Journal of Child Language (2018), **45**, 939–958
doi:10.1017/S0305000917000605

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Validity of caregiver-report measures of language skill for Wolof-learning infants and toddlers living in rural African villages

Ann M. WEBER[1]*, Virginia A. MARCHMAN[2], Yatma DIOP[3], and Anne FERNALD[2]

[1]Department of Pediatrics, Stanford University School of Medicine, USA, [2]Department of Psychology, Stanford University, USA and [3]Human Development and Family Studies, Michigan State University, USA
*Corresponding author. E-mail: annweber@stanford.edu

(Received 14 September 2016; revised 3 July 2017; accepted 22 November 2017;
first published online 9 March 2018)

Abstract

Valid indigenous language assessments are needed to further our understanding of how children learn language around the world. We assessed the psychometric properties and performance of two caregiver-report measures of Wolof language skill (language milestones achieved and vocabulary knowledge) for 500 children (ages 0;4 to 2;6) living in rural Senegal. Item response models (IRM) evaluated instrument- and item-level performance and differential function by gender. Both caregiver-report measures had good psychometric properties and displayed expected age and socioeconomic effects. Modest concurrent validity was found by comparing the caregiver-report scores to transcribed child language samples from a naturalistic play session. The caregiver-report method offers a valid alternative to more costly tools, such as direct behavioral assessments or language sampling, for measuring early language development in non-literate, rural African communities. Recommendations are made to further improve the performance of caregiver-report measures of child language skill in these settings.

Keywords: Parent report methodology; Early language milestones; West African Languages (Wolof)

Introduction

With a global focus on the importance of early childhood development (ECD), there is an increasing demand for adequate developmental measures – especially for young children under three years of age – that can be used for program evaluation and program planning worldwide. Language is an extremely important domain of ECD to assess since successful language skills are crucial to children's cognitive and socio-emotional development (Kuhl & Rivera-Gaxiola, 2008) and predictive of later cognitive function and school achievement in both high-income (Duncan *et al.*, 2007; Marchman & Fernald, 2008) and low-income countries (Fernald, Kariger, Engle, & Raikes, 2009; Pollitt & Triana, 1999). However, few tools are available for use with young children learning the indigenous languages spoken in most low-income countries. Here, we report on two measures that were developed to evaluate the effectiveness of a parenting program run by the international non-profit

organization Tostan in subsistence-level villages in the Kaolack region of Senegal, West Africa. A central goal of the evaluation was to test whether program participation was associated with significant change in caregivers' behavior, and importantly, whether these changes were associated with change in children's language learning (Weber, Fernald, & Diop, 2017). To address the latter question, cost-effective, valid, and reliable measures for assessing early language development in more than 500 Wolof-learning infants and young children were needed.

Over 2,000 rural communities in six West African countries have participated in Tostan's Community Empowerment Program (CEP), a three-year human rights-based education program (Cislaghi, Gillespie, & Mackie, 2016). In a study commissioned by Tostan, parents in many Senegalese villages reported that engaging verbally with young children was frequently discouraged (Zeitlin, 2011), consistent with ethnographic observations in other West African societies (e.g., Richman, Miller, & LeVine, 1992). Moreover, there was concern that low levels of parent–child interaction early in life might be related to the low literacy rates of Wolof-speaking children in the customarily French or French–Arabic schools in Senegal. Therefore, Tostan developed a new parent education program, the Reinforcement of Parental Practices (RPP) program, to build on their three-year CEP (Tostan, n.d.). With a focus on caregiving for children from birth to six years, the RPP program comprised group sessions and bi-monthly home visits over a 9–10 month period, drawing on scientific discoveries about the crucial role of early language experience in cognitive development and later school achievement. The RPP program initially aimed to enable caregivers to engage more effectively in verbal interactions with their infants, thus building a stronger foundation for their children's learning skills.

A central goal of the evaluation was to assess changes in the language skills of children whose caregivers participated in the RPP program, compared to children in a matched control group (Weber *et al.*, 2017). To develop a language assessment appropriate for use in the Wolof language and culture, we began by selecting two caregiver-report instrument formats that have been successfully adapted for use in other low-income settings: (a) Milestones Checklist: a checklist of language and communication skills, or language milestones, listed in order of increasing difficulty; and (b) Vocabulary Inventory: an inventory of vocabulary words of varying levels of difficulty (i.e., some words would be expected to be known by most children, whereas other words on the list would be known by a smaller set of children) that are potentially produced by young Wolof-learning children. Although the structure and underlying logic differ for the two measures, both formats take advantage of the fact that caregivers are typically keen observers of their child's early communicative behaviors. Caregivers have the opportunity to observe the child in a variety of situations and can thus provide data that may be more representative of a child's developmental level than what would be observed in a laboratory-based language sample or on a standardized test. In addition, both formats focus on emerging behaviors, asking caregivers to report on what the child can do 'now', as opposed to retrospective questions such as "At what age did your child hold his head up?" Many studies have demonstrated that caregivers can provide reliable and valid information about a child's language ability in American English (for review, see Fenson, Marchman, Thal, Dale, Reznick, & Bates, 2007) and other languages (e.g., Alcock, Rimba, Holding, Kitsao-Wekulo, Abubakar, & Newton, 2015; Hamadani *et al.*, 2010), including children learning two languages simultaneously (Marchman & Martínez-Sussmann, 2002).

Here, we compare estimates of children's language level from the Milestones Checklist and Vocabulary Inventory to estimates based on direct observation of children's abilities

during a naturalistic play session. In traditional studies of language development, naturalistic observation is a common way to observe children's abilities, as the aim is to create conditions where the child will produce language in a manner that will accurately reflect their abilities. Of course, the behavior observed in a play session is highly dependent on the context, for example, the particular toys or activities involved and the style and skill of the interlocutor. Moreover, while engaging one-on-one with an interlocutor is a frequent activity for many children, it may be much less familiar to others. For children and caregivers living in rural Senegal, a one-on-one play session may not reflect the way that caregivers and children engage on a regular basis. Nevertheless, given the goals of our analyses, it was important to obtain a direct measure of children's language to supplement the caregiver-report instruments.

This study has three objectives: First, we assess the psychometric properties of the Milestones Checklist and Vocabulary Inventory as measures of language skill for children aged 0;4 to 2;6 living in Wolof-speaking villages in rural Senegal. For this objective, we rely on item response models, which describe the probability of a given response to an item as a function of the properties of the item (e.g., difficulty) and the characteristics of the respondent (e.g., ability). In this way, we can test, for example, whether children with large vocabularies have a higher probability of knowing hard words than children with small vocabularies. An additional advantage of item response models is that they can be augmented to test for systematic differences in responses to individual items for different subgroups (also known as 'differential item function' or DIF) (Baranowski, Allen, Masse, & Wilson, 2006; Wilson, Allen, & Li, 2006). DIF is a common concern in test development that can lead to bias. For example, even though underlying ability is the same for boys and girls, more boys than girls may respond correctly to a test item using a sports analogy because boys may be more familiar with the sport. This would constitute item bias by gender for this item. Since very young girls (<20 months) have been found to learn new words more rapidly than boys in the US (Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991), we augmented the item response models to test for DIF by gender in the new Wolof language instruments.

Our second objective was to explore the validity of the two measures in several ways by comparing instrument scores: (a) to each other in a subsample for which both scores were available; (b) to child age and family socioeconomic status; and (c) to direct measures of the children's language ability (number of utterances, number of word tokens, and word-level mean length of utterance) obtained from a 5-min naturalistic language sample during a play session.

