Exploring Characteristics of Quality in Language Teaching & Learning: The Mother Tongue Adapted Coding Scheme (MACS)

Beth O'Brien[ab], Nurul Taqiah Yussof[a], Poorani Vijayakumar[a],
Malikka Begum Habib Mohamed[a], Leong Xuan En Rachel[c]

[a]Centre for Research in Child Development, National Institute of Education
[b]Centre for Research and Development in Learning,
Nanyang Technological University, Singapore
[c]Singapore University of Social Sciences

Correspondence should be directed to Beth A. O'Brien, beth.obrien@nie.edu.sg, (ORCHID #0000-0002-1187-5908), 1 Nanyang Walk, National Institute of Education, Nanyang Technological University, Singapore, 637616

Acknowledgement: This study was funded by the Education Research Funding Programme, National Institute of Education (NIE), Nanyang Technological University, Singapore, project no. OER 09/14RB. The views expressed in this paper are the author's and do not necessarily represent the views of NIE. The authors would like to thank the participating children, parents and schools.

Declarations of interest: None.

Abstract

Teacher-child interactions are an integral factor influencing the quality of early childhood education, and multilingualism is increasingly prevalent in many contexts. In the multilingual society of Singapore this is particularly relevant because early childhood classrooms follow a bilingual policy and include the teaching and learning of Mother Tongue languages. To evaluate what constitutes quality for preschool language teaching in this multilingual context, the Mother Tongue Adapted Coding Scheme (MACS) was developed to examine teacher-child interactions during blocks of Mother Tongue language lessons. The comprehensive observational rating scheme for language learning included four domains: language input, language output, varied teaching strategies, and English use for teaching Mother Tongue language. This rating scheme was applied to fifty-one classrooms where teachers were videotaped while they conducted Chinese, Malay, or Tamil language lessons. Within the observed classrooms, four hundred and ninety-one children were assessed on their Mother Tongue language learning over three years for receptive vocabulary, reading, and morphological awareness. Overall interrater reliability on the MACS was high across language classrooms (K > 0.72). Correlation and exploratory factor analyses indicated a main factor for the language input and output domains, and separate factors for English language use. Teachers' factor scores were found unrelated to their scores on the CLASS. Higher factor scores on input/output quality were correlated with having more high progress learners (compared to peers) for receptive vocabulary, but less high progress learners for reading. Implications for future applications and adaptations of the MACS are discussed in light of the use of rating tools to examine and better understand the constituents of ECE quality within multilingual contexts.

Introduction

Early childhood is a critical period of development, and children are spending an increasing amount of time in classrooms in their early years. Based on evidence that the quality of early childhood education (ECE) impacts children's development (e.g., Burchinal, Magnuson, Powell & Hong, 2015), there is heightening interest in understanding, assessing, and ensuring quality in ECE. Quality in ECE is currently conceptualized in terms of process quality involving caregiver-child interactions, and structural quality including teacher experience and characteristics of the ECE program (Burchinal, 2018). In the quest to measure such components of quality of the ECE environment there is also an awareness that either universal features or socio-culturally contextualized ones may play a role (Schwartz, 2018). This paper focuses on what constitutes quality of early bilingual education, and preschool classroom practices that best support language outcomes (Palviainen et al., 2016). As this study is situated in Singapore - a multilingual context with unique sociopolitical factors contributing to the convergence of diverse language competencies - this question is key yet unresolved. Here, as in many multilingual societies, the school and home language may differ for some students, creating a complex classroom environment in which individuals may vary widely on proficiency levels. This translates to a need for teachers to cater to different language competencies using a variety of instructional strategies, to provide comprehensible input for all (Jiang, Garcia & Willis, 2014; Probyn, 2015). Many of the currently established ECE quality rating systems were developed within Western contexts and may not be sensitive to these unique teaching and learning needs.

The multilingual society of Singapore includes a 'bilingual education' model, in which English is the main medium of instruction, but students must also demonstrate proficiency in a designated Mother Tongue language (MTL) by the end of primary school. The official MTLs are

Mandarin (Chinese), Behasa Melayu (Malay) or Tamil, and are assigned based on children's ethnicity, though these languages are currently used less frequently in homes. A proportion of class time is allocated for MTL lessons (which we refer to as the "target" language), and this so-called MTL at times does not match the actual home language (Dixon, 2009), which has become increasingly English-dominant (Bolton & Ng, 2014). Consequently, at primary school entry children are differentially prepared for the subject of MTL due to variation in home language exposure (Sun, Yin, Amsah & O'Brien, 2018). This places the preschool classroom in a pivotal role for children's early MTL learning, acquisition and school readiness.

Given this key period of preschool for child development, coupled with the specific socio-cultural demands and educational objectives in the Singapore context, we sought to examine characteristics of quality ECE regarding language learning within a bilingual context. We focus on MTL learning and teaching within the ECE context given the concerns with increasing English dominance. To this end, we constructed a measure of early education quality specific to MTL learning called the Mother Tongue Adapted Coding Scheme (MACS). We then applied the coding scheme to videotaped classroom language lessons. In sections of this paper we describe the process of developing the coding scheme, the content and rating scales of the coding scheme, and several approaches for examining the validity of the coding scheme.

Conceptual Framework

The MACS was developed as part of the Singapore Kindergarten Impact Project (SKIP, Ng et al., 2014), which is a large-scale longitudinal study of children's school readiness. A key research question of SKIP asks what characterizes quality preschool instruction, including language instruction. Through several phases (see Figure 1), the rating scheme was developed with a broad framework to be comprehensive, as an initial attempt to characterize the ECE environment in Singapore. We began with a review of existing observational tools for preschool and bilingual contexts (summarized in Appendix A) and extended these with consideration of local language planning and language-specific factors, and with characteristics based on language or second language learning theory. These sources contributed to the conceptualization of the MACS (Figure 2) across 4 broad domains – teachers' language *input*, children's elicited language *output*, *varied* pedagogical *strategies* for language learning, and *use of English* as bilingual children's stronger language. They are described in more detail in the following sections: (1) considering linguistic frameworks for the input and output strategies domains, (2) considering pedagogical frameworks for the varied strategies domain, (3) considering context-specific factors for the English use domain, and (4) overall adaptation from existing tools.

1. Considering linguistic frameworks. Language input and use are widely accepted as the basis for language acquisition (Unsworth, Persson, Prins & de Bot, 2015). This occurs ultimately through comprehensible input (i.e., the Input Hypothesis, Krashen, 1985), and so teachers' provision of simplified input is key for learner comprehension. At the same time, language use or 'output' is considered necessary for facilitating linguistic knowledge of the learner (i.e., the Output hypothesis, Swain, 1985), and for leading to further input from teachers (Pearson, 2007) (refer to Figure 2).

Input quantity and quality are consistently linked to bilingual children's language and literacy outcomes (Hoff & Core, 2013; Huang & Kuo, 2020). Optimal input stems from highly proficient speakers: input from native speakers in the home was found most beneficial for young children's development (Place & Hoff, 2011), and input from highly proficient speakers was reported as an integral factor in quality language instruction (Canh & Renandya, 2017; Richards, 2010). Such proficient input includes the qualities of correct usage (pronunciation, syntax and grammar) as well as linguistic richness (variety of words and sentence types) (Bornstein, Haynes & Painter, 1998). Quality input is especially important for languages with insufficient exposure (De Houwer, 2005).

Acquiring both language form (knowledge about organizing sounds, words and sentences) and content (meaning) knowledge are important for language learning (Lahey, 1988). Hence, language teaching may focus on either or both of these, as the target language may be the subject (teaching proper oral and written language forms) or the medium of instruction for teaching other content (discussing meaning and developing concepts) (Ellis & Shintani, 2014). For example, proficient teachers, who not only model oral and written language, but also elaborate and explain words and concepts, may increase children's learning and continued engagement (Pakarinen et al., 2010; La Paro, Pianta & Stuhlman, 2004).

Output is also considered key to language learning, including for bilingual vocabulary development and proficiency (e.g., Bohman, Bedore, Pena, Mendez-Perez & Gillam, 2010; Swain, 2005). Output allows learners to notice gaps in their knowledge (de Bot, 1996), to access meaning and to generalize, which can serve as building blocks for language representations (e.g., Ibbotson, 2013). Encouraging children to use the target language and providing feedback on their output enhances learning. This is particularly important because even with extended input, children's target language output often lags behind their listening comprehension (e.g., DePalma, 2010).

