston-Q

Using the results of the confirmatory factor analysis, we created scaled scores for Spanish and English self-perception of proficiency and bilingual experience. A scale of 0 to 10 was used to describe Self-Perception of Proficiency in Spanish and English and a scale from 0 to 20 to describe Bilingual Experience. Importantly, we weighted the contribution of each item within each scale to align with its unique factor loading, given that the items did not equally reflect the underlying constructs of interest. We tested whether the items could be scaled to contribute equally but found that this significantly worsened the reliability of the questionnaire. Forcing the items to contribute equally resulted in substantial bias (i.e., 28 – 41%) in each subscale score. In other words, weighting items equally resulted in substantially different subscale scores when compared to varying the item weights. These results suggest that some of the questionnaire items were more important indicators of children's self-perception proficiency and bilingual experience than others. For example, question #1, "Are you good at speaking Spanish?" was a more robust and consistent indicator of self-perception of Spanish proficiency across children than question #6, "Do you have friends who only speak Spanish?". For question #1, the response "yes" reliably reflected a higher overall self-perception of proficiency in Spanish. Children who received a high score on self-perception of proficiency in Spanish generally responded "yes" to question #1. On the other hand, for question #6, more friends who speak Spanish typically but not always reflected higher self-perception of Spanish proficiency. There was a weaker association between children's total scores for self-perception of Spanish proficiency and their responses to question #6. This variation in item contributions was observed for all three subscales and is evident in the standardized item loadings. Our scoring system reflects this variation by weighting each item differently.

The scoring system also allows children to receive scores on each of the subscales even if they do not respond to individual questionnaire items. We incorporated this design feature because the results of the present work revealed no patterns in children's missing data, suggesting that children randomly skipped questions throughout the questionnaire. Children did not frequently skip items, and when they did, there was no apparent reason why they skipped. We believe this may be attributable to normal lapses in attention. Consequently, it is reasonable to obtain a subscale score even when children skip a few items across the questionnaire.

Convergent Validity

Children's self-perception of Spanish proficiency correlated positively with the Spanish language measures. Correlations with sentence repetition and productive morphology were moderate, and correlations with receptive vocabulary were weak-to-moderate. Similarly, self-perception of English proficiency positively correlated with the English language measures overall. In English, the correlations with sentence repetition, receptive vocabulary, and productive morphology were weak-to-moderate. Although replication with an independent, larger sample is necessary to establish the magnitude of these associations more definitively, the direction of the correlations is consistent. It is important to note that the receptive vocabulary measures were normed on monolingual children and, therefore, are not appropriate estimation of the vocabulary knowledge of the bilingual children in this study, which may have lowered the magnitude of the correlations between the Houston-Q subscale scores and vocabulary, particularly for Spanish.

As expected, children's bilingual experience scores on the Houston-Q, which ranged from 0 to 20, with 10 indicating fully balanced experience in Spanish and English, also

correlated with the external standardized measures. Bilingual experience correlated moderately positively with self-perception of proficiency by language. Bilingual experience values between 0 and 10, which indicate more self-reported experience in Spanish, generally corresponded with higher Spanish language scores. Further, bilingual experience values between 10 and 20, which indicate more self-reported experience in English, generally corresponded with higher English scores. These results suggest the bilingual experience metric functions as expected, with self-reported exposure and use to each language aligning with norm-referenced scores in each respective language.

We interpret these small to medium correlations and the direction of the associations to be good indicators of the validity of the Houston-Q (Strauss & Smith, 2009). These correlations indicate that proficiency measures using behavioral tasks and the children's perception of their proficiency shared some properties, but they represent distinct constructs. This finding might be explained by the fact that the behavioral tasks tap into specific language skills, like children's ability to recall sentences, which might not necessarily be what children consider would qualify them as good speakers of a language. Because we did not design the study a priori to compare the strength of the correlations, we cannot make specific claims about what correlations are stronger or weaker than others. However, the directionality of the correlations provides us with necessary evidence informing the validity of Houston-Q. Recall that children's self-perception of Spanish proficiency using the Houston-Q correlated positively with receptive vocabulary, productive morphosyntax, and sentence repetition in Spanish while correlating negatively with the same measures in English. Namely, children who rate themselves as good speakers of a language tend to have higher scores from behavioral tasks in that language than children who consider themselves not to be good at speaking that language. Further, children who rated

themselves as high in both languages tended to have high scores in both languages. These findings suggest that children's responses to the Houston-Q rating are tapping into their proficiency in each language.