Finally, because the measures assess language constructs using different caregiver-report formats, a central question of interest was whether the two instruments make overlapping contributions, as well as independent contributions, to predicting children's productive speech during the language sample.

Method

Participants

We analyzed language data for 500 monolingual Wolof-learning children of caregivers in 24 communities in the Kaolack region of Senegal who were enrolled in the evaluation of Tostan's parenting program. To obtain the most accurate estimates of children's language ability, we interviewed the caregiver who spent the most daylight hours

https://doi.org/10.1017/S0305000917000605 Published online by Cambridge University Press

caring for the child. In the final sample of reporters, most caregivers (84.7%) had no classroom-based education, 11.5% had some primary school education, and less than 4% had any secondary or higher-level education. Almost all caregivers (93.9%) were the child's mother; the remaining caregivers were other adult female relatives.

Children who were identified as bilingual (i.e., they were spoken to more than 10% of the time in a language other than Wolof), who had a serious developmental delay, or who had a hearing, speech, or vision impairment, were excluded from the study. If a mother had twins, only one twin was included. Data for the Milestones Checklist were obtained for 501 eligible children (0;4 to 2;6), but we excluded one child with an implausibly high score for age. Data for the Vocabulary Inventory were obtained for a cohort of 241 older children in a restricted age range (1;8 to 2;6) whom we expected to be producing recognizable words. One vocabulary score was excluded for a child aged 2;0 whose caregiver responded 'yes' to all words in the list. Data from the direct observation assessment were available for all 500 children across both age cohorts (0;4–2;6).

Data collection

Data presented here were obtained at baseline, prior to the onset of the RPP program. Extensive survey, video, and audio data were collected for the participants by two field teams, who kept to a demanding 10-week schedule of work and travel during data collection. Given the large size of our study sample and the high prevalence of illiteracy in this population, both caregiver-report measures were administered in an interview format by trained field assessors.

To maximize the accuracy of estimating children's skill in the Wolof language, the caregiver-report measures were written in Wolof, and the assessors were trained to read the Wolof exactly as written. It is noteworthy that this protocol represented a significant change from common practice in survey research in Senegal. Although Wolof is the most widely spoken national language in Senegal, the Latin orthography for Wolof is not taught in schools and remains highly variable (De Swaan, 2001). As a result, although around 60% of Senegalese children attend primary school (UNICEF, n.d.), and may learn to read French, most Senegalese are unable to read and write in their own native language. For this reason, it is customary for survey questionnaires to be written in French and then translated 'online' into Wolof as the survey is being administered to local participants, a practice that can result in substantial variability due to differences in interviewers' wording of the questions, as well as variability in their translations over time. To eliminate these sources of variability, we began with an intensive one-week training session in which the field team mastered the rules of written Wolof and reviewed each questionnaire in detail. This training was then followed by extensive pilot testing of the questionnaires with Wolof-speaking families in three villages not involved in the RPP intervention, to give the field team ample practice in administering all assessments, and to solicit feedback from participants. Although the questionnaires were initially translated by certified Wolof translators, they underwent considerable revision during these training and pilot-testing phases based on responses from Senegalese mothers and other Wolof language experts. The iterative training and translation processes used here were time- and labor-intensive, but well worth the extra effort to enhance the ecological validity of both our interview procedures and our measures. These survey tools are available upon request in both French and Wolof.

Language measures

Three types of measures of children's language proficiency analyzed here were collected to assess children's language skill: (1) Milestones Checklist, a caregiver-report checklist of language milestones achieved by children in both younger and older cohorts of children, aged 0;4 to 2;6; (2) Vocabulary Inventory, a caregiver-report measure of expressive vocabulary knowledge for an older cohort of children aged 1;8 to 2;6, and (3) Naturalistic Observation, a direct measure of child speech production obtained from audio- and video-recordings of a 5-min caregiver–child interaction when the children were 0;4 to 2;6.

Milestones Checklist

The Milestones Checklist was designed to assess early communication ability across the full range of ages of children in the study (0;4–2;6). For each skill in the checklist, caregivers were asked if their child typically responded in a certain way (e.g., turning her head to the sound of a familiar voice) or was able to perform a specific activity (e.g., following a verbal instruction such as "bring me the spoon"). This type of milestones checklist is often used as a screening tool for developmental delay, but our goal was to create a continuous measure with items ordered by increasing level of difficulty or skill. With a single instrument, we aimed to assess a wide age range of children, while avoiding floor or ceiling effects for low- or high-achieving children.

To develop the Milestones Checklist, we first created an English-version checklist of 52 items to ask the caregiver about language skills in infants and young children that were increasingly more advanced. We based the instrument on similar measures used in the US to assess children's communication skills, such as the LENA™ Developmental Snapshot (Gilkerson & Richards, 2008) and the Ages and Stages Questionnaire (Squires, Potter, & Bricker, 1995). The checklist was translated into Wolof and adapted for use in the Senegalese context of rural villages in Kaolack with the help of local experts in child development and the Wolof language. Adaptations included modifying descriptions for correct use of grammar in Wolof, such as the use of determiners and plurals. In a pilot study in three Wolof-speaking villages, caregivers for 20 children (0;2–1;10) were administered the preliminary checklist. For each item, mothers were asked to report whether or not their child demonstrated the particular skill, and their response was scored either 0 (no) or 1 (yes). Three items were dropped due to parent confusion and the difficulty of explaining the skill. For example, when asked "Does your child bring toys or objects to his/her mouth?", respondents thought we wanted to know whether their children put 'dirty' things in their mouths, so they consistently said no, their child does not do that. Four additional items were removed to shorten the questionnaire because they were redundant with information captured by the vocabulary inventory (e.g., "Does your child say at least 10 distinct words?"). Two items related to naming objects in books were also removed because books are rare in rural village households. Finally, five items thought to be too advanced for our sample were cut (e.g., "Can your child retell a story or event with a beginning, middle, and end without using pictures?"). The remaining 38 milestone items were ordered by difficulty in the instrument in the same order as administered in similar instruments in the US. However, it is highly probable that Wolof children achieve certain milestones earlier or later than expected for US children due to differences in their language environment. For example, knowing the names of colors might come later for Wolof children because

https://doi.org/10.1017/S0305000917000605 Published online by Cambridge University Press

these children do not have colorful books and toys to play with, and caregivers do not frequently name the color of natural objects for their children.

Administration procedures for the Milestones Checklist were adapted for use with this population. In US-validated tests, age of the child is used to determine the first item to administer, with administrations for younger children starting at or near the beginning of the list and administrations for older children beginning later in the list. In addition, administration typically is halted after the caregiver provides a certain number of 'no' responses in a row. Because the developmental ordering of the items was not expected to be identical for Wolof-learning children and English-learning children from the US, we started all caregivers with the first item. In addition, we implemented a generous stopping rule of six 'no' responses in a row to reduce the chances that age-appropriate items would fail to be administered. A raw score total was calculated as the sum of 'yes' responses over all possible items administered. Items not administered after the stopping rule were given a score of 0.