Schwartz and Gorbatt (2016) noted that bilingual preschool teachers need to create a low-anxiety, secure environment to encourage children's target language production. Thus, encouraging children to use the target language within a positive classroom climate was considered an important aspect of eliciting language output. Teachers' use of language modelling and scaffolding learner output was also considered essential (e.g., Gibbons, 2006).

2. Considering pedagogical frameworks. Besides the linguistics perspective, pedagogical factors were included in a set of *Varied Strategies* that teachers may use to further assist children's understanding of the target language. A socio-cultural perspective holds social interactions in the classroom as central to language acquisition and instruction (Vygotsky, 1978), and is considered in terms of the input-output cycle and the role of feedback for second language learning (Gass & Mackey, 2007). Since the current study focuses on preschoolers who may lack exposure to MTL input, or who may be learning the MTL for the first time, we considered different strategies recommended for beginning language learners (see Table 1), including adapted (child-directed) speech (Wesche, 1994), and non-verbal cues like gestures and facial expressions (Kersten, Steinlen, Tiefenthal, Wippermann & Matsson, 2010; Weitz, Pahl, Mattsson, Buyl, & Kalbe, 2010).

Another set of strategies of particular relevance to multilingual contexts is the use of the dominant or first language in target language teaching. Teachers may switch to children's dominant or stronger language (often English) for explaining and translating to ensure target language comprehension (Chimbutane, 2013; Lin, 2005), for managing classroom behaviors, and facilitating task completion (Enama, 2016; Gort & Pontier, 2013). These strategies are detailed in Table 1 and consider purposes of switching to English (instructional vs. managerial) (see Figure 2).

3. Considering context-specific factors. In multilingual societies with changing demographics and differences between generational cohorts, the term 'Mother Tongue' language (MTL) is not necessarily synonymous with the first and most proficient language (Wright, Boun

& Garcia, 2015). In Singapore schools, the MTL may not always correspond to children's home or dominant language: for example, home languages could include Telugu, Urdu, Hindi, Hokkien, or Teochew (Dixon, 2009); and younger generations tend to come from more English dominant homes (Bolton & Ng, 2014; Singapore Department of Statistics, 2018). As such, 'MTLs' may appear more like a heritage language for many Singaporean school children, where the MTL curriculum serves the education policy role of 'cultural ballasts', ensuring transmission of cultural knowledge (Ministry of Education, 2015), while the teacher serves the role of a good model of MTL use. A policy for English use in the MTL classroom is not widely acknowledged, and the common approach is to minimize switching to English and to maximize MTL input (e.g., Mukhlis & Pang, 2015). We therefore explored the prevalence of English use in preschool MTL classrooms with the MACS.

Second, there are aspects of the MTLs themselves to be considered, in terms of the impact on instruction. Two of Singapore's MTLs, Malay and Tamil, are characterized by diglossia, with distinct spoken and written forms. An educator's guide for Malay language teachers (Ministry of Education, 2015) expresses the need for using the standardized 'baku' form of Malay during teaching, especially for phonological instruction and reading aloud. At the same time, studies indicated that for spoken language a Johor-Riau variety of Malay is still more prevalent than the standard form in Singapore (Subhan, 2013). For the Tamil language, a spoken variety in Singapore referred to as Standard Spoken Tamil (SST) sees prevalent use in Tamil-speaking households and the wider speech community, while a literary form of Tamil is used in classrooms (Lakshmi & Saravanan, 2009; Schiffman, 2004). Teachers are recommended to use SST in classrooms to bridge the gap with the home language and to facilitate children's understanding and expression (Lakshmi & Saravanan, 2009). Thus, communicative language in and out of classrooms differs from the literary forms taught in MTL classrooms for Malay and Tamil. These sociolinguistic aspects influenced the conceptualization of the Teacher Input indicator within MACS, as illustrated in Figure 3.

Varieties of spoken Chinese, by teachers who come from outside of Singapore (Teng, 2018), may also lead to gaps with children's community language. For example, Taiwanese Mandarin has the tendency towards the neutralization of contrasts between past and present (Cheng, 1985), while for mainland Chinese Mandarin, a past specific action is almost always marked by *le* (completive or past). Another difference between these varieties of Mandarin is with reduplication of the verb, which is generally used to express the tentative aspect in mainland Chinese Mandarin, but not in Taiwanese Mandarin, which uses the term *yixia* (a while) instead. This is seen in the phrase 'wait a minute', which is expressed as *deng-deng* (wait-wait), while in Taiwanese Mandarin it is expressed as *deng yi-xia* (wait a while). As such, the Teacher Input indicator for correctness was considered in relation to Chinese Mandarin (outlined in Table 1).

4. Adapting from existing tools. Existing observation schemes that informed the development of the MACS include the IQOS (Weitz et al., 2010), the LISn (Atkins-Burnett, Sprachman, & Caspe, 2010), and the CLASS (Pianta, La Paro, & Hamre, 2008) (see Appendix A). The IQOS (Weitz et al., 2010) is geared towards classroom quality in bilingual preschool contexts and shares the aim in the current paper to compare teachers' language use as it relates to children's development, using quantitative observation methods. The LISn (Atkins-Burnett, Sprachman, & Caspe, 2010) similarly focuses on the bilingual preschool context, but in diverse settings where teachers cater to children from households with varying levels of target language exposure. The CLASS (Pianta et al., 2008) was developed to assess different aspects of children's preschool experience and teacher-child interactions holistically rather than focusing on language strategies, and was adopted as a widely used scoring scheme within the larger SKIP study (Ng et al., 2014).

With regard to the IQOS, we adapted many of the behavior indictors, including teachers' comprehensible, simplified and contextualized input that is rated under ritualized language, adapted speech, contextualization strategies under the domain of varied strategies, rich and varied input indicators, and output indicators of positive climate, whereby children were encouraged but not forced to produce and maintain target language output. After piloting and discussions to reduce rating subjectivity, we use a reduced number of indicators and changed some items from 'high' to 'low inference', through observing frequency rather than quality of the indicators for ritualized language, adapted speech, and contextualization strategies.

Also, due to differences in sociolinguistic context, we did not adopt the IQOS language separation stance where minimal translation or language switching reflected higher quality input. With the MACS, both the potentially negative and positive instances of teacher language switching to English were considered. Teachers were scored lower if they switched to English without instructional or administrative purposes on the MACS Teacher Input indicator, particularly in the case of colloquial English mixed language use. Within the MACS Use of English domain, teachers' translations between English and Mother Tongue were observed for providing learning opportunities or for classroom management purposes. Separate ratings of these frequencies allow us to consider the benefits, or lack thereof, for each type.

With regard to the LISn (Atkins-Burnett, Sprachman, & Caspe, 2010), which captures the varying ways that language switching occurs (repetition, confirmation, elaboration, and so on), we were also interested in the ways teachers switched to English from Mother Tongue, but not to the same level of detail. The LISn was used and scored differently, involving five-minute ‘snapshots’ per observed child, whereas MACS scoring was focused at the class level for the 10- to 20-minute video observations. While more intensive, child-focused rating has its merits, the approach was not intended to capture other elements of classroom language exchanges, such as the quality or complexity of each type of talk, which was the focus for the MACS. The LISn goal was to closely observe the type of individualized input and interactions a child would receive or be engaged in, and thus provided a useful model for types of talk data and activity types that are important for observation.

With the CLASS (Pianta et al., 2008), one of its three domain scores was most relevant to the objectives of MACS: Instructional Support, which includes the indicators of Language Modelling and Quality of Feedback. These were both incorporated within the Language Modelling and Scaffolding indicator of the MACS. However, the MACS descriptors were revised to include effectiveness of teacher strategies through the observation of the proportion of children who spoke in class, and the extent to which they spoke (words, phrases or sentences). The Varied Strategies domain of the MACS also adapted activities and directions from the CLASS Instructional Learning Format dimension (La Paro, Pianta, & Stuhlman, 2004), but these were further detailed in the MACS (see Table 1), while CLASS adopted a more general evaluation. Contrary to the purposes of the other schemes above, CLASS was not intended for rating bilingual practices in classrooms and did not include language switching.