Sample-Specific Considerations

It is important to consider that most children in this study rated themselves as speaking both Spanish and English well, although the degree of their ratings varied. This consideration is particularly important because about 40% of the children in this study had standard scores for morphosyntax (BESA/BESA-ME) and receptive vocabulary (PPVT/TVIP) within 10 points of each other, which suggests that their proficiency in each language was at similar levels. These data need to be interpreted within the context. The data for this study was collected in Houston, a city where 39.3% of the overall population speak Spanish at home (U.S. Census Bureau, 2019). Bilingual education is available for children with limited English language ability because Texas law mandates bilingual instruction for elementary schools with 20 or more children who need English language support (Bilingual Education and Training Act). Spanish immersion is also available in some schools in Houston, but it is not mandated by law. Notably, a significant proportion of children in our sample attended bilingual programs and immersion programs. This strong bilingual context might impact the child's ability to rate themselves in each language since their everyday experiences include both Spanish and English, which might be different from other contexts in the U.S. Therefore, future studies should be conducted in other bilingual populations to examine the effect of the context on the reliability and validity of the Houston-Q.

The finding that self-perception of Spanish proficiency was associated with age and our operationalized metric of language ability is worth noting. The shift into more English-focused

environments as bilingual children get older in the U.S. may explain the negative relationship between age and participants' self-perception of Spanish proficiency. Recall that we included children between 4 and 8 years of age in this study. At age 4, children tend to spend more time in the home with their family, whereas by age 8, they are likely spending more time in the community and with friends. Although bilingual education offers a protective effect on the maintenance of Spanish language skills, it is not sufficient for some children (Castilla-Earls et al., 2019). This interpretation is supported by the finding that age was positively associated with children's English receptive vocabulary and productive morphology since we also observed that older children tended to have higher English language scores.

There is an important finding regarding children with low language ability that must be considered carefully. Although 42% of the children in this study were receiving speech/language services at the time of data collection, all children generally tended to rate themselves as speaking both languages well, although the degree of their ratings varied. It is crucial to design questionnaires with multiple questions from a measurement perspective. For example, in looking at the factor loadings (Table 3), the question "are you good at speaking Spanish?" was a strong indicator of "self-perception of Spanish" (i.e., .93 loading), whereas this was slightly weaker for "are you good at speaking English?" (i.e., .71 loading). These loadings can be roughly interpreted similarly to correlations with the overall factor. Although children tended to respond positively to both of these items, there was additional variation in their self-perceptions captured by the other questionnaire items. From a questionnaire design perspective, we did not design the Houston-Q to capture variation in language ability. We expected that even children with low language ability (i.e., language disorders) would rate themselves as speaking well in at least one of their languages. Our results suggested that this was the case. The Houston-Q cannot identify

children with low language ability but can potentially help identify a child, for example, with stronger language proficiency in Spanish than in English and who has more experiences in Spanish regardless of their language ability.

Language Awareness

The results of this study suggest that bilingual children have enough language awareness to complete a questionnaire about their perceptions of their bilingual experiences and proficiency. When children between the ages of 4 and 8 complete this questionnaire, they do so reliably, and their responses are in general agreement with their proficiency in each language. These results support previous studies that suggest that young children have enough language awareness to self-report their relative language proficiency and bilingual experience (e.g., Babino & Stewart, 2016; Rojo & Echols, 2017). This finding is of interest because children are usually not asked to provide this information, and instead, this information is often sought from parents and teachers. We did not compare whether parents, teachers, or children provide the most accurate information about the children, so we cannot make judgments about the overall accuracy of the different reports. However, our results suggest that children might have a role in providing this information because they are direct observers of their bilingual experience and might be able to estimate their knowledge in each language compared to what parents and/or teachers can report.

Clinical Application

An important piece of information during the assessment of language skills in bilingual children is to understand how bilingual experience and proficiency in each language may play a

role in the child's overall language ability. This understanding is key to differentiating language disorders from limitations or differences due to variability in proficiency and language exposure. Administering the Houston-Q to children as part of the bilingual assessment could provide important information about the child's perception of their current bilingual experience and general proficiency in each language, which might facilitate identification of children's baseline language experiences and strongest language prior to direct comprehensive language assessment. Since this study included children with various levels of language ability, we recommend that this questionnaire could be used by children with and without language disorders. Using the child's self-reported bilingual experience and proficiency in each language may support clinical assessment to consider the child's abilities in their strongest language for more accurate identification of language disorders. However, it is important to note that this questionnaire was not designed to identify children with low language ability.