Vocabulary Inventory

For the Vocabulary Inventory, we adapted the short-form of the *MacArthur-Bates Communicative Development Inventory* (CDI), *Words & Sentences* (Level II; Fenson *et al.*, 2007), one of the most widely used parent-report instruments worldwide for assessing expressive vocabulary in children under three years of age (e.g., Alcock *et al.*, 2015; Hamadani *et al.*, 2010; Jackson-Maldonado, Marchman, & Fernald, 2013). This version asks parents to report on the child's ability to produce single words (from a list, or inventory, of words) and to combine words into phrases. In our study, we focused only on vocabulary production of children 1;8 to 2;6.

The development of a Wolof CDI was a multi-phase process for which we obtained the approval of the CDI Advisory board. A pilot version of the Wolof CDI was developed based on an existing version for use with the languages of Krobo, Ewe, and Twi, which are spoken primarily in southern regions of Ghana (Prado *et al.*, 2016). In consultation with Wolof-speaking colleagues, we examined each item for cultural and linguistic relevance to Wolof, and substituted those items that were not appropriate to our study population, avoiding items that were homonyms or which were highly variable across dialects. This phase culminated in the development of a final 105-item short form. Following guidelines from the MacArthur-Bates Advisory Board, the final set of items reflected a distribution of approximately 20% easy, 60% middle, and 20% hard items for children aged 1;8 to 2;6 living in rural villages.

When administering the CDI in interview format, caregivers were asked to report if their child could "both understand and say" the word. Because children often develop their own idiosyncratic pronunciations of words for common objects, the guidelines for administering the CDI allow caregivers to report idiosyncratic or child-like pronunciations of such words (e.g., *wow-wow* for *xaj* or 'dog'). When caregivers reported a positive response, they were asked to provide examples of specific times when their child used that word. These probes served to make sure that the caregivers understood the instructions. Raw score totals were calculated as the sum of CDI words that the caregiver reported that the child "understands and says" out of the 105 words in the list.

Naturalistic observation

All children were also video-recorded in a structured play session with their primary caregiver. In each village, recordings took place in the same testing locale, which was as private, quiet, and well lit as possible. The caregiver and child sat on a floor-mat

https://doi.org/10.1017/S0305000917000605 Published online by Cambridge University Press

and household objects were provided as toys. A microphone was attached to the caregiver's clothing and a video camera set up on a tripod about 5 to 6 feet from the dyad. The caregiver was asked to engage with her child just as she would at home, and was then left undisturbed in the room for 15 min.

Due to financial and time constraints, the transcription of speech produced by caregiver and child during the play session was limited to the middle 5 min of each recording. Following CHILDES conventions (MacWhinney, 2000), all word tokens and utterances in these 5-min segments were transcribed by Senegalese research assistants, who were trained in Wolof orthography and transcription protocols by a Wolof linguist. Every transcription was then reviewed by one of our gold-standard transcribers for consistency of spelling and parsing of utterances. Transcription was difficult because the children were very young and did not always articulate words clearly. In addition, background noise from nearby conversations among adults could not be entirely eliminated in the village field conditions, making it difficult in some recordings to hear exactly what the child said. Word tokens that could not be accurately transcribed were marked as unintelligible, but were still included in the word and utterance counts. Since we were working with short language samples under difficult conditions, we did not discard words or utterances that were unclear, as this would have underestimated children's productive abilities, thereby introducing bias for the youngest children. Each complete transcript was then processed using CLAN (MacWhinney, 2000) to obtain three measures of child speech: number of utterances, number of word tokens, and mean length of utterance-words (MLU-w), calculated as the total number of word tokens divided by the number of utterances. It is generally standard practice to annotate relations across morphological variants of word tokens in order to also derive word types (e.g., the inflected forms *walked* and *walking* would be considered two tokens of the word type *walk*). We were unable to estimate number of word types because limited resources precluded adequate training on Wolof grammar that would be required to develop a morphology for assigning tokens to types. Thus, our measure of word tokens reflects the total number of word tokens produced by each child during the language sample. While the use of tokens will inflate the estimates of the number of different words that the children produced, the rank order of the children will likely remain the same.

Results

To evaluate the quality and performance of the Milestones Checklist and Vocabulary Inventory, we first looked at the psychometric properties of these two instruments, or how well they measured language skill for young children living in Wolof-speaking villages in rural Senegal. In a second set of analyses, we then explored several forms of validity of the two measures. Finally, we asked whether the two instruments made independent, as well as overlapping, contributions to predicting children's productive speech during the language sample. Summary descriptive statistics for the language measures obtained from caregivers' reports and from language samples are presented in Table 1 for the entire sample, and stratified by younger and older age cohorts.

Psychometric properties of caregiver-report measures

The exploration of the psychometric properties of the two caregiver-report measures of language skill is presented here in three stages. In the first stage, we evaluated the overall

https://doi.org/10.1017/S0305000917000605 Published online by Cambridge University Press

Table 1. Descriptive statistics (*M* (*SD, min-max*)) on language measures for the cohorts of younger and older children and the full sample

	Younger cohort (n = 255)	Older cohort (n = 245)	Full Sample (N = 500)
Age (months)	0;11 (0;4, 0;4–2;6)	2;0 (0;2, 1;8–2;6)	1;6 (0;7, 0;4–2;6)
Language Milestones[a]	13.1 (3.8, 4–30)	20.6 (5.4, 8–38)	16.8 (6, 4–38)
Expressive vocabulary[b]	–	48.2 (28.6, 1–105)	–
Utterances[c]	8.5 (14.9, 0–80)	47.4 (32.9, 0–150)	27.5 (31.9, 0–150)
Word tokens[c]	10.1 (19.5, 0–182)	62.1 (50, 0–233)	35.5 (45.7, 0–233)
MLU-w[c]	1.17 (0.39, 1–4)	1.25 (0.37, 1–5)	1.22 (0.38, 1–5)

Notes. [a] Number of Language Milestones achieved reported on the Wolof Milestones Checklist; [b] Number of words reported as produced on the Wolof Vocabulary Inventory; [c] Number of child utterances, word tokens, and mean length of utterance (MLU-w) produced during a 5-min caregiver–child free-play session.

instrument performance, to determine how well the items worked together (i.e., by testing internal consistency and reliability) and also whether the range of item difficulties was well-matched to the children's ability, given that better matching allows for more precise estimates of ability. Second, we evaluated the performance of each item by asking how well responses to the items fit a statistical model – for example, whether individual items were useful in discriminating between low- and high-ability children. In the third stage, we used the item response model to test for possible bias from differential item function by child gender, an important concern in the development of a gender-fair instrument.

For the analyses in the first and second stages, we applied the simplest model in item response modeling, known as the Rasch model, a simple logistic model for which an observed response is a function of the difference between person ability and item difficulty (Wilson *et al.*, 2006). The model provides an estimate of each person's ability and each item difficulty (with standard errors) in the same log-odds units (logit). ACER ConQuest version 2.0 (Wu, Adams, Wilson, & Haldane, 2007) was used for the Rasch modeling, with the mean of the item difficulty parameters constrained to zero for identification and maximum likelihood used for estimation.