The scoring scales of the CLASS were adapted into the MACS, in terms of the ‘depth, duration and frequency’ of teacher input and strategies for each indicator coded. This was due to the similar way the video data was captured and prepared for scoring, whereby 10- to 20-minute video segments were scored, and the focus of the scoring was on teacher-child interactions.

However, for MACS' focus on specific aspects of mother tongue language learning, necessary modifications were made to the rating scales after piloting (as described below).

Current Study

The contextual factors discussed above reflect diversity in preschoolers' MTL proficiencies and learning needs that teachers have to address in order to provide quality language instruction. Therefore, the MACS was developed somewhat differently from other tools, in order to assess a range of teacher input and strategies. This was planned to better understand aspects of quality MTL instruction and input, given the unique local context with variation in children's MTL proficiency. This study explores the degree to which the MACS achieves this aim with the following research questions (RQ):

RQ 1. What contributes to the conceptualization and development of a classroom quality scoring scheme in preschool language classrooms for bilingual learners?

RQ 2. To what extent can different raters reliably code classrooms with the MACS?

RQ 3. Do the dimensions of the MACS fit with theoretical expectations?

RQ 4. Do ratings across the dimensions predict child learning outcomes?

Different methods were applied to address each research question. The first research question involves steps in the construction of the MACS from initial development, to revisions, and finally the application of the scheme, as seen in Figure 1. The literature review, summarized above, was followed up with revising the scheme through piloting, consensus discussions and consultation with an expert panel, resulting in a manual with descriptors for each indicator and examples derived from the first set of classroom videos (Table 1). These phases are elaborated in the results section. To address the second research question, a new group of raters was exposed to a set of training videos, and their interrater reliability was established across the full sample of the classrooms. To address the third research question, we use points outlined in Lahti, Elicker, Zellman, and Fine (2015) and Zellman and Fiene (2012) to establish reliability, and construct and concurrent validity. Finally, to examine research question four regarding the predictive validity of the MACS, child language measures were examined across the rated classrooms.

The initial version of MACS consisted of the four domains noted above: *input strategies, output strategies, varied strategies* and the *use of English*. These consisted of 12 behavioral indicators initially, and 11 in the final version (Figure 2).

Methods

The samples, measures and procedures for the study are provided below for each research question, based on the different phases of the study. Samples include the teachers in the observed classrooms, the coders who rated the classroom teachers with the MACS, and students from the classrooms. Overall there were 51 classroom teachers, with 27 observed for the initial phase (RQ 1) and 48 double scored for estimation of interrater reliability (RQ 2), and 491 students (RQ 4).

Sampling

Across the 51 classrooms selected from a larger longitudinal study (SKIP, Ng et al., 2014), all teachers are female and acquired at least a diploma to teach in Singapore preschools (Early Childhood Development Agency, 2018). Teachers' self-reported proficiency indicated that they are fluent in MTL and conversational in English as rated on a 1-7 scale (English - *M* = 4.11, *SD* = 1.85; MTL - *M* = 6.16, *SD* = 1.32), speaking (English - *M* = 4, *SD* = 1.78; MTL – *M* = 6.11, *SD* = 1.35), reading (English - *M* = 4.03, *SD* = 1.91; MTL – *M* = 6.14, *SD* = 1.46) and writing (EL - *M* = 3.57, *SD* = 2.08; MTL – *M* = 5.92, *SD* = 1.46); where 3 = "can communicate routine, basic information", 4 = "can independently understand/use language to carry on conversation", and 5 = "can independently interact with fluency and spontaneity").

From the larger study, all Malay and Tamil classes and a random subset of Chinese classes were included for coding and analysis. Differences in the number of classes per language are due to population demographics (74.3% Chinese, 13.4% Malay, and 9.1% Indian; Singapore Department of Statistics, 2018). Classrooms included in the sample of 51 videos were from government-run (21.6%), public (17.6%), not-for-profit (45.1%), and commercial (15.7%) preschool education providers recruited within the larger study.

During each observation session, one microphone and receiver was attached to a teacher-focused camera to capture teacher-child interactions. Typical MTL lessons last 30 to 60 minutes, depending on the school program, and at least one video session with the best angles and sound quality was chosen per class for coding. If there was more than one video for the MTL lesson, all videos (up to three) were coded and item scores were averaged. The maximum duration for each video recording is 20 minutes, and the average duration across the 51 classrooms was 18min 20s (*SD* = 2min 31s).

The average number of children per class was 14.8 (*SD* = 3.99) for Malay, 15.18 (*SD* = 3.45) for Tamil, and 15.08 (*SD* = 3.32) for Chinese classes. Out of the four hundred and ninety-one children with outcome data, 96.5% matched in their ethnic background and classroom MTL assignment. Assignment to MTL classes is based on ethnicity, except in cases where the child's MTL is not taught in their school due to low enrollment, they will join an available MTL class (Wong, 2018), with the teacher sometimes codeswitching to the shared language of English (Yussof & Sun, 2020).

All coders had at least a bachelor's degree in early childhood, education or psychology. All were Singaporean, and were fluent bilinguals in English and their MTL, having completed their formal education locally, and having acquired and practiced spoken and academic registers of their MTL. Thus, all Malay coders used the Johor-Riau variant of spoken Malay (Subhan, 2013), and learned formal Malay in school. The Tamil coders used Standard Spoken Tamil (Lakshmi & Saravanan, 2009), and literary Tamil for reading and writing. These language experiences allowed them to distinguish and code teachers' use of different registers within the classroom.

RQ 1. Phase 1-2 scheme development and revision.

Teachers. Twenty-seven Kindergarten 1 (K1) MTL teachers (6 Malay, 7 Tamil, 14 Chinese) were observed. Their mean age = 41.37 (*SD* = 11.80), and average preschool teaching experience = 11.05 years (*SD* = 6.93).

Coders. Six coders (2 per MTL) rated classroom videos from Phase 1, while fifteen total coders rated videos in Phase 2 (refer to Figure 1).

RQ 2. Phase 3 scheme application to classrooms.

Teachers. Forty-eight observed classrooms were double coded, including15 K1 teachers and 33 K2 teachers (12 Malay, 4 Tamil, 32 Chinese). Their ages ranged from 21 – 69 years (K1 - *M* = 41.00, *SD* = 11.22; K2 – *M* = 41.88, *SD* = 12.44), and preschool teaching experience ranged from 1 – 30 years (K1 - *M* = 12.45., *SD* = 7.43; K2 – *M* = 10.85, *SD* = 7.09).

Coders. Eighteen coders (6 Malay, 3 Tamil, 9 Chinese) rated the 48 classroom videos using the final version of the MACS. Nine coders (2 Malay, 2 Tamil, 5 Chinese) were also trained in CLASS.

RQ 3. Phase 3 scheme fit with theoretical expectations.

Teachers. Ratings for all 51 MTL teachers (13 Malay; 6 Tamil; 32 Chinese) were conducted by the eighteen coders described above.

RQ 4. Phase 3 scheme prediction of child outcomes.

Teachers. Ratings were used for the same 51 teachers from the 18 coders.

Children.

Classrooms Sample. There were 491 children within the 51 classrooms with teachers' MACS ratings, including 325 Chinese, 75 Malay, and 91 Tamil learners, with an average of 10.13 children within each classroom cluster (*SD* = 6.72, Range of 1 – 41). They were assessed on the MTL vocabulary, reading, and morphological awareness measures at three time points: Kindergarten 1 (K1), Kindergarten 2 (K2) and Primary 1 (P1).

Total Comparison Sample. To appraise the learning of these children, rather than using raw test scores, we considered their learning trajectories relative to same-language peers from the larger sample of SKIP. This was to account for differences in the assessments across MT languages, and because of a lack of locally-normed assessments. Data from the children in the MACS classrooms along with the larger SKIP sample (*N* = 1538) were included in growth mixture models (GMM) to determine latent classes of learners (high versus low progress). Separate GMMs were run per language group (CL-Chinese, ML-Malay, TL-Tamil) for each outcome. The total samples for these models included: (1) 1075 CL, 195 ML, 238 TL for vocabulary; (2) 323 CL, 137 ML, 161 TL for reading; and (3) 319 CL, 137 ML, 168 TL for morphological awareness. Total sample size on each measure varies due to missing data (e.g., child absences) and to task administration (subsamples of SKIP). All children participating in SKIP were typically developing learners.