Limitations

There are limitations to the interpretation of this study that are important to acknowledge. This work provides preliminary evidence of the reliability and validity of the Houston-Q for gaining some insight into Spanish-English speaking bilingual children's language experience and proficiency. Although we believe the current scoring system is functional for clinical and research use, further vetting with independent samples of bilingual children in the U.S. (and other countries) is needed to better understand how children with different bilingual language experiences respond to the Houston-Q. The questions may elicit different patterns of responses in different contexts, and there may be outside factors that influence these patterns. For example, question #10, which asks about having friends who speak Spanish and English, may be a more

effective indicator of self-perception of Spanish proficiency in areas with less bilingual language support compared to Houston. Or, in contexts where Spanish is the primary societal language, question #10 could reflect self-perception of English proficiency. These differences are essential to examine carefully, to better understand the information that can be obtained from the Houston-Q in various contexts.

Given the size of the current sample, we were not able to test for differences in scale functioning by individual differences among children within the sample. Specifically, although we examined overall associations between children's age and their subscale scores on the Houston-Q, we did not have sufficient power to assess measurement invariance by language ability level or age. Consequently, it is important to recognize that this study provides initial evidence that children can complete the Houston-Q and that their responses broadly reflect valuable information. Further specific examination of the scale (and subscale) functioning across diverse samples of bilingual children, particularly among children at risk for language disorders, is needed to inform the clinical utility of the measure in diagnostic contexts. For future users of the Houston-Q, we recommend starting with the initial scoring system provided in this study. A careful examination of the robustness of the provided item parameters will be needed to validate it for use in other contexts and samples.

Finally, it is important to note the limitations of current measurement modeling, particularly in quantifying distinct but related factors using a combination of dichotomous and polytomous response options. We prioritized the establishment of a practical scoring approach for the Houston-Q so that it could be easily used with basic computer software by both clinicians and researchers. Specifically, we developed item weights for the Houston-Q are based on the identified loadings from the categorical confirmatory factor analysis. The results of this work do

clearly suggest that this approach is preferable compared to weighting the items equally. Still, the generalizability of the loadings is limited to the extent to which the participant sample is representative. As more sophisticated techniques for scoring and representative sampling of participants become more accessible, a more generalizable scoring approach may be implemented to obtain scores quickly and reliably for individual children.

Conclusion

In this study, we examined the internal consistency reliability and preliminary criterion validity of the Houston Questionnaire. The Houston-Q was created to gather information from the child's perspective about their bilingual experience and proficiency in each language. Our results provide evidence in support of the reliability and validity of the Houston-Q when used with bilingual children between the ages of 4 and 8 with various levels of language ability and different bilingual proficiency profiles.

References

Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A systematic review and meta-analysis of the cognitive correlates of bilingualism. *Review of Educational Research, 80*(2), 207–245. https://doi.org/10.3102/0034654310368803

Archibald, L. M. D., & Joanisse, M. F. (2009). On the Sensitivity and Specificity of Nonword Repetition and Sentence Recall to Language and Memory Impairments in Children. *Journal of Speech, Language and Hearing Research, 52*(4), 899-914. https://doi.org/10.1044/1092-4388(2009/08-0099)

Arias, G., & Friberg, J. (2017). Bilingual language assessment: Contemporary versus recommended practice in American schools. *Language, Speech, and Hearing Services in Schools, 48*(1), 1–15. https://doi.org/10.1044/2016_LSHSS-15-0090

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association

Artiles, A. J., Rueda, R., Salazar, J., & Higareda, I. (2002). English- language learner representation in special education in California urban school districts. In D. Losen & G. Orfield (Eds.), *Racial inequality in special education* (pp. 117–136). Cambridge, MA: Harvard Education Press.

Babino, A. & Stewart, M. A. (2016). "I like English better": Latino dual language students' investment in Spanish, English, and bilingualism. *Journal of Latinos and Education, 16*(1), 18–29. https://doi.org/10.1080/15348431.2016.1179186

Bedore, L. M., & Peña, E. D., García, M., & Cortez, C. (2005). Conceptual versus monolingual scoring: When does it make a difference? *Language, Speech, and Hearing Services in Schools, 36*(3), 188-200. https://doi.org/10.1044/0161-1461(2005/020)

Bedore, L. M., & Peña, E. D. (2008). Assessment of bilingual children for identification of language impairment: Current findings and implications for practice. *International Journal of Bilingual Education and Bilingualism, 11*(1), 1–29. https://doi.org/10.2167/beb392.0

Bedore, L. M., Peña, E. D., Gillam, R. B., & Ho, T.-H. (2010). Language sample measures and language ability in Spanish English bilingual kindergarteners. *Journal of Communication Disorders, 43*(6), 498–510. https://doi.org/10.1016/j.jcomdis.2010.05.002

Bedore, L. M., Peña, E. D., Joyner, D., & Macken, C. (2011). Parent and teacher rating of bilingual language proficiency and language development concerns. International *Journal of Bilingual Education and Bilingualism, 14*(5), 489–511.