Instrument performance

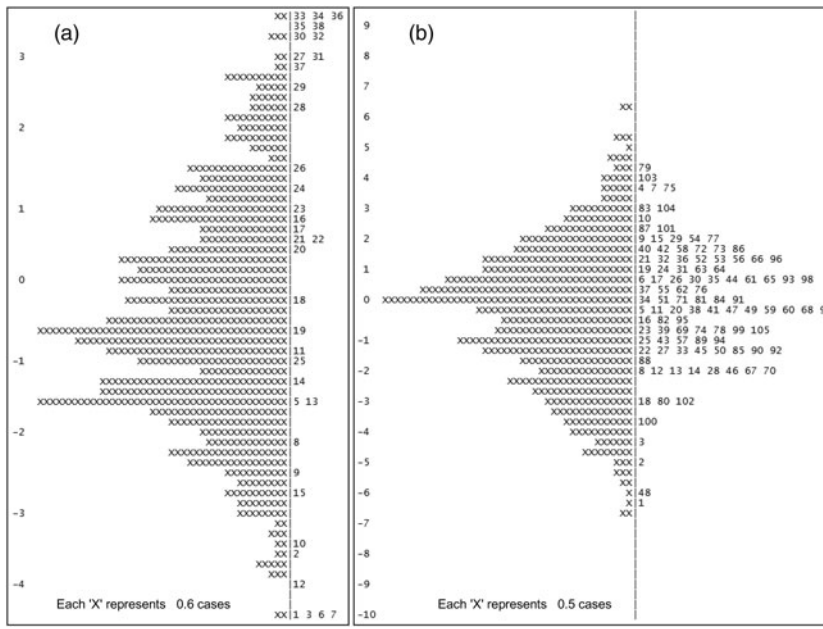
To test the overall performance of the two caregiver-report instruments, we used two indicators of reliability: Cronbach's alpha (Cronbach, 1951) for internal consistency (i.e., how well the items work together), and person-separation reliability (i.e., how well test scores 'separate' children of low and high ability). Person-separation reliability was calculated from the Rasch model as the difference in observed and unexplained variance in estimated abilities, divided by the variance explained by the model. A higher value for person-separation reliability indicates more precise estimates of ability, allowing for better distinction among persons – a desirable property for the instrument. We expected internal consistency to be comparable to person-separation reliability, and in fact we found both to be excellent for the Milestones Checklist at .86 and .88, respectively, as well as for the Vocabulary Inventory at .98 and .91, respectively.

https://doi.org/10.1017/S0305000917000605 Published online by Cambridge University Press

To examine how well the range of item difficulties for the instruments was matched to children's ability in our sample, we plotted the distribution of estimated item difficulties against the distribution of person abilities on the same logit scale, using what is referred to as a Wright map. As shown in Figure 1, the left-hand side of the map is a histogram of Xs illustrating the distribution of ability estimates from low (at the bottom of the map) to high, where each X represents a subset of children. Similarly, the items are enumerated on the right side of the map from easiest (at the bottom) to hardest, and are mapped to ability at the point where a person has a 50% chance of succeeding on the item. Children with abilities above that point have a greater than 50% chance of succeeding on the item, and persons below have less than a 50% chance.

Rasch model estimates of item difficulty on the Milestones Checklist (right side of vertical line) and child ability (left side of vertical line) are represented in Figure 1a. The Wright map shows that item difficulties generally followed the order as designed, with the easiest items at the bottom (lowest item number and logit value) and the more difficult items at the top (highest item number and logit value), although there were exceptions (e.g., item 5 was as difficult as item 13). The difficulty for item 4 ("Does your child laugh?") was not estimated because all caregivers responded 'yes' to the question, and some variation is necessary for estimation. However, the distribution of person abilities indicates that ability is approximately normally distributed, and the item difficulties adequately span the range of the sample children's ability.

Rasch model estimates of item difficulty and person ability on the Vocabulary Inventory for 240 children are represented in Figure 1b. Unlike the Milestones Checklist, the items on the Vocabulary Inventory were not ordered by difficulty:



Figures 1a and b. Wright maps of item difficulties and person abilities for (a) Milestones Checklist and (b) Vocabulary Inventory.

https://doi.org/10.1017/S0305000917000605 Published online by Cambridge University Press

caregivers reported on all 105 words in the list. From the Wright map in Figure 1b we can see that the bulk of the items are concentrated in the middle of the distribution of child estimated ability, with perhaps too few items to assess children at the low and high ends of the scale.

Item performance

Next we checked for individual item performance of the two caregiver-report measures by evaluating the fit of each item to the Rasch model. Items that fit the model will have an increasing probability of a successful response, with increasing language skill in a smooth S-shaped, or logistic growth curve (see examples in Figure 2). Deviations from the model were assessed both visually and with a fit statistic for each item, referred to as the weighted mean square (MNSQ) fit, or infit (Wilson, 2005) (the ratio of the variances of the observed over expected residuals from the model). An infit equal to one for an item indicates that deviations from the model vary as much as would be expected by chance. Infit values above one denote positive misfit, or more variation than expected, with a response pattern that is flat with respect to ability (i.e., responses are random). Infit values of less than one denote negative misfit, or less variation than expected, with a steep transition from low to high probability of success for an item (i.e., the item is highly discriminating for ability). Some deviation from one is expected due to random error. We reviewed model fit for items with infit outside of the range of 0.75 and 1.33 (3/4 and 4/3) (Wilson, 2005). We used a large *t*-statistic (based on a transformation of the infit into a standard normal distribution) to assess whether evidence of poor fit was statistically significant. For the visual inspection of fit, we used item characteristic curves (ICC) which are plots of the probability of a correct response as a function of ability. Deviations of the empirical ICC (based on the observed data) from the modeled ICC are indicative of a lack of fit of the item. We explore possible reasons to explain poorly fitting items.

The weighted MNSQ (infit) for all but seven of the Milestones items were within the acceptable boundaries of 0.75 and 1.33. Four items evidenced statistically significant negative infit (<0.75), with less variation than expected, suggesting a more discriminating item. Three of these items were related to object naming, specifically:

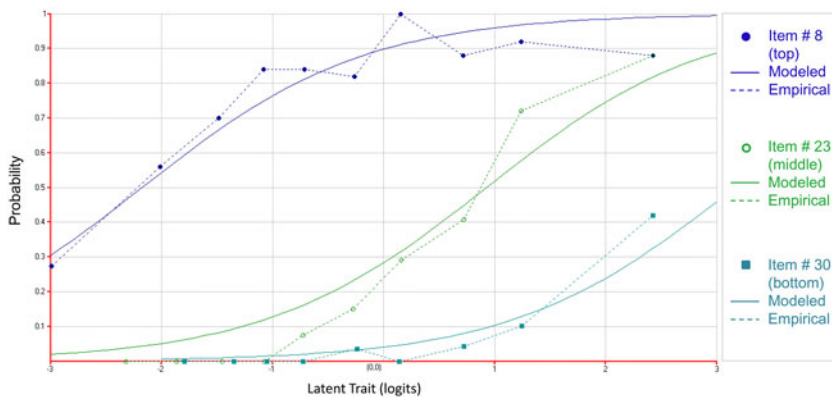


Figure 2. Item characteristic curves for milestone items with good fit at low (item #8, top lines), medium (item #23, middle lines), and high (item #30, bottom lines) difficulty levels.

https://doi.org/10.1017/S0305000917000605 Published online by Cambridge University Press

retrieving a named object (#13), pointing to a named object (#14), or naming a familiar object (#21). In general, items exhibiting significant negative infit contribute less to distinguishing among children, but are not a serious concern.

Three items evidenced statistically significant positive infit (>1.33), with more variation than expected, suggesting random responses from caregivers. All three positive infit items were related to early child vocalizations, such as: produces 2 or more vowel sounds (#5) or repeats the same sound twice in a row (#9). Caregivers may have been unable to accurately report on these items if they were not accustomed to identifying such sounds.