The classroom sample did not differ from the total sample in terms of age, gender proportion, non-verbal intelligence, or family income: age in years at K1 *M* = 5.28 (*SD* = 0.32) vs. 5.26 (*SD* = 0.32); 46.7% vs. 49.7% female; Raven's Coloured Progressive Matrices (Raven, 1947) raw scores *M* = 15.91 (*SD* = 5.10) vs. *M* = 15.76 (*SD* = 5.23); and average monthly household income rating *M* = 13.47 (*SD* = 5.85) vs. *M* = 12.01 (*SD* = 5.97) on a scale of 19, corresponding to an average range of S$7000 – S$7499 vs. S$6500 – S$6999.

Measures

MACS Ratings. There are 12 behavioral indicators initially, and 11 indicators included in the final rating scales of MACS (see Figure 2). The final indicators were: 3 Input (teacher input, language forms, concept development); 2 Output (positive climate, language modelling/scaffolding); 4 Varied Strategies (ritualized language, adapted speech, use of gestures, use of pictures/objects); and 2 Use of English (learning opportunities, behavior and task management) indicators. The Input and Output domains were scored on a 1 to 7 scale, similar to CLASS (Pianta et al., 2008). This was done to observe the level of complexity in teacher input and strategies (low – medium – high). Indicators under these domains are high inference items, and coders were instructed to gauge consistency and depth of teacher input and strategies across the whole video segment. Given this element of judgement, the MACS adopted the within-one reliability scoring approach (allowing coders to be one point off from each other or from the consensus score), a procedure used in other coding systems (e.g., CLASS). The Varied Strategies and English Use domains were scored on a scale of 1 to 3, as these were low inference items. These domains constitute behavior frequency ratings that quantify the occurrence of the described behaviors indicated at different levels (absent – minimal – regular). Since items in these domains have to be definitively observed, coder scores had to be exact to achieve inter-rater reliability. Scores per indicator were entered into the analyses. The mean ratings across all classrooms for the dimensions of MACS are indicated in Table 2.

For the final version of the MACS (used in Phase 3), a training manual included the following operationalized definitions per indicator, with examples from sample classes to illustrate the target behaviors. Table 1 provides a summarized version of all MACS descriptors.

Input Strategies. This domain includes 3 aspects of teacher linguistic behaviors and language instruction strategies constituting the MTL input children are exposed to.

Teacher input. Given the importance of teacher proficiency for providing quality instruction and facilitating children's language acquisition, this indicator firstly observed 'correctness' of teachers' input in terms of pronunciation, vocabulary use, sentence formation and grammar (Canh & Renandya, 2017; Richards, 2010). Additionally, this indicator considered if input included 'consistently varied vocabulary and sentence structures', since richness of input is important (Bornstein et al., 1998). Further, teacher input was coded for 'appropriateness' of formal and informal register use in relation to the context. For instance, the formal register was expected during reading and phonological instruction, while the informal register was expected in social interactions to build rapport with children (Lakshmi & Saravanan, 2009; Ministry of Education, 2015; Schiffman, 2004; Subhan, 2013; see Figure 3 for descriptors of 'appropriate' language register use). The Teacher Input indicator was thus scored using the 1 to 7 scale to differentiate between teachers who provided consistently incorrect, inappropriate, colloquial, and repetitive MTL input at the low end of the scale, and those who consistently provided correct, appropriate and varied MTL input at the high end.

Language forms. Instructional methods for developing children's phonological, morphological, syntactic, semantic, pragmatic, and vocabulary knowledge (Ishwaran et al., 2005; Sze & Leung, 2010) were rated on a scale geared towards capturing the different levels of teacher facilitation strategies. In the Language Forms section in Table 1, teachers were scored low when they either did not engage children in learning or discussing language forms at all, or any requests for children to read or spell that were not scaffolded. Teacher scaffolding strategies included linking letters to letter sounds, and breaking words into syllables to facilitate decoding. Teachers were scored higher when they were consistently observed to provide scaffolding for decoding and making meaning.

Concept development. While the local curriculum framework on MTL teaching and learning places heavy emphasis on the role of MTL in providing cultural education (Ministry of Education, 2013), we considered application of the MTLs across a wider range of topics that could facilitate and encourage broader language use and understanding. This indicator thus focused on teachers' MTL use for developing concepts and modelling thinking processes in other relevant topics such as science, math, and art (e.g. Pianta et al., 2008). With this indicator, teachers were scored low if classroom talk is teacher-dominated, with little to no opportunity for children to express their own ideas and opinions. Use of only close-ended questions and decontextualized discussion of topics were considered as limiting to children's opportunities to speak, think, and understand topics discussed in the MTL. Teachers who employed these strategies were thus scored lower. Conversely, teachers who endeavored to elicit and encourage children's speaking, thinking and understanding of different topics discussed in MTL through, for instance, brainstorming together, asking open-ended questions and linking topics to children's daily life, were scored higher.

Output Strategies. This domain includes 2 indicators related to how teachers elicit children's MTL use, and how effective the strategies were, as evidenced by children's responses.

Positive Climate. For this indicator, we considered if teachers were patient, encouraging, and if they scaffolded children's MTL output (Pianta et al., 2008). Coders observed if teachers were effective in making children comfortable and confident enough to use MTL, and to what degree. Coders also considered if teachers were permissive rather than punitive with children's English use, since children's language mixing is indicative of emerging language skills (Byers-Heinlein & Lew-Williams, 2013; Weitz et al, 2010). Teachers who limited children's expression, especially through ignoring or scolding children when they switch to English and provided no help as children struggled to use MTL, were scored low. When teachers were patient, incorporated and affirmed children's talk regardless of the language used, and provided help and wait time to support children's MTL use, they were scored higher.

Language Modelling and Scaffolding. As with positive climate, the effectiveness of teacher strategies for modelling and scaffolding language acquisition and production were assessed through child output. Coders considered if these strategies could effectively elicit more child output: open and closed-ended questions, repetitions and extensions of children's talk, mapping language to actions, introduction and contextualization of new words and concepts (Pianta et al., 2008). Teachers were scored for consistency and variety of different language modelling and scaffolding strategies used. When more children were observed to use more MTL in response to these strategies, the teachers were scored higher.

Varied Strategies. Varied strategies included 4 indicators frequently discussed in existing literature, and included in other observation tools: *ritualized language* (phrases, songs or rhymes used for routine activities), *adapted speech* (child-directed talk in terms of tone, pitch or repetition), use of *gestures and expressions*, and use of *pictures, objects, and realia* (Kersten et al., 2010; Weitz et al., 2010). As in Table 1, the scale used indicated whether the varied strategies were absent, minimal (1 to 2 observed instances) or regular (more than 2 observed instances).

***Use of English*.** Two forms of English use for supporting MTL teaching were rated for this domain. First is the presence and frequency of teachers' English use for *learning opportunities*. These 'learning opportunities' were considered not only when teachers provided explanations or translations of target vocabulary and concepts in English, but also when teachers translated instructions between MTL and English to facilitate task completion and behavior management. In these situations, the English use was still considered an opportunity for learners to be exposed to MTL input. Second, English used for *behavior and task management* was coded when only English was used for the sole purpose of disciplining children and facilitating task completion (Gort and Pontier, 2013). Here it was considered that children did not have the 'opportunity' to make connections between English and MTL as part of learning MTL. Scores indicated the presence and amount of English use.

CLASS Ratings. Videos were previously coded using the Classroom Assessment Scoring System (CLASS) for the SKIP study (Ng et al., 2014). All coders who used CLASS were Teachstone certified at the time of coding. Scores for these videos, on the three CLASS domains of emotional support, classroom organization and instructional support, were derived according to the training manual (Pianta et al., 2008). The mean CLASS ratings across all classrooms were 3.46 (*SD* = 0.726) for the ES, 4.51 (*SD* = 0.640) for the CO, and 2.25 (*SD* = 0.824) for the IS domains respectively.

Child outcome measures. Assessments of children's MTL vocabulary, reading, and morphological awareness were conducted across the noted three time points: K1, K2, P1. Details of the assessment measures are included in Appendix B. These data were used to identify learning trajectories over time. Five children (0.01%) completed tasks in the MTL assigned in school rather than that used at home.