Bedore, L. M., Peña, E. D., Summers, C. L., Boerger, K. M., Resendiz, M. D., Greene, K., ... & Gillam, R. B. (2012). The measure matters: Language dominance profiles across measures in Spanish–English bilingual children. *Bilingualism*, *15*(3), 616-629.

Bishop, D. V., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & Catalise Consortium. (2016). CATALISE: A multinational and multidisciplinary Delphi consensus study. Identifying language impairments in children. *PLOS One*, *11*(7), e0158753.

Bridges, K., & Hoff, E. (2014). Older sibling influences on the language environment and language development of toddlers in bilingual homes. *Applied Psycholinguistics*, *35*(2), 225–241. https://doi.org/10.1017/S0142716412000379

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñones, H. R., & Young, S. L. (2018) Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health, 6*(149). https://doi.org/10.3389/fpubh.2018.00149

Bohman, T. M., Bedore, L. M., Peña, E. D., Mendez-Perez, A., & Gillam, R. B. (2010). What you hear and what you say: Language performance in Spanish-English bilinguals. *International Journal of Bilingual Education and Bilingualism, 13*(3), 325–344. https://doi.org/10.1080/13670050903342019

Castilla-Earls, A., Francis, D., Iglesias, A., & Davidson, K. (2019). The Impact of the Spanish-to-English Proficiency Shift on the Grammaticality of English Learners. *Journal of Speech, Language, and Hearing Research*, *62*(6), 1739–1754. https://doi.org/10.1044/2018_JSLHR-L-18-0324

Castilla-Earls, A., Bedore, L., Rojas, R., Fabiano-Smith, L., Pruitt-Lord, S., Restrepo, M. A., & Peña, E. (2020). Beyond scores: Using converging evidence to determine speech and language services eligibility for dual language learners. *American Journal of Speech-Language Pathology*, *29*, 1116–1132. https://doi.org/10.1044/2020_AJSLP-19-00179

Carroll, S. E. (2017). Exposure and input in bilingual development. *Bilingualism: Language and Cognition*, *20*(1), 3–16. https://doi.org/10.1017/S1366728915000863

de Ayala, R. J. (2013). *Factor analyses with categorical indicators*. In Y. Petscher, C. Schatschneider, & D. L. Compton (Eds.), Applied quantitative analyses in the educational and social sciences (pp. 208–242). New York, NY: Routledge.

De Houwer, A. (2014). *The absolute frequency of maternal input to bilingual and monolingual children: A first comparison*. In T. Grüter & J. Paradis (Eds.), Input and experience in bilingual development (pp. 37–58). Amsterdam, The Netherlands: John Benjamins.

De Houwer, A. (2017). Early multilingualism and language awareness. In J. Cenoz (Ed.), Encyclopedia of Language and Education (pp. 83-94). Switzerland: Springer.

DeMars, C. E. (2012). Confirming testlet effects. *Applied Psychological Measurement, 36*(2), 104–121. https://doi.org/10.1177/0146621612437403

DiStefano, C., Zhu, M., & Mîndrilă, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research and Evaluation, 14*(20). https://doi.org/10.7275/da8t-4g52

Dunn, L. M., & Dunn, D. M. (2007). *PPVT-4: Peabody Picture Vocabulary Test-Fourth Edition*. Pearson Assessments. https://search.library.wisc.edu/catalog/999616587302121

Dunn, L. M., Padilla, E. R., Lugo, D. E., & Dunn, L. M. (1986). *Tvip : Test De Vocabulario En Imagenes Peabody : Adaptacion Hispanoamericana = Peabody Picture Vocabulary Test : Hispanic-American Adaptation*. American Guidance Service. https://search.library.wisc.edu/catalog/999767172102121

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399-412. https://doi.org/10.1111/bjop.12046

Duursma, E., Romero-Contreras, S., Szuber, A., Proctor, P., Snow, C., August, D., & Calderon, M. (2007). The role of home literacy and language environment on bilinguals' English and Spanish vocabulary development. *Applied Psycholinguistics, 28*(1), 171–190. https://doi.org/10.1017/S0142716406070093

Farver, J. A. M., Lonigan, C. J., & Eppe, S. (2009). Effective early literacy skill development for young Spanish-speaking English language learners: An experimental study of two methods. *Child Development, 80*(3), 703–719. https://doi.org/10.1111/j.1467-8624.2009.01292.x

Gollan, T. H., Weissberger, G. H., Runnqvist, E., Montoya, R. I., & Cera, C. M. (2012). Self-ratings of spoken language dominance: A Multilingual Naming Test (MINT) and preliminary norms for young and aging Spanish-English bilinguals. *Bilingualism: language and cognition*, 15(3), 594-615.