In Figure 2, we show ICC for three Milestones items with good fit at low (item #8, top lines), medium (item #23, middle lines), and high (item #30, bottom lines) difficulty levels. The plots demonstrate that the empirical ICCs (dotted lines) based on the observed data follow the modeled ICCs (solid lines) closely. In Figure 3, the ICC for item #13 is shown as an example of negative infit, and we see that the empirical ICC deviates from the model with a steeper transition from low to high probability of success on the item. The ICC for item #5 is shown as an example of positive infit in Figure 3, and we see that the empirical ICC deviates from the model with nearly equal probability of success on the item across the range of person ability (flat line).

The weighted mean square fit statistics (infit) for 90 of the vocabulary words were within the acceptable boundaries of 0.75 and 1.33. Three more were outside these bounds, but the lack of fit was not statistically meaningful. Six of the vocabulary items demonstrated statistically significant negative infit (<0.75), indicating higher discrimination. Of these items, four were body parts: eyes, mouth, head, and stomach. Another six items demonstrated statistically significant positive infit (>1.33), where the three poorest fitting items were: 'animal' (#26), 'this' (#39), and 'bye bye' (#81). The poor fit for 'animal', or *Bàyyima* in Wolof, may be explained by the fact that the word is pronounced the same way as the expression *Bàyyi ma*, which means 'Leave me alone'. For the demonstrative 'this', the interviewer may have emphasized the action of pointing instead of explaining the question, in which case mothers would have responded randomly. Finally, the practice of teaching children to say 'bye bye' may differ by family context and not be related to age or vocabulary size.

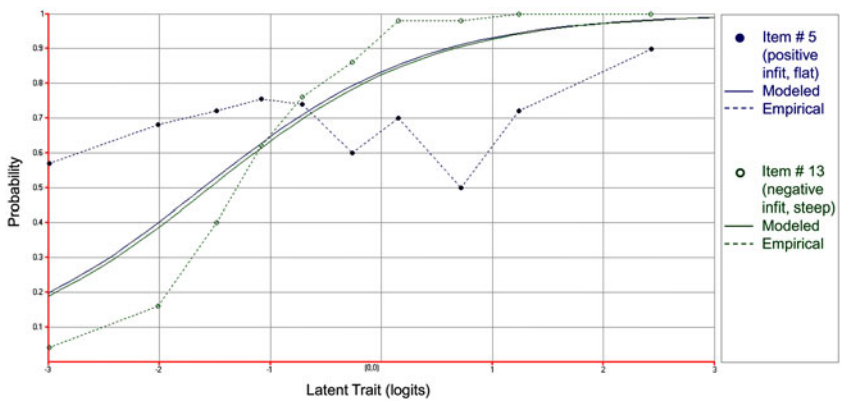


Figure 3. Item characteristic curves for milestone items #13 with negative infit (steep transition), or less variation than expected, and #5 with positive infit (flat line), or more variation than expected.

https://doi.org/10.1017/S0305000917000605 Published online by Cambridge University Press

In summary of the item performance, most Milestone and Vocabulary items had either good fit to the Rasch model or demonstrated acceptable fit but were more discriminating than expected, which is not a serious concern. For the Milestones Checklist, caregivers appeared to respond randomly to three early items related to infant vocalizations. However, it is important to keep these items as there are no clear replacement options for language skills in this early period of development. Instead, in the future, we recommend additional training of the interviewers to minimize confusion on the part of the respondent, and the use of audio-recording prompts during the interview to improve the performance of milestones items about infant babbling. For the Vocabulary Inventory, three items were particularly problematic: 'animal', 'this', and 'good-bye'. In future studies of Wolof, these three items should be removed or replaced with items of similar difficulty because they are adding random noise to the data.

Differential item function by gender

As a final evaluation of the psychometric properties of the caregiver-report measures, we tested for differential item function by child gender. In order to evaluate the Wolof language instruments for item-level gender bias, we added two terms to the Rasch model: a term for gender group membership and an interaction term between each item and the group (for DIF). We estimated DIF effect size from the interaction term for each item and considered it statistically significant if the magnitude was greater than 1.96 times the standard error of the DIF effect estimate. Items were considered to be exhibiting DIF if the DIF was found to be both statistically significant and the effect size was large (>.638 logits, based on the system developed and used by the Educational Testing Service (Paek & Wilson, 2011)).

For the Milestones, two items exhibited large and statistically significant DIF by gender: boys had a higher probability than girls of the same ability to be credited for the item 'vocalizes to attract attention', and girls were more likely than boys of the same ability to be credited for the item 'indicates past tense when speaking'. The absolute value of the magnitude of the DIF effect was about 1 logit in both cases (where 1 logit is equivalent to about .5 SD of the distribution of milestones ability in this sample). Added together, the net gender DIF effect from these two milestones items was near zero, such that the milestones instrument as a whole was unbiased with respect to gender.

For the Vocabulary Checklist, 18 items exhibited large and statistically significant DIF by gender. Six items favored boys: 'father', 'horse', 'dog', 'ball', 'more', and 'drink' (the verb). Twelve items favored girls, including four food items: 'fish', 'carrot', 'onion', and 'rice'; four words related to social etiquette: 'yes', 'no', 'thank you', and 'bye bye'; and four others: 'bowl', 'hair', 'cry', and 'wake-up'. The absolute value of the magnitude of the DIF effect was quite large (>1 logit per item, where 1 logit is equivalent to about .4 SD of the distribution of vocabulary knowledge) for four of the items favoring boys, whereas the magnitude of the DIF effect was generally smaller for the items favoring girls (<1 logit per item). However, added together, the net gender DIF effect favored girls.

We rely here on the item statistics from the Rasch model to identify items that behaved differentially for boys and girls with the same estimated underlying ability. However, the model is not very useful for determining why those items are a problem. We cannot know if a gender difference in item performance was a true difference in girls' or boys' knowledge and production of the word, or whether it

https://doi.org/10.1017/S0305000917000605 Published online by Cambridge University Press

represents a difference in parents' perception of girls' vs. boys' knowledge. Cognitive interviews with caregivers or more extensive evaluation of children's productive language would be required to disentangle these two possibilities. In future studies, we recommend replacing or balancing the number of such gender-biased items.

Validity of reported measures

The goal of the second analyses was to explore the validity of the two caregiver-report measures in a rural Wolof-speaking population. We first tested the strength of the association between the two measures. Although the measures were designed to assess different subdomains of language, we expected that children with a larger vocabulary would also be using more complex grammar, and that the two measures would be highly correlated. However, language skill is also strongly correlated with age – increasing rapidly during this early period of child development – such that age could be a confounder of the estimated association between the two instruments. Therefore we also verified that the measures remained correlated when controlling for the age effect.

Using Pearson's correlation, vocabulary size was found to be strongly correlated with scores on the milestones achieved in the older cohort ($r = .79$, $p < .001$), for whom both measures were assessed. In addition, language ability increased with increasing age of the child, as we would expect. The number of milestones achieved for all 500 children was positively and strongly associated with age with a correlation of .73 ($p < .001$). This correlation was smaller for the subset of older children ($r = .35$, $p < .001$) than for the full sample. Reported productive vocabulary size for 240 children in the older cohort also increased with increasing age ($r = .46$, $p < .001$). The inter-correlation of the two measures remains high ($r_p = .76$, $p < .001$) when removing the effect of age. There was no evidence of a floor or ceiling effect for either measure.