Receptive vocabulary. The Bilingual Language Assessment Battery (BLAB; Rickard-Liow & Sze, 2008) is a locally developed test, which has been widely used for vocabulary assessment in Singapore. For each language, the test was administered on a tablet, where the participant chooses one of four pictures corresponding to a word presented aurally in MTL through the tablet. Children completed three practice trials with corrective feedback, then 80 trials for a score of total correct responses. Spearman-Brown split-half reliabilities at each time point were 0.85, 0.78 and 0.80, respectively.

Reading. Tasks were developed per MTL script that included grapheme recognition (discrimination), naming and word naming components (see Appendix B for details). For Chinese, the total score was calculated as the number of total correct responses across the three tasks (character discrimination, stroke naming, character naming). Spearman-Brown split half-reliabilities at each time point were 0.95, 0.94, and 0.96, respectively. For Malay, the total score was calculated as the number of total correct responses across two tasks (letter naming and word reading). Spearman-Brown split-half reliabilities at each time point were 0.98, 0.98, and 0.92, respectively. For Tamil, the total score was calculated as the number of total correct responses across the three tasks (letter discrimination, letter naming, and word reading). Spearman-Brown split-half reliabilities at each time point were 0.98, 0.98, and 0.98, respectively.

Morphological awareness. Tasks to evaluate children’s understanding on the morphemic structure of words were developed per MTL, and included two measures adapted from previous research: Compound word production, and compound structure judgement (Tong, McBride-Chang, Shu, & Wong, 2009; Chen, Hao, Geva, Zhu, & Shu, 2008) (see Appendix B for details). A total morphological awareness score was calculated as the sum of compound production and compound structure task scores. Spearman-Brown split half-reliabilities at each time point were 0.75, 0.60, 0.73, respectively.

Procedures

Classroom observations and coding.

MTL lessons were video-recorded for offline coding. Figure 1 describes the phases of development for the MACS. Two rounds of coding were conducted in Phase 1. The second author introduced the initial scheme to 2 coders per MTL who then each coded one pilot video with the MACS, and noted potential issues for coding. This feedback contributed to minor elaborations and edits, then these original six coders plus one Chinese language coder reviewed and applied the scheme to two videos each. They attained an average inter-rater reliability agreement of 80% across these 2 videos. They next coded a total of 6 Malay, 7 Tamil and 14 Chinese language sessions.

In Phase 2, MACS revisions incorporated the feedback from the two coding rounds in Phase 1, plus consultation with local language curriculum and linguistic experts, as well as transcriptions and preliminary analyses of 27 language classrooms. A manual with detailed descriptors and exemplars from the classroom videos was produced as part of finalizing the MACS. This manual was used to train 15 coders (5 per MTL) for coding in Phase 2. Training and reliability testing included two days of reviewing the scoring manual, discussion of all indicators and examples, and some coding practice. Coders then scored 5 classroom videos and were required to obtain on average 80% or better inter-rater agreement with the consensus scores. Disagreements in ratings were resolved with consensus discussions, and these discussions also informed the revised coding scheme.

In Phase 3, additional coders were trained to the reliability rate of 80% or better across 5 training videos. A total of 6 Malay, 3 Tamil, and 9 Chinese coders viewed 51 classroom videos. Of these, 48 videos were double coded by randomly assigning pairs of coders to score them. The dual ratings on these 48 classrooms were entered into an inter-rater reliability analysis. Across the videos, final ratings were reached by consensus discussion if inter-rater agreement was below 80% (e.g., coders compared notes and provided justifications for their scores). For double-scored videos with inter-rating reliability above 80%, scores from the first coder were used. Where needed, transcript and video data were double-checked to ascertain the score to be given. These consensus scores were then included in the data used for the exploratory factor analyses. Scores from all 51 coded classrooms were used for the analyses of child outcomes.

Analysis Plan

We examined the 4 research questions using as a rough guide the components of validation approaches that have been utilized in other quality measurement studies (e.g., Lahti et al., 2015; Zellman & Fiene, 2012; see also Halle, Whittaker & Anderson, 2010) (keeping in mind that our approach involves a quality rating *tool* rather than a *system*). RQ 1 involved the key concepts of the face or content validity of the MACS, and these are reviewed in the results. RQ 2 involved reliability of scoring the MACS. RQ 3 involved a latent structural analysis and exploration of convergent and divergent validity with another coding measure. RQ 4 involved predictive validity with learning outcomes.

RQ 2. Inter-rater reliability was calculated using the Kappa statistic, or the weighted Kappa for the low inference items such as ritualized language, adapted speech, use of gestures, use of pictures/objects and English use for learning opportunities and behavior management. The Kappa statistic is conservative in that it corrects for "chance" agreement, with more fine-grained scales (such as our 7-point scales) scores that are close but fall within one point of each other were considered to be in disagreement. Therefore, Kappa values were recalculated using scores of the first 5 items (teacher input, language forms, concept development, positive climate, language modelling/scaffolding) counted as in agreement if they were within one point of each other. The other 3-point scale items were not included in the within-one re-calculation.

RQ 3. Inter-item correlations were run for all items on the MACS, using the consensus scores from multiple raters of the 51 classrooms. To further examine the structure of the coding scheme, and with the objective of reducing the number of items, we took a factor analytic approach. While confirmatory factor analysis would be an ideal method to check whether items align according to the four domains of the MACS, we opted to use exploratory factor analysis to observe data-driven relations in the case that they would not fit according to our theoretical domains, and due to our limited sample size (for samples smaller than 50 cases, EFA could yield reliable results with 3-4 factors and moderate (0.7) factor loadings for about 12 variables; de Winter, Dodou & Wieringa, 2009). Following these expectations, we proceeded to conduct the EFA using all 11 items with a weighted least squares (WLSMV) estimator and oblique rotation (Mplus version 8, Muthén & Muthén, 2017). Data were entered as ordered categorical variables, having been rated on a scale of “low” to “high” dimensions of correctness/sophistication (described above where the high dimension was considered to be more aligned with the goals of the kindergarten curriculum framework from the Ministry of Education). To determine the optimal number of factors, a scree plot was first examined, along with indices for model fit (χ^2 p-value > 0.05; RMSEA < 0.05, CFI > 95; SRMR < 0.08) (Clark & Bowles, 2018).

RQ 4. We first conducted growth mixture modelling (GMM) to identify latent classes of learners based on their growth in several MTL skills: receptive vocabulary, reading, and morphological awareness. Separate GMMs were conducted for each language group using available data from the total SKIP (Ng et al., 2014) sample of children using Mplus v.8 Software with an MLR estimator (Muthén & Muthén, 2017). This approach was taken because the measures between MT languages differed in terms of the number of items and difficulty of items, such that the forms were not parallel. The outcomes of these GMMs (provided in Appendix C) supported a 2-class solution for each model, with one class having steeper slopes (and higher intercepts in most cases, for vocabulary and reading), and another class with shallower slopes (and lower intercepts in most cases). We refer to these classes, respectively, as the high progress and low progress learners, and the latent class assignment was attributed to each individual child for the outcome measures. These categorical data were then used to determine the proportion of high progress learners within each classroom, and this was correlated with the MACS classroom scores.

Results

RQ 1. Content for classroom quality

Between Phases 1 and 2 (Figure 1), the following sets of revisions were made between the initial and final versions of the MACS.

Teacher input indicator and informal language. The teacher input indicator initially required coders to consider only correctness and variety in teachers' classroom talk, with colloquial forms included at the low end of the scale. The informal language indicator was initially a separate indicator under the varied strategies domain. Observations from the first round of coding found prevalent use of non-standard and informal language forms such as the Johor-Riau variant of Malay and Standard Spoken Tamil, which made the teacher input indicator difficult to score. Following consultation with language and curriculum experts from the

Ministry of Education and National Institute of Education, the objectives of MTL instruction (Ministry of Education, 2013) were highlighted: to develop communicative skills, it is essential for teachers to also expose children to spoken language forms. Thus, the scheme was modified to include informal language use within appropriate classroom settings, such as in social talk for teacher-child rapport building (see Figure 3).

MTL specific coding descriptors. The second major revision was to produce descriptors specific to each MTL. Using the initial version, coders found the descriptors inadequate to inform language specific factors. Hence, the teacher input indicator was adapted to incorporate different registers and variants of the standard language for Tamil and Malay (Lakshmi & Saravanan, 2009; Subhan, 2013), while Chinese grammar and vocabulary used by teachers hired from outside Singapore was taken into account. Also, the language forms indicator added descriptors that differentiated between modelling at the phoneme, alpha-syllable and syllable level for Tamil and Malay language, versus discussion of strokes and stroke patterns for writing characters and teaching meaning-based forms that aids character recognition for Chinese script.