Hammer, C. S., E. Komaroff, B. L. Rodriguez, L. M. Lopez, S. E. Scarpino, and B. Goldstein. (2012) "Predicting Spanish–English Bilingual Children's Language Abilities." Journal of Speech, Language, and Hearing Research 55 (5): 1251–1264.

Hoff, E., & Core, C. (2013). Input and language development in bilingually developing children. *Seminars in Speech and Language, 34*(4), 215–226. https://doi.org/10.1055/s-0033-1353448

Hoff, E. (2017). How bilingual development is the same as and different from monolingual development. *OLBI Working Papers*, 3–16.

Hoff, E., Burridge, A., Ribot, K. M., & Giguere, D. (2018). Language Specificity in the Relation of Maternal Education to Bilingual Children's Vocabulary Growth. *Developmental Psychology, 54*(6), 1011-1019. https://doi.org/10.1037/dev0000492

Kan, P. F., & Windsor, J. (2010). Wording learning in children with primary language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research, 53*(3), 739-756. https://doi.org/10.1044/1092-4388(2009/08-0248)

Kapantzoglou, M., Restrepo, M. A., Gray, S., & Thompson, M. S. (2015). Language ability groups in bilingual children: A latent profile analysis. *Journal of Speech, Language, and Hearing Research, 58*(5), 1549–1562. https://doi.org/10.1044/2015_JSLHR-L-14-0290

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Brief Intelligence Test* (Second Edition ed.). Pearson, Inc.

Kohnert, K. (2010). Bilingual children with primary language impairment: Issues, evidence, and implications for clinical actions. *Journal of Communication Disorders, 43*(6), 456–473. https://doi.org/10.1016/j.jcomdis.2010.02.002

Leonard, L.B. (2014). Children with specific language impairment., 2nd Edition. Cambridge, MA: MIT Press.

Logan, J. A. R., Jiang, H., Helsabeck, N., & Yeomans-Maldonado, G. (2019, June 25). Should I Allow my Confirmatory Factors to Correlate During Factor Extraction? Implications for the Applied Researcher. https://doi.org/10.31219/osf.io/zcsnv

Lomax, R. G. (2013). *Introduction to structural equation modeling*. In Y. Petscher, C. Schatschneider, & D. Compton (Eds.), Applied quantitative analysis and the social sciences (pp. 245-264). New York: Routledge.

Lutz, A. (2008). Negotiating home language: Spanish maintenance and loss in Latino families. *Latino(a) Research Review, 6*, 37-64.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research, 50*(4), 940–967.

Martin, B. (2012). Coloured language: identity perception of children in bilingual programmes. *Https://Doi.Org/10.1080/09658416.2011.639888*, *21*(1–2), 33–56. https://doi.org/10.1080/09658416.2011.639888

Melo-Pfeifer, S. (2015). Multilingual awareness and heritage language education: children's multimodal representations of their multilingualism. *Http://Dx.Doi.Org/10.1080/09658416.2015.1072208*, *24*(3), 197–215. https://doi.org/10.1080/09658416.2015.1072

McNeish, D. (2017). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*, 412–433. http://dx.doi.org/10.1037/met0000144

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavioral Research Methods, 22*, 2287-2305. https://doi.org/10.3758/s13428-020-01398-0

Muthén, L. K., & Muthén, B. O. (1998-2019). *Mplus user's guide* (8th Ed). Los Angeles, CA: Múthen & Múthen

National Kids Count (2020). *Children living in linguistically isolated households by family nativity in the United States*. The Annie E. Casey Foundation. https://datacenter.kidscount.org/data/tables/129-children-living-in-linguistically-isolated-households-by-family-nativity

Obied, V. M. (2009). How do siblings shape the language environment in bilingual families? *International Journal of Bilingual Education and Bilingualism*, *12*(6), 705–720. https://doi.org/10.1080/13670050802699485

Osika, W., Friberg, P., & Wahrborg, P. (2007). A new short self-rating questionnaire to assess stress in children. *International journal of behavioral medicine*, *14*(2), 108–117. https://doi.org/10.1007/BF03004176

Pratt, A. S., Peña, E. D., & Bedore, L. M. (2020). Sentence repetition with bilinguals with and without DLD: Differential effects of memory, vocabulary, and exposure. *Bilingualism: Language and Cognition*, *24*(2), 305-318. https://doi.org/10.1017/s1366728920000498

Peña, E. D., Bedore, L. M., Gutierrez-Clellen, V. F., Iglesia, A., & Goldstein, B. A. (2008). *Bilingual English-Spanish Assessment - Middle Extension Experimental Test Version (Besa-Me).* Unpublished manuscript.