Another question of interest was whether the language scores were associated with two indicators of household socioeconomic status: median household education and household wealth (a composite score of dwelling characteristics, assets, and animal ownership). The evidence for a socioeconomic effect on child developmental outcomes is strong in both high-income (Currie, 2009; Shonkoff, Boyce, & McEwen, 2009) and low-income (Currie & Vogl, 2013; Fernald, Kariger, Hidrobo, & Gertler, 2012) settings. Even in low-income contexts, children living in households with relatively more wealth and better educated parents have been shown to perform better on developmental assessments than children living in households with relatively less wealth and less parental education (Fernald, Weber, Galasso, & Ratsifandrihamanana, 2011). We therefore wanted to evaluate this expected association in our sample of children, all of whom were living in poor rural villages in Senegal.

Mean adult education was significantly, but weakly, correlated with scores on both the Milestones Checklist and Vocabulary Inventory when adjusting for child age ($r_p = .19$ and .20, respectively, with both ps <.01). The household wealth index was not significantly associated with either caregiver-report measure, a surprising finding, but which may be due to the relative homogeneity in the measure of wealth used in our sample as compared to the heterogeneity in how children learn language at a given age. Alternatively, it could be that education level of adults in the household is a better predictor of early language development than is household wealth among young children in a rural Senegalese context, as it is in the US (Blau, 1999; Dearing, McCartney, & Taylor, 2001).

https://doi.org/10.1017/S0305000917000605 Published online by Cambridge University Press

As a final check of validity, we asked how strongly the caregiver-report measures were correlated with children's language level as assessed using a direct measure of child productions during a naturalistic language sample. Other studies have found moderate to high correlations (ranging from .55 to .65) between language sample measures and adaptations of the CDI: Words and Sentences (Thal, Jackson-Maldonado, & Acosta, 2000; Alcock *et al.*, 2015). Correspondingly, we expected that children with higher scores on the Wolof Milestones Checklist and Vocabulary Inventory would also produce more language in a naturalistic sample.

Restricting our analyses to only the older cohort of children for whom both caregiver-report and language sample data were available, both caregiver-report measures were significantly correlated with the direct measures of language obtained from the play session. Milestones was significantly, albeit moderately, correlated with the number of utterances ($r = .23$), word tokens ($r = .29$), and MLU-w ($r = .30$). After adjusting for age, the correlations between Milestones and the language sample measures dropped to .16, .20, and .22 for utterances, word tokens, and MLU-w, respectively, but remained significant ($ps < .01$). The correlations between the language sample measures and scores on the Vocabulary Inventory were somewhat stronger than for the Milestones measure: utterances ($r = .35$), word tokens ($r = .41$) and MLU-w ($r = .35$). After adjusting for age, the correlations again dropped to .27, .30, and .25 for utterances, word tokens, and MLU-w, respectively.

Prediction of language sample

In this last analysis, the goal was to test if the two caregiver-report instruments make independent, as well as overlapping, contributions to predicting the number of word tokens and utterances that children produce during the naturalistic language sample. Was there added value in administering two caregiver-report measures over just one? To answer this question, we performed a series of linear regressions, controlling for child age, gender, median household education, and household wealth, and adjusted for clustering at the village level. We report the R-squared for the percentage of variance in the outcome measures (utterances, word tokens, and MLU-w), which can be explained by the factors included in the models (demographics and caregiver-report measures). We use the likelihood-ratio (LR) test to determine if the addition of one or both of the caregiver-report measures significantly improves the model fit (i.e., the *p*-value associated with the LR test statistic is <.05). We restrict these analyses to the older cohort of children with all non-missing language and demographic data ($n = 216$).

In Table 2, we first look at the predictors of the number of utterances produced by the child during the language sample. Demographics (age, gender, maternal education, household wealth) explained about 11% of the variance in the number of utterances produced by the child. Adding either of the caregiver-report measures improved the model fit significantly. The addition of the Milestones to the model increased the R^2 to 13.5%, whereas the addition of Vocabulary Inventory scores increased the percentage to 18% of the variance in utterances. Adding both Milestones and Vocabulary improved the model significantly over adding the Milestones alone (LR test $p < .0001$), but did not improve the model over adding Vocabulary alone (LR test $p = .16$).

Next, Table 2 shows that demographics explained about 16% of the variance in the number of word tokens produced by the child. The addition of the Milestones score to the model increased the R^2 to 19%, and the addition of the Vocabulary score increased

https://doi.org/10.1017/S0305000917000605 Published online by Cambridge University Press

Table 2. Proportion of variance in children's spontaneous speech (number of utterances, word tokens, and MLU-w) explained by two caregiver-report measures of children's language skill

Outcome: number of utterances produced in 5-min play session			
Model	Adjusted R^2	Change in R^2 from demographics only	Change in R^2 from demographics + other language measure
Demographics[a] only	0.114	–	–
Demographics + Milestones[b]	0.135	0.021**	0.049***
Demographics + Vocabulary[c]	0.184	0.070***	0.008
Demographics + Milestones + Vocabulary	0.192	0.078***	–
Outcome: number of words produced in 5-min play session			
Model	Adjusted R^2	Change in R^2 from demographics only	Change in R^2 from demographics + other language measure
Demographics[a] only	0.157	–	–
Demographics + Milestones[b]	0.189	0.032***	0.051***
Demographics + Vocabulary[c]	0.237	0.080***	0.003
Demographics + Milestones + Vocabulary	0.24	0.083***	–
Outcome: MLU-w in 5-min play session			
Model	Adjusted R^2	Change in R^2 from demographics only	Change in R^2 from demographics + other language measure
Demographics[a] only	0.094	–	–
Demographics + Milestones[b]	0.135	0.041***	0.015*
Demographics + Vocabulary[c]	0.148	0.054***	0.002
Demographics + Milestones + Vocabulary	0.15	0.056***	–

Notes. * $p < .1$; ** $p < .05$; *** $p < .01$; [a] Child age, gender, median household education, and household wealth index; [b] Number of language milestones achieved reported on the Wolof Milestones Checklist; [c] Number of words reported as produced on the Wolof Vocabulary Inventory.

https://doi.org/10.1017/S0305000917000605 Published online by Cambridge University Press

the percentage to 24% of the variance in word tokens. As with utterances, adding both reported measures improved the model significantly over adding the Milestones alone (LR test $p = .0002$), but not over adding Vocabulary alone (LR test $p = .34$), suggesting that the vocabulary measure did a better job of explaining number of utterances and word tokens produced in the language sample.

Finally, for MLU-w, demographics explained about 9% of the variance. Once again, adding either caregiver-report measure improved the model over demographics alone. However, adding both caregiver-report measures did not improve the model significantly at the 5% level over adding either report measure alone, suggesting that the measures perform comparably in explaining MLU-w in the language sample.

How can these results be explained? Given that the Vocabulary Inventory was designed specifically to capture a representative sample of words that Wolof-learning children in the target age range can say, it might not be surprising that this instrument was somewhat more strongly related to counts of number of utterances and word tokens than the Milestones Checklist. Recall also that the Vocabulary Inventory had greater internal consistency than the Milestones Checklist, another factor that could account for the somewhat stronger correlations. In contrast, the Milestones Checklist was designed to capture a broader spectrum of children's language skill, including their use of grammar and certain types of words, as well as their ability to follow verbal instructions, and thus may be less directly related to utterance and word token counts. However, as the high correlation between the two measures suggests, those children who knew more vocabulary words are also those who used better grammar and produced longer, more complex sentences (e.g., types of items in the milestones checklist). This is consistent with many studies reporting a high correlation between children's vocabulary development and use of more complex morphosyntactic forms (e.g., Bates & Goodman, 2001).