Reducing subjectivity. Steps were taken to make descriptors more explicit, including the MTL-specific descriptors. The first set of videos from K1 were transcribed, and a selection of relevant examples were drawn out to illustrate the different descriptors and provide more detail for coders. These transcriptions also served as a written account, reference check that facilitated consensus-building for final ratings and for reliability training videos for new coders. To further reduce subjectivity, quantitative guidelines were included where possible. For instance, under the positive climate indicator, a medium score meant at least ¼ of children were comfortable responding, while at least ⅓ were required for a high score. Similarly, the proportion of children responding was used in the language modelling and scaffolding ratings. Behavior frequency ratings were applied to the domains of varied strategies and English use, as ‘absent’ (0), ‘minimal’ (1-2), or ‘regular’ (> 2 occurrences).

RQ 2. Rater Reliability

Interrater reliability on the 48 double-coded classrooms showed overall acceptable Kappa values using an exact estimate for Malay and Tamil ratings, but this was low for the Chinese language group: Chinese = 0.34, Malay = 0.60, Tamil = 0.60. Using within-one estimates yielded generally higher Kappa values, which were within acceptable ranges for all languages: Chinese = 0.75, Malay = 0.82, and Tamil = 0.72. Estimates per indicator are provided in Table 1.

RQ 3. Construct Validity: Dimensional Structure of the MACS

Cross correlation between MACS indicators. Relations between items within each of the 4 MACS domains are shown in the shaded portions in Table 3. These inter-domain relations were expected to be stronger than cross-domain relations, outside of the shaded areas in the table. Moderate significant inter-domain correlations were found for indicators of Input (1-3), showing that teacher’s input quality covaried with both language forms and concept development emphases. However, these indicators of input also showed strong correlations with those of other domains, especially modelling/scaffolding. The two indicators for Output (4, 5) were significantly correlated with each other, indicating that teachers who use one strategy tended to use the other as well. As noted, these indicators also correlated with the input indicators. The varied strategies domain indicators (6-9) were less coherent, especially for pictures/objects, which was only correlated with concept development from Input. Significant moderate inter-domain correlations here were found only for gesture with ritualized language, and with adapted speech strategies. Adapted speech was also related to cross-domain indicators of input and output. The final domain of English use showed no intra-domain relations between the two indicators (10, 11), suggesting teachers selectively used English for different purposes.

Factor Analysis of MACS ratings. In the initial model with all items included, the *Strategic use of gestures/expression* was highly correlated with and not statistically distinguishable from two other variables, and the model was then re-run without this variable. The EFA supported a three-factor structure for the MACS data, according to a scree plot and fit indices: χ^2 (18) = 14.30, p = 0.71, RMSEA = 0 (CI = 0.00-0.096), SRMR = 0.059, CFI = 1.00. This model yielded a better fit than a 2-factor model, χ^2 (26) = 23.51, p = 0.60, RMSEA = 0 (CI = 0.00-0.098), SRMR = 0.081, CFI=1.00, or a 1-factor model χ^2 (35) = 40.29, p = 0.25, RMSEA = 0.054 (CI = 0.00-0.119), SRMR = 0.119, CFI= 0.988. Factor loadings from the final 3-factor model are shown in Table 4. ‘Input’ and ‘output’ categories converged onto one factor along with ‘varied strategies’ of ritualized language and adapted speech (F1), while the ‘English use’ variables split into separate factors (F2, F3) in agreement with the intra-domain correlation noted above (see Table 2). Instructional English use was coupled with teachers’ lesser focus on language forms, while task management English use was related to more of this focus plus teachers’ use of pictures/objects/realia (see Table 4).

Relation between rating schemes. Correlational analysis was run between the MACS Factor Scores from the EFA and the CLASS domain scores. As shown in Table 5, there were no significant relations between the three MACS factors and the three CLASS domains. This was in spite of the similar indicators that were adapted from the CLASS; the instructional support dimensions of language modelling and concept development, and the emotional support dimension of positive climate.

RQ 4. Predictive Validity: Link to Child Learning Outcomes

Correlations were run between teachers' MACS factor scores and their proportion of high progress learners within the classroom. In Table 5, it can be seen that MACS Factor 1 scores were significantly and positively correlated with a greater proportion of high progress learners in terms of MTL vocabulary, but significantly and negatively correlated with the proportion of high progress learners for MTL reading. Scatterplots of these significant MACS factor score relations to high progress learners are shown in Figure 4. Note that 1 case in vocabulary and 2 cases in reading had significant Cook's distance and these classes were omitted. None of the other relations were significant (all *p*'s > 0.1). For comparison, we also examined the correlations of teachers' CLASS ratings with their proportion of high progress learners, and none of these relations were significant (all *p*'s > 0.1), as shown in Table 5.

Discussion

Given the pivotal role that early childhood education plays in multilingual learning, there is a clear need to identify key aspects of quality language instruction. Available ECE quality rating systems may not be sensitive to the specific issues related to bilingual learning across different socio-cultural contexts such as Singapore. The current study describes the development process for a new ECE rating tool to address this gap and to highlight characteristics of quality ECE for MTL learning in an increasingly English-dominant multilingual society. The overriding purpose for this tool is to inform research and practice in early bilingual development and education. The MACS is not a teacher evaluation for high-stakes purposes. Rather, the MACS is a first step toward understanding the features of ECE quality for multilingual outcomes, with longer-term implications for bilingual education policy.

As a first step to identifying characteristics of MTL ECE quality, the MACS observational tool is intended to be comprehensive. Multiple sources contributed to the content of the coding scheme, which resulted in a four-domain conceptualization of quality that we felt encompassed important aspects for MTL learning in the preschool setting: teachers' (1) input strategies, (2) output strategies for eliciting child talk (3) other relevant varied strategies for early language learning in bilingual contexts, and (4) use of English for classroom learning and management purposes. Given the well-founded emphasis on comprehensible input for language learning, in the current context we needed to account for teachers' use of different registers for different purposes – as more instructional or more conversational talk (see Figure 3). Our input domain also included indicators related to the purpose of the language lesson – as focused on language forms, or concept development for topics such as science or arts (Pianta et al., 2008). Emphasis in the literature on output for language learning also influenced the inclusion of positive climate and language modelling and scaffolding strategies for increasing child output. The recognition that children need to feel secure and supported in their language production is established in preschool studies (Schwartz & Gorbatt, 2010; Weitz et al., 2010), and scaffolding their efforts in language output is important to bilingual learning (Bohman, Bedore, Pena, Mendez-Perez & Gillam, 2010). For each of these indicators, we arranged the examples of high ratings to reflect the literature in terms of beneficial strategies for teaching and learning, while also aligning them with the educational goals of the Ministry of Education's framework.

Rounding out the content of the MACS from these two key components of input and output, we added domains with pedagogical indicators specific to the preschool and local context. Strategies to increase comprehensible input for young children, who vary in MTL proficiency were observed in teachers' simplified and child adapted speech, nonverbal gestures, and English (Weitz et al., 2010; Atkins-Burnett et al., 2010). The literature is equivocal on which strategies are more beneficial, such as the use of language switching, so we rated the frequency with which such observed behaviors were used for different purposes.

To test and refine this conceptualization of MTL classroom quality, we then applied the MACS to a set of video-recorded MTL kindergarten classroom sessions. After addressing initial challenges with coding reliability, by detailing a coding manual with language-specific illustrative examples and by modifying rating scales to include frequency ratings, overall interrater agreement was within acceptable limits for within-one estimates (although there is recent concern surrounding the "within one" approach, Mashburn, 2017). Reliability per indicator varied, however, with better reliability on the frequency-based ratings (absent to regular) (although some were still low, Kappa < 0.6), and improved reliability on the finer-grained low-high scale items for within-one estimates.