Peña, E. D., Bedore, L. M., & Kester, E. S. (2016). Assessment of language impairment in bilingual children using semantic tasks: two languages classify better than one. *International Journal of Language & Communication Disorders, 51*(2), 192–202. https://doi.org/10.1111/1460-6984.12199

Peña, E. D., Bedore, L. M., Gutierrez-Clellen, V. F., Iglesia, A., & Goldstein, B. A. (2016). *Bilingual English-Spanish Assessment - Middle Extension Field Test Version (Besa-Me).* Unpublished manuscript.

Peña, E. D., Gillam, R. B., & Bedore, L. M. (2014). Dynamic assessment of narrative ability in English accurately identifies language impairment in English language learners. *Journal of Speech, Language, and Hearing Research, 57*, 2206–2220. https://doi.org/10.1044/2014_JSLHR-L-13-0151

Peña, E. D., Gutierrez-Clellen, V. F., Iglesias, A., Goldstein, B., & Bedore, L. M. (2018). *Bilingual English-Spanish Assessment (Besa)*. Brookes Publishing.

Perez-Leroux, A. T., Cuza, A., & Omas, D. (2011). From parental attitudes to input conditions Spanish-English bilingual development in Toronto. In K. Potowski (Ed.), *Bilingual youth: Spanish in English-speaking societies* (pp. 149–176). John Benjamins.

Place, S., & Hoff, E. (2011). Properties of Dual Language Exposure That Influence 2-Year-Olds' Bilingual Proficiency. *Child Development*, *82*(6), 1834–1849. https://doi.org/10.1111/J.1467-8624.2011.01660.X

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

Revelle, W., & Condon, D. M. (2019). Reliability from alpha to omega: A tutorial.

Psychological Assessment, 31(12), 1395 – 1411. https://doi.org/10.1037/pas0000754

Rojas, R., Iglesias, A., Bunta, F., Goldstein, B., Goldenberg, C., & Reese, L. (2016). Interlocutor differential effects on the expressive language skills of Spanish-speaking English learners. *International journal of speech-language pathology*, *18*(2), 166–177. https://doi.org/10.3109/17549507.2015.1081290

Rojo, D. P., & Echols, C. H. (2017). Accepting labels in two languages: Relationships with exposure and language awareness. *OLBI Journal*, *8*. https://doi.org/10.18192/OLBIWP.V8I0.2115

Rujas, I., Mariscal, S., Murillo, E., & Lázaro, M. (2021). Sentence repetition tasks to detect and prevent language difficulties: A scoping review. *Children*, 8(7), 578.

Samson, J., & Lesaux, N.K. (2009). Language minority learners in special education: Rates and predictors of identification for services. *Journal of Learning Disabilities*, 42(2), 148-1

Semel, E., Wiig, E. H., & Secord, W. A. (2006). *Clinical Evaluation of Language Fundamentals–Fourth Edition, Spanish Version*. Pearson Education, Inc.

Solans, M., Pane, S., Estrada, M. D., Serra-Sutton, V., Berra, S., Herdman, M., Alonso, J., & Rajmil, L. (2008). Health-related quality of life measurement in children and adolescents: a systematic review of generic and disease-specific instruments. *Value in health: the journal of the International Society for Pharmacoeconomics and Outcomes Research*, *11*(4), 742–764. https://doi.org/10.1111/j.1524-4733.2007.00293.x

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology*, *5*, 1-25. https://doi.org/10.1146/annurev.clinpsy.032408.153639

Svalberg, A., M-L. (2007). Language awareness and language learning. *Language Teaching, 40*(4), 287-308. https://doi.org/10.1017/S0261444807004491

Tomoschuk, B., Ferreira, V. S., & Gollan, T. H. (2019). When a seven is not a seven: Self-ratings of bilingual language proficiency differ between and within language populations. *Bilingualism: Language and Cognition*, 22(3), 516-536.

U.S. Census Bureau. (2021) *2019: American Community Survey 1-Year Subject Tables: S1601 Language Spoken at Home*. Retrieved from: https://data.census.gov/cedsci/table?q=Houston&t=Language%20Spoken%20at%20Home&tid=ACSST1Y2019.S1601

Vagh, S. B., Pan, B. A., & Mancilla-Martinez, J. (2009). Measuring growth in bilingual and monolingual children's English productive vocabulary development: The utility of combining parent and teacher report. *Child Development, 80*(5), 1545–1563. https://doi.org/10.1111/j.1467-8624.2009.01350.x

Wiig, E. H., Semel, E., & Secord, W. A. (2013). *Clinical Evaluation of Language Fundamentals–Fifth Edition*. Pearson Education, Inc.