Discussion

The development of reliable assessment instruments is critical to furthering our understanding of the differences and commonalities of how children learn language around the world. Because no such instruments existed for studying language development by Wolof-learning children in the rural Kaolack region of Senegal, we developed two caregiver-report instruments that have been successfully adapted for use in other low-income settings: (a) Milestones Checklist: a checklist of language and communication skills, or language milestones, listed in order of increasing difficulty; and (b) Vocabulary Inventory: an inventory of vocabulary words of varying levels of difficulty that are typically produced by young Wolof-learning children. In this report, we show that both caregiver-report instruments were psychometrically robust and valid for use in rural Senegal, although minor modifications are recommended for future work. Both instruments showed expected relations with age, child gender, and some indices of family socioeconomic status.

The use of video-taped play sessions was a strength of this study, as it provided a direct measure of children's language ability with which to complement the caregiver-report measures, without requiring a highly trained assessor to extract speech from shy or reticent children. The Vocabulary Inventory was more strongly related to the number of utterances and word tokens produced by the child during a naturalistic play session than was the Milestones Checklist. However, both measures explained a significant proportion of variance in child MLU-w.

https://doi.org/10.1017/S0305000917000605 Published online by Cambridge University Press

Nevertheless, the correlations between the caregiver reports and the language sample measures were low relative to those reported for adaptations of the CDI: Words and Sentences in Spanish (Thal *et al.*, 2000) and two Kenyan languages (Alcock *et al.*, 2015). The correlations reported for these other studies were based on a much smaller number of children (10–20) for whom longer language samples (30–60 min) were recorded and transcribed. In this study, the use of number of tokens (not types) from a short 5-min language sample in a single setting is one likely explanation of the generally weaker correlations that we found. It is also possible that the free-play context that we provided was not familiar to these caregivers and their children, limiting the representativeness of the language samples that were obtained.

We cannot rate the relative benefit of using the caregiver-report measures with other types of direct behavioral assessments, such as the *Bayley Scales of Infant Development* (Bayley, 2006). These assessments typically involve more time commitment and training of personnel than were involved in collecting the naturalistic language samples used here. In many cultures and low-income settings, children can be very shy about talking in the presence of a stranger, which can result in censoring of the data due to non-response. Caregiver-report measures are not subject to this particular bias. Given sufficient resources, future studies should continue to include some type of direct observations in their protocols to further explore the conditions under which these methods will yield information that is more consistent with the caregiver-report measures.

Despite the limitations of the direct measure of child language used here, the short recordings provided a wealth of data beyond the child's spontaneous productive speech, including the number of words and utterances produced by the caregiver – data that were critical for the evaluation of Tostan's parenting program. In fact, results showed that caregivers in program villages nearly doubled the amount of child-directed speech during the play session as compared to baseline, while caregivers in matched comparison villages showed no change (Weber *et al.*, 2017).

The work from this project can inform future efforts to study languages spoken in challenging contexts, such as ones where caregivers are mostly illiterate and for which appropriate measures do not currently exist. Here we discuss some of the issues to consider in choosing to adapt one or both instruments for a study of early child language development in such a situation. In particular, the development of the instruments was challenging and time-intensive in our study. Fortunately, we did not have to start from scratch: the initial set of candidate items for each instrument was based on pre-existing measures developed for other contexts and languages. However, careful translation and adaptation of the measures was still required, as described in the 'Method' section, and both instruments were piloted in the field together with a sample of about 30 caregivers and their data analyzed similarly. Additional updates to the items were made as issues were encountered and resolved during the enumerator training and fieldwork testing, which were performed together.

In terms of training, both instruments required literacy and fluency in the local language on the part of the interviewers, who were master's level students or graduates in sociology. They also required a clear understanding of the protocol for administration, and opportunity to practice and receive feedback. Training of the interviewers occurred over a two-week period that included work on improving their reading proficiency in Wolof and practice sessions in the field in the presence of a Wolof language expert and field supervisor. The administration of the Milestones Checklist was relatively easy to train and administer: questions were

read out loud and 'yes/no' responses recorded. In the event that the respondent did not understand the question, interviewers were provided with appropriate examples. The checklist took a maximum of 15 min if all items were administered, and shorter if the stopping rule was applied. We avoided the use of a starting rule, which other investigators might find necessary if the age range of their sample is very broad, but which is more complicated to train for and prone to errors in field administration.

In contrast, all words in the Wolof CDI were administered, which eliminated potential errors from the use of start and stopping rules, but also took longer: about 20 to 30 min. However, the Vocabulary Inventory instructions for what constitutes a 'yes' vs. 'no' response for a word that the child 'says' are more complex and context dependent than the Milestones Checklist. For example, caregivers are asked to distinguish between words that the child says on his/her own, rather than imitates, so that it is clear that the child knows what the word means. Credit is given for words that the child does not pronounce correctly and for words that mean the same thing as the words on the list, but that are specific to a certain child, dialect, or region. To improve the chances of receiving a valid 'yes' response, caregivers are regularly queried with: "What was your child doing when he/she said that?"

In studies using caregiver-report measures of child language ability across a wide range of ages, from preverbal to verbal children, there are advantages to using both types of caregiver-report formats, a Milestones Checklist and a Vocabulary Inventory, in tandem. However, if time is limited to develop, adapt, and validate both instruments in a new context and language, investigators may prefer to choose just one of the two caregiver-report formats. This choice would primarily be guided by the research objectives, but also by the age range of children in the study. The Vocabulary Inventory was more predictive of our direct measure of children's productive language, but it is inappropriate for preverbal children. Additional emphasis should also be placed on assuring gender balance in the list of words. In contrast, the Milestones Checklist is applicable to preverbal as well as verbal children, but additional care is needed to assure caregiver understanding of all of the items. Items that involved caregivers' judgment of the level of sophistication of children's early babbling were more likely to result in random caregiver responses. Providing audio-recorded examples of children's early vocalizations may improve caregivers' understanding of the types of behaviors that are targeted by such items. In addition, if the Milestones Checklist is used by itself, items that report on the number or types of words that the child is able to produce (e.g., 'Does your child say at least 10 meaningful words?') should be added. Such items were in the English-language measures (Gilkerson & Richards, 2008; Squires *et al.*, 1995) that the Wolof-language measure was based on, but were excluded in the Wolof version because they were considered redundant with the information captured by the Vocabulary Inventory.

In sum, the current study provided evidence of strong psychometric properties of a Milestones Checklist and a Vocabulary Inventory, two methods that capture information regarding the early language development of children in Wolof-speaking populations in Senegal. The adaptation of the instruments for this population was a complex process in which cultural and linguistic factors were taken into account to improve the validity of the measures in the local context. Moreover, it was critical to provide extensive training to ensure and maintain consistent administration

protocols. Future studies should continue to explore methods of administration which would further enhance the reliability and validity of the measures. In spite of limited literacy in this population, the caregiver-report method is likely to offer a valid alternative to more costly tools for assessing early language development in rural African communities.