Using consensus scores, we then examined the domain structure for the MACS with correlation and factor analyses. The correlations between items did not follow the expected pattern, whereby indicators within a domain would be more highly correlated than between domains. Indicators across input and output domains were moderately related, while indicators within the English use domain were not related to each other. Results from the exploratory factor analysis coincide with this pattern of results, showing moderate to high factor loadings of the input and output indicators on the first factor (F1), along with the adapted speech strategy. The pattern of F1 factor loadings is suggestive that both comprehensible input (through teachers' rich, accurate, and adapted MT speech and language modelling), and elicited child speech (through a positive climate and scaffolding), contribute to quality language instruction. That both of these domains load onto a single factor might also indicate the importance of an input-output cycle for language learning and eliciting speech production (e.g., Gass & Mackey, 2007;

Pearson, 2007). The English use indicators loaded on separate factors (F2, F3), on the other hand. English for learning opportunities loaded onto F2 with a negative contribution of language forms strategies, which may be due to a tendency to emphasize oral language more than literary forms and thereby using English translation to teach new vocabulary, as observed in some classrooms. Thus, a re-organization of the MACS domain scores may be in order. These preliminary results require replication with a larger sample of teachers, however.

Comparison of general classroom quality and our measure of MTL classroom quality showed no overlap in these scores. The CLASS teacher ratings, which were completed on the same classrooms as part of the larger study, did not correlate with the factor scores from the MACS. This was in spite of the adapted indicators from the CLASS IS dimension (language modelling and concept development) and ES (positive climate) into the MACS, and indicates that our measure is unique from more broadly defined teacher-child interaction process quality.

Finally, we examined how the MACS ratings of quality for MTL learning in the classroom may be related to the learning outcomes of students in those classrooms. We examined several measures of MTL learning over time, including high progress (relative to peers) in receptive vocabulary, reading, and morphological awareness. Teacher ratings on F1 (input and output) were correlated with having more high progress learners of MT vocabulary, but negatively correlated with these types of learners for MT reading. This may follow from the emphasis on oral language development in the kindergarten framework for MTL (Ministry of Education, 2013) and the stronger teacher input and child elicited speech scores in some classrooms may disproportionally affect oral vocabulary as compared with reading of scripts. The majority of teachers (85%) were rated in the medium range for teacher input a (Table 1), whereas there was a broader range of ratings across the low to medium ranges for language forms (with 43% in the low range). Extending reading and writing lessons to a context of making meaning with print (high rating) rarely occurred, in only 2% of teachers. Thus, explicit linkages between oral and written language may not be made at this age – meaning that a focus on oral language vs. written language may be more exclusive. On the other hand, children's scores on language and literacy were positively related, consistent with the extant literature, so the converse effects might instead be related to the MACS scales, which centered on spoken language indicators to a greater degree. Teacher scores on the other factors, and on the CLASS domains were unrelated to the proportion of high progress learners. Though preliminary, these results suggest that the MACS ratings may capture elements of language classroom quality that are important for children's MTL outcomes.

Implications

We identified teachers' language and elicitation of child speech as key components for ECE quality related to MTL learning. Strategies of language modelling, scaffolding and adapted speech, as well as using the MTL for concept development and creating a positive classroom climate, appear to have positive effects on children's vocabulary progress. Scaffolding may be especially important in the context of diverse language learner competencies and in the face of shifting home language trends. On the other hand, a range of teachers' English use was observed in the present classrooms, and for different purposes. These results do not inform the question about language switching benefits, but may recommend training teachers to be aware of their language quality and purpose for English use.

Limitations

The findings need to be considered in light of several limitations. First, the sample of classrooms was relatively small and at the limits for conducting the analyses we wished to enlighten about the MACS' latent structure. Also, Tamil, and Malay teachers were less represented, but showed that some quality indicators needed adjustment to account for language-specific issues such as diglossia use. Second, the data sampling method followed standard protocols for observation schemes, with roughly 20-minute segments from a "typical day" in the classroom. Other approaches may yield behavior stability over observations or at finer time scales. Third, while interrater reliability improved over iterations of the coding scheme, many of the indicators yielded poor agreement unless a with-one estimate was used for the 7-point items, suggesting more master-code feedback may be required during training. Fourth, the evaluation regarding student outcomes was limited by a lack of standardized assessments, so learners were categorized as high- or low-progress relative to peers. However, in some cases the groups differed in incoming scores (e.g., reading) and we cannot rule out extra-classroom variables on their outcomes, which would require further study.

Conclusion

The MACS was developed to apply generally across bilingual preschool contexts – as compelled by the local Singapore context, wherein multiple Mother Tongue languages are incorporated into ECE. Although we aimed for a generalizable coding scheme, this was balanced by the need for language-specific aspects of the rating tool. By providing a detailed account of the conceptualization and development phases of the MACS, we hope to provide a demonstration of how rating tools can be adapted to assess and describe quality language teaching in bilingual ECE contexts. The current intention was using this tool to understand the constituents of ECE quality as related to child outcomes within multilingual contexts. Further research applying the rating tool to larger, representative samples of teachers in classrooms may contribute to bilingual education policy and teacher professional development.

References

Atkins-Burnett, S., Sprachman, S., & Caspe, M. (2010). *Language Interaction Snapshot + End of Visit Ratings (LISn + EVR)*. Princeton, NJ: Mathematica Policy Research.

Bohman, T. M., Bedore, L. M., Pena, 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.

Bolton, K., & Ng, B. C. (2014). The dynamics of multilingualism in contemporary Singapore. *World Englishes, 33*(3), 307-318.

Bornstein, M. H., Haynes, M. O., & Painter, K. M. (1998). Sources of child vocabulary competence: A multivariate model. *Journal of Child Language*, *25*(2), 367-393.

Burchinal, M. (2018). Measuring Early Care and Education Quality. *Child Development Perspectives, 12*(1), 3-9.

Burchinal, M., Magnuson, K., Powell, D., & Hong, S. S. (2015). Early child care and education and child development. In M. Bornstein, R. Lerner & T. Leventhal (Eds.), *Handbook of child psychology and developmental science* (Vol. 4, 7th ed., pp. 223–267). Hoboken, NJ: Wiley.

Byers-Heinlein, K., & Lew-Williams, C. (2013). Bilingualism in the early years: What the science says. *Learning Landscapes, 7*(1)*,* 95-112.

Canh, L.V., & Renandya, W. (2017). Teachers’ English proficiency and classroom language use: A conversation analysis study. *RELC Journal*, *48*(1), 67-81.

Chen, X., Hao, M., Geva, E., Zhu, J., & Shu, H. (2008). The role of compound awareness in Chinese children's vocabulary acquisition and character reading. *Reading and Writing: An Interdisciplinary Journal, 21*, 559–586.

Cheng, R. L. (1985). A comparison of Taiwanese, Taiwan Mandarin, and Peking Mandarin. *Language*, *61*(21), 352-377.

Chimbutane, F. (2013). Codeswitching in L1 and L2 learning contexts: Insights from a study of teacher beliefs and practices in Mozambican bilingual education programmes. *Language and Education*, *27*(4), 314-328. doi: 10.1080/09500782.2013.788022

Clark, D.A. & Bowles, R.P. (2018). Model fit and item factor analysis: Overfactoring, underfactoring, and a program to guide interpretation. *Multivariate Behavioral Research*, *53*(4), 544–558.

de Bot, K. (1996). The psycholinguistics of the output hypothesis. *Language Learning*, *46*(3), 529-555. doi: 10.1111/j.1467-1770.1996.tb01246.x.

de Houwer, A. (2005). Early bilingual acquisition. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 30-48). New York: Oxford University Press.

DePalma, R. (2010). Language Use in the Two-way Classroom: Lessons from a Spanish-English Bilingual Kindergarten. Clevedon, UK: Multilingual Matters.

de Winter, J. C. F., Dodou, D., & Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research*, *44*(2), 147-181.

Dixon, L.Q. (2009). Assumptions behind Singapore's language-in-education policy: Implications for language planning and second language acquisition. *Language Policy, 8*(2)*,* 117-137.

Early Childhood Development Agency. “Requirements for teacher certification”. ECDA.gov.sg. Published November 2018. Accessed 12 December 2018. https://www.ecda.gov.sg/Documents/Requirements%20for%20Teacher%20Certification.pdf

Ellis, R., & Shintani, N. (2014). *Exploring language pedagogy through second language acquisition research*. London: Routledge.

Enama, P. (2016). The impact of English-only and bilingual approaches to EFL Instruction on low-achieving bilinguals in Cameroon: An empirical study. *Journal of Language Teaching and Research*, *7*(1), 19-30. doi: 10.17507/jltr.0701.03.

Gass, S. M., & Mackey, A. (2007). Input, interaction, and output in second language acquisition. In B. Van Patten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 175–199). New York/London: Routledge.