Wood, Hoge, Schatschneider & Castilla-Earls (2018). Predictors of item accuracy on the *Test de Vocabulario en Imagenes Peabody* for Spanish-English speaking children in the United States. *International Journal of Bilingual Education and Bilingualism*, https://doi.org/10.1080/13670050.2018.1547266

Table 1. Demographics and Language measure scores for children in the study (*n*=113)

	n	*M*	*SD*	%
Age (in months)	113	71.05	12.46	
Gender				
Male	64			56.6 %
Female	49			43.4 %
Mother's Level of Education				
No college	63			54.5 %
At least some college	50			45.5 %
Does the child qualify for free/reduced lunch?				
No	36			30.0 %
Yes	77			70.0 %
Child has received/is receiving services for speech/language?				
No	66			58.4 %
Yes	47			41.6 %
Language Spoken at Home				
English	12			10.6%
Spanish	56			49.6%
Both English and Spanish	45			39.8%
School Programs				
English-only	5			4.4%
Bilingual or Immersion	101			89.4%
Other: Saturday Spanish School	7			6.2%
Language Measures Norm-referenced assessments				
BESA/BESA-ME Morph Spanish		80.93	18.66	
BESA/BESA-ME Morph English		84.78	19.47	
BESA/BESA-ME Morph best language		92.19	17.06	
TVIP Spanish		86.72	17.93	
PPVT English		85.26	20.12	
CELF RO Spanish		6.75	3.11	
CELF SR English		6.76	3.56	

Note. BESA/BESA-ME = Bilingual English-Spanish Assessment/Bilingual English-Spanish Assessment-Middle Extension. Morph = Morphosyntax. TVIP = Test de Vocabulario en Imágenes Peabody. PPVT = Peabody Picture Vocabulary Test, 4th Edition; CELF = Clinical Evaluation of Language Fundamentals. RO = Recordando Oraciones / Recalling Sentences. SR = Sentence Repetition.

Table 2

Fit Indices for Hypothesized Models Underlying Questionnaire

Model		χ^2	*df*	$\Delta\chi^2$	Δdf	ΔSig.	RMSEA	LB	UB	CFI	TLI
A	4-Factor with Items 10-11 crossed[1]	97.898	96		---		0.013	<.001	0.052	0.981	0.976
	2-Factor with Items 10-11 crossed	103.470	101	5.895	5	.317	0.015	<.001	0.052	0.975	0.970
B	2-Factor with 10-11 on Spanish	105.984	103	2.625	2	.269	0.016	<.001	0.052	0.970	0.965
	2-Factor with 11 only on Spanish[2]	91.864	89		---		0.017	<.001	0.054	0.974	0.970
A	2-Factor: Bilingual Experience[1]	63.727	43		---		0.066	0.026	0.098	0.948	0.934
B	1-Factor: Bilingual Experience[2]	64.360	44	0.233	1	.630	0.064	0.024	0.096	0.949	0.936

Note. $\Delta\chi^2$ is reported for the model comparisons against the previous (above) model.

[1]Depicted in Figure 1A.

[2]Depicted in Figure 2. Finalized through discussion of item functioning, global fit, and consistency with theoretical expectations. The decrease in degrees of freedom reflects the full removal of question #10 from the measurement model.

Table 3

Standardized Item Loadings and Thresholds for Final Model

Factor	Questionnaire Item	Loading (SE)	Thresholds (SE)
Self-Perception of Spanish	1. Speak Spanish well (0/1)	0.93 (0.11)	-1.25 (0.16)
	2. Degree of speaking Spanish well	0.63 (0.10)	-1.13 (0.15)
			-0.84 (0.14)
			-0.60 (0.13)
			-0.49 (0.13)
	20. Spanish easiness (0/1)	0.79 (0.12)	-1.11 (0.15)
	21. Degree of Spanish easiness	0.69 (0.09)	-1.06 (0.15)
			-0.88 (0.14)
			-0.69 (0.13)
			-0.47 (0.12)
	6. Friends who speak Spanish (0/1)	0.37 (0.13)	-0.52 (0.12)
	7. Number of friends who speak Spanish	0.40 (0.14)	-0.77 (0.14)
			-0.28 (0.1)
			-0.04 (0.13)
			0.07 (0.13)
	24. Quantity of Spanish heard each day.	0.42 (0.12)	-0.96 (0.14)
			-0.62 (0.13)
			-0.32 (0.12)
			-0.16 (0.12)
	11. Number of Spanish-English speaking friends	0.27 (0.12)	-0.91 (0.14)
			-0.65 (0.13)
			-0.11 (0.12)
			0.23 (0.12)