Acknowledgements. This work would not have been possible without the hard work and many contributions of the study team members in Senegal and the Stanford University Language Learning Lab. We gratefully acknowledge the generous support of all those who made this research possible, including members of the non-profit organization Tostan, Dalberg Global Development Advisors in Dakar, and the caregivers, children, and other community members in Kaolack, Senegal. We also wish to thank the William and Flora Hewlett Foundation for their financial support for this study.

References

Alcock, K. J., Rimba, K., Holding, P., Kitsao-Wekulo, P., Abubakar, A., & Newton, C. R. J. C. (2015). Developmental inventories using illiterate parents as informants: Communicative Development Inventory (CDI) adaptation for two Kenyan languages. *Journal of Child Language*, *42*(4), 763–85.

Baranowski, T., Allen, D. D., Masse, L. C., & Wilson, M. (2006). Does participation in an intervention affect responses on self-report questionnaires? *Health Education Research*, *21*(Supplement 1), i98–i109.

Bates, E. & Goodman, J. C. (2001). On the inseparability of grammar and the lexicon: evidence from acquisition. In M. Tomasello & E. Bates (Eds.), *Language development: the essential readings* (pp. 134–62). Malden, MA: Blackwell Publishing Inc.

Bayley, N. (2006). *BSID-III: Bayley Scales of Infant Development*, 3rd edn. San Antonio, TX: Harcourt Assessment.

Blau, D. M. (1999). The effect of income on child development. *Review of Economics and Statistics*, *81*(2), 261–76.

Cislaghi, B. F., Gillespie, D., & Mackie, G. (2016). *Values deliberations and collective action: community empowerment in rural Senegal*. New York: Palgrave MacMillan.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Currie, J. (2009). Healthy, wealthy, and wise: socioeconomic status, poor health in childhood, and human capital development. *Journal of Economic Literature*, *47*(1), 87–122.

Currie, J., & Vogl, T. (2013). Early-life health and adult circumstances in developing countries. *Annual Review of Economics*, *5*, 1–36.

De Swaan, A. (2001). *Words of the world: the global language system*. Cambridge: Polity Press and Blackwell.

Dearing, E., McCartney, K., & Taylor, B. A. (2001). Change in family income-to-needs matters more for children with less. *Child development*, *72*(6), 1779–93.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., … Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428–46.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, S. J., & Bates, E. (2007). *MacArthur–Bates Communicative Development Inventories: user's guide and technical manual*, 2nd edn. Baltimore, MD: Brookes Publishing.

Fernald, L. C. H., Kariger, P., Engle, P. L., & Raikes, H. A. (2009). *Examining early child development in low-income countries*. Washington, DC: The World Bank.

Fernald, L. C. H., Kariger, P. K., Hidrobo, M., & Gertler, P. J. (2012). Socio-economic gradients in child development in very young children: evidence from India, Indonesia, Peru and Senegal. *Proceedings of the National Academy of Science* (*PNAS*), *109*(Supplement 2), 17273–80.

Fernald, L. C., Weber, A., Galasso, E., & Ratsifandrihamanana, L. (2011). Socioeconomic gradients and child development in a very low income population: evidence from Madagascar. *Developmental Science*, *14*(4), 832–47.

Gilkerson, J., & Richards, J. A. (2008). *The LENATM developmental snapshot* (Technical Report No. LTR-07-2). Boulder, CO: LENA Research Foundation.

Hamadani, J. D., Baker-Henningham, H., Tofail, F., Mehrin, F., Huda, S. N., & Grantham-McGregor, S. M. (2010). Validity and reliability of mothers' reports of language development in 1-year-old children in a large-scale survey in Bangladesh. *Food and Nutrition Bulletin*, *31*(2 Supplement 2), S198–S206.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: relation to language input and gender. *Developmental Psychology*, *27*(2), 236–48.

Jackson-Maldonado, D., Marchman, V. A., & Fernald, L. C. H. (2013). Short-form versions of the Spanish MacArthur–Bates Communicative Development Inventories. *Applied Psycholinguistics*, *34*, 837–68.

Kuhl, P., & Rivera-Gaxiola, M. (2008). Neural substrates of language acquisition. *Annual Review of Neuroscience*, *31*(1), 511–34.

MacWhinney, B. (2000). *The CHILDES Project: tools for analyzing talk*. Mahwah, NJ & London: Lawrence Erlbaum.

Marchman, V. A., & Fernald, A. (2008). Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. *Developmetal Science*, *11*(3), F9–F16.

Marchman, V. A., & Martínez-Sussmann, C. (2002). Concurrent validity of caregiver/parent report measures of language for children who are learning both English and Spanish. *Journal of Speech, Language, and Hearing Research*, *45*, 983–97.

Paek, I., & Wilson, M. (2011). Formulating the Rasch Differential Item Functioning Model under the marginal maximum likelihood estimation context and its comparison with Mantel-Haenszel Procedure in short test and small sample conditions. *Educational and Psychological Measurement*, *71* (6), 1023–46.

Pollitt, E., & Triana, N. (1999). Stability, predictive validity, and sensitivity of mental and motor development scales and pre-school cognitive tests among low-income children in developing countries. *Food and Nutrition Bulletin*, *20*(1), 45–52.

Prado, E. L., Adu-Afarwuah, S., Lartey, A., Ocansey, M., Ashorn, P., Vosti, S. A., & Dewey, K. G. (2016). Effects of pre- and post-natal lipid-based nutrient supplements on infant development in a randomized trial in Ghana. *Early Human Development*, *99*, 43–51.

Richman, A. L., Miller, P. M., & LeVine, R. A. (1992). Cultural and educational variations in maternal responsiveness. *Developmental Psychology*, *28*(4), 614–21.

Shonkoff, J. P., Boyce, W. T., & McEwen, B. S. (2009). Neuroscience, molecular biology, and the childhood roots of health disparities: building a new framework for health promotion and disease prevention. *Jama*, *301*(21), 2252–9.

Squires, J., Potter, L., & Bricker, D. (1995). *The ASQ user's guide for the Ages & Stages Questionnaires: a parent-completed, child-monitoring system*. Baltimore, MD: Paul H Brookes Publishing.

Thal, D., Jackson-Maldonado, D., & Acosta, D. (2000). Validity of a parent-report measure of vocabulary and grammar for Spanish-speaking toddlers. *Journal of Speech, Language, and Hearing Research*, *43*(5), 1087–100.

Tostan (n.d.). Online: http://www.tostan.org/program/reinforcement-parental-practices-module.

UNICEF (n.d.). Senegal – Statistics. Online: http://www.unicef.org/infobycountry/senegal_statistics.html#117 (last accessed 1 September 2011).

Weber, A., Fernald, A., & Diop, Y. (2017). When cultural norms discourage talking to babies: effectiveness of a parenting program in rural Senegal. *Child Development*, *88*(5), 1513–26.

Wilson, M. (2005). *Constructing measures: an item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, M., Allen, D. D., & Li, J. C. (2006). Improving measurement in health education and health behavior research using item response modeling: introducing item response modeling. *Health Education Research*, *21*(Supplement 1), i4–i18.

Wu, M. L., Adams, R., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest: generalized item response modeling software (version 2)*. Camberwell: ACER Press.

Zeitlin, M. (2011). *New information on West African traditional education and approaches to its modernization*. Dakar: Tostan.

Cite this article: Weber AM, Marchman VA, Diop Y, Fernald A (2018). Validity of caregiver-report measures of language skill for Wolof-learning infants and toddlers living in rural African villages. *Journal of Child Language* **45**, 939–958. https://doi.org/10.1017/S0305000917000605