Factor	Questionnaire Item	Loading (SE)	Thresholds (SE)
Self-Perception of English	3. Speak English well (0/1)	0.71 (0.16)	-1.35 (0.17)
	4. Degree of speaking English well	0.60 (0.12)	-1.33 (0.17)
			-0.93 (0.14)
			-0.70 (0.13)
			-0.43 (0.13)
	22. English easiness (0/1)	0.58 (0.14)	-0.76 (0.13)
	23. Degree of English easiness	0.45 (0.19)	-1.02 (0.15)
			-0.77 (0.14)
			-0.34 (0.13)
			-0.05 (0.12)

	8. Friends who speak English (0/1)	0.56 (0.17)	-0.77 (0.13)
	9. Number of friends who speak English	0.17 (0.16)	-0.74 (0.14)
			-0.32 (0.13)
			-0.18 (0.13)
			-0.08 (0.13)
	25. Quantity of English heard each day.	0.45 (0.13)	-1.11 (0.15)
			-0.62 (0.13)
			-0.223 (0.12)
			0.05 (0.12)
Factor	**Questionnaire Item**	**Loading (SE)**	**Thresholds (SE)**
Bilingual Experience	5a. Language used with a family member (1).	0.80 (0.06)	-0.18 (0.12)
			1.11 (0.15)
	5b. Language used with a family member (2).	0.42 (0.10)	-0.34 (0.12)
			0.76 (0.13)
	5c. Language used with a family member (3).	0.74 (0.07)	-0.01 (0.12)
			0.88 (0.14)
	12. Language spoken with bilingual friends.	0.65 (0.07)	-0.45 (0.12)
			0.45 (0.12)
	13. Language used with teacher.	0.24 (0.11)	-0.82 (0.14)
			0.88 (0.14)
	14. Language used for learning to write.	0.56 (0.08)	-0.79 (0.13)
			1.15 (0.15)
	15. Language used for watching TV.	0.62 (0.08)	-1.01 (0.15)
			0.52 (0.13)
	16. Language used when playing at the park.	0.79 (0.06)	-0.56 (0.13)
			0.51 (0.13)
	17. Language used at parties/family reunions.	0.65 (0.08)	-0.41 (0.13)
			0.66 (0.13)
	18. Language used to read books.	0.62 (0.07)	-0.51 (0.13)
			1.01 (0.15)
	19. Language used for learning to read.	0.52 (0.09)	-0.58 (0.13)
			1.12 (0.15)

Note. The underlying latent trait mean was set to zero, with a variance of 1.
For the Bilingual Experience latent factor, -1 = Experience in Spanish, 0 = Experience in Spanish and English, and 1 = Experience in English.

Table 4

Means, standard deviations, and correlations with confidence intervals

	Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10
Houston-Q	1. Spanish SP	7.73	2.15										
	2. English SP	7.69	2.09	-.24*									
	3. Bilingual Exp	8.94	4.53	-.61**	.42**								
Spanish	4. CELF RO	6.75	3.11	**.36****	-.10	**-.27****							
	5. TVIP	86.72	17.93	**.23***	-.08	**-.27****	.64**						
	6. BESA Morph	80.93	18.66	**.42****	-.11	**-.35****	.81**	.69**					
English	7. CELF SR	6.76	3.56	**-.43****	**.32****	**.36****	.30**	.19*	.16				
	8. PPVT	85.26	20.12	**-.40****	**.24***	**.38****	.07	.21*	.05	.73**			
	9. BESA Morph	84.78	19.47	**-.39****	**.23***	**.35****	.12	.15	.11	.78**	.80**		
	10. Best BESA	92.19	17.06	**-.22***	.15	.14	.44**	.40**	.50**	.70**	.61**	.78**	
	11. Age (mos)	70.05	12.46	**-.19***	-.13	.02	-.15	.01	-.05	.10	.21*	.47**	.38**

Note. M and *SD* are used to represent mean and standard deviation, respectively. SP = Self-Perception. Exp = Experience. CELF RO = Clinical Evaluation of Language Fundamentals. RO = Recordando Oraciones / Recalling Sentences. TVIP = Test de vocabulario en imagenes Peabody. BESA = Bilingual English-Spanish Assessment. Morph = Morphosyntax. SR = Sentence Repetition. PPVT = Peabody Picture Vocabulary Test, 4th Edition. * indicates $p < .05$. ** indicates $p < .01$