Gibbons, P. (2006). *Bridging discourses in the ESL classroom: Students, teachers and researchers*. London: Continuum Books.

Gort, M., & Pontier, R. (2013). Exploring bilingual pedagogies in dual language preschool classrooms. *Language and Education*, *27*(3), 223-245.

Halle, T., Whittaker, J. E. V., & Anderson, R. (2010). *Quality in Early Childhood Care and Education Settings: A Compendium of Measures, Second Edition.* Washington, DC: Child Trends. Prepared by Child Trends for the Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Hoff, E., & Core, C. (2013). Input and language development in bilingually developing children. *Seminars in speech and language*, *34*(4), 215–226.

Huang, B. H., & Kuo, L. (2020). The role of input in bilingual children's language and literacy development: Introduction to the special issue. *International Journal of Bilingualism, 24*(1), 3-7.

Ibbotson, P. (2013). The scope of usage-based theory. *Frontiers in Psychology*, *4*, 1-15.

Ishwaran, S., Shanmugam, K., Varaprasad, N., Sankaran, C., Lakshimi, S., Saravanan, V., & Peng, H. (2005). *Report of the Tamil Language Curriculum and Pedagogy Review Committee* (Rep.) Republic of Singapore: Ministry of Education.

Jiang, Y., García, G., & Willis, A. (2014). Code-Mixing as a bilingual instructional strategy. *Bilingual Research Journal*, *37*(3), 311-326. doi: 10.1080/15235882.2014.963738.

Kersten, K., Steinlen, A. K., Tiefentahl, C., Wipperman, I., & Mattsson, A. F. (2010). Guidelines for language use in bilingual preschools. In K. Kersten, A. Rohde, C. Schelletter & A. K. Steinlen (Eds.), *Bilingual preschools: Best practices* (pp. 103-116). Trier: WVT Wissenschaftlicher Verlag Trier.

Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. London: Longman.

Lahey, M. (1988). *Language disorders and language development*. London, UK: Macmillan.

Lahti, M., Elicker, J., Zellman, G., & Fiene, R. (2015). Approaches to validating child care quality rating and improvement systems (QRIS): Results from two states with similar QRIS type designs. *Early Childhood Research Quarterly*, *30*, 280-290.

Lakshmi, S., & Saravanan, V. (2009). *An examination of the use of Standard Spoken Tamil in Singapore – in the school and media domains in Tamil classrooms in order to establish SST as and additional resource for the teaching and learning of Tamil.* Final Research

Report for Project no. CRP 6/04 SL & CRP 10/06 SL. Centre for Research in Pedagogy & Practice, National Institute of Education, Nanyang Technological University, Singapore.

La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The Classroom Assessment Scoring System: findings from the prekindergarten year. *The Elementary School Journal*, *104*(5), 409-426.

Lin, A. M. Y. (2005). Critical, transdisciplinary perspectives on language-in-education policy and practice in postcolonial contexts: The case of Hong Kong. In A. M. Lin & P. W. Martin (Eds.), *Decolonisation, globalisation: Language-in-education policy and practice* (pp. 38–54). Clevedon, UK: Multilingual Matters.

Mashburn, A. J. (2017). Evaluating the validity of classroom observations in the Head Start Designation Renewal System. *Educational Psychologist, 52*(1), 38-49.

Ministry of Education. (2013). *Nurturing early learners – A curriculum for kindergartens in Singapore. Framework for Mother Tongue Languages.* Republic of Singapore.

Ministry of Education. (2015). Memupuk pelajar pada peringkat awal – Kurikulum untuk tadika di Singapura. Panduan pendidik Bahasa Melayu untuk prasekolah. Republic of Singapore.

Mukhlis, A. B., & Pang, E. (2015). Learning to be biliterate in English and Malay using dual-language books. (NIE Research Brief Series No. 15-013). Singapore: National Institute of Education.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide.* Eighth edition. Los Angeles, CA: Muthén & Muthén.

Ng, E. L., O’Brien, B. A., Khng, K. H., Poon, K. L. K., Karuppiah, N., Bull, R., Pang, E., Lee, K., Hwee, L. M., Tan, C.T., & Tan, G. H. (2014). *Singapore Kindergarten Impact Project*

(SKIP). OER 09/14RB, Office of Education Research, National Institute of Education, Singapore.

Pakarinen, E., Lerkkanen, M., Pikkeus, A., Kiuru, N., Siekkinen, M., Rasku-Puttonen, H., & Nurmi, J. (2010). A validation of the class assessment scoring system in Finnish Kindergartens. *Early Education and Development, 21*(1), 95-124. Doi: 10.1080/10409280902858764

Palviainen, A., Protassova, E., Mård-Miettinen, K., Schwartz, M. (2016). Two languages in the air: A cross-cultural comparison of preschool teachers' reflections on their flexible bilingual practices. *International Journal of Bilingual Education and Bilingualism, 19*(6), 614-630.

Pianta, R., La Paro, K., & Hamre, B. (2008). *Classroom Assessment Scoring System manual: Pre-K.* Baltimore, MD: Paul H. Brookes Pub. Co.

Pearson, B. Z. (2007). Social factors in childhood bilingualism in the United States. *Applied Psycholinguistics, 28*, 399-410.

Place, S., & Hoff, E. (2011). Properties of dual language exposure that influence 2-year-olds' bilingual proficiency. *Child Development, 82*(6), 1834-1849.

Probyn, M. (2015). Pedagogical translanguaging: Bridging discourses in South African science classrooms. *Language and Education*, *29*(3), 218-234. doi: 10.1080/09500782.2014.994525.

Raven, J. C. (1947). *Progressive matrices. Set A, Ab, B, book form*. London: H.K. Lewis.

Rickard-Liow, S. J., & Sze, W. P. (2008). *Bilingual Language Assessment Battery (BLAB).* In D. o. P. Singapore: Psycholinguistics Lab (Ed.): National University of Singapore.

Richards, J. C. (2010). Competence and performance in language teaching. *RELC Journal*, *41*(2), 101-122.

Schiffman, H. F. (2004). The Tamil case system. *South Indian horizons: felicitation volume for Francois Gros on the occasion of his 70th birthday*, 293-322.

Schwartz, M. (2018). Preschool bilingual education: Agency in interactions between children, teachers and parents. In M. Schwartz (Ed.), *Preschool Bilingual Education* (pp. 1-24). Cham, Switzerland: Springer.

Schwartz, M., & Gorbatt, N. (2016). 'Why do we know Hebrew and they do not know Arabic?' Children's meta-linguistic talk in bilingual preschool. *International Journal of Bilingual Education and Bilingualism*, *19*, 1–21.

Singapore Department of Statistics, Ministry of Trade and Industry. (2018). Population Trends, 2018. Retrieved from: https://www.singstat.gov.sg/-/media/files/publications/population/population2018.pdf

Subhan, M. A. (2013). *Bilingualism and its effects on Malay language planning. (Unpublished PhD thesis).* National Institute of Education Nanyang Technological University, Singapore.

Sun, H., Yin, B., Amsah, N. F. B. B., & O'Brien, B. A. (2018). Differential effects of internal and external factors in early bilingual vocabulary learning: The case of Singapore. *Applied Psycholinguistics*, *39*(2), 383-411.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. M. Gass & C. M. Madden (Eds.), *Input in Second Language Acquisition* (pp. 235-253). Rowley, MA: Newbury House.

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 471-483). Mahwah, NJ: Lawrence Erlbaum Associates.

Sze, P., & Leung, F. F. (2010). Enhancing learners' metalinguistic awareness of language form: The use of eTutor resources. *Assessment and Learning, 3,* 79-96.

Teng, A. (2018). Pre-schools turn to native speakers to meet demand. *The Straits Times.* Retrieved from: https://www.straitstimes.com/singapore/education/pre-schools-turn-to-native-speakers-to-meet-demand .

Tong, X., McBride-Chang, C., Shu, H., & Wong, A. M. Y. (2009). Morphological awareness, orthographic knowledge, and spelling errors: Keys to understanding early Chinese literacy acquisition. *Scientific Studies of Reading, 13*(5), 426-452.

Unsworth, S., Persson, L., Prins, T., & de Bot, K. (2015). An investigation of factors affecting early foreign language learning in the Netherlands. *Applied Linguistics, 36*(5), 527-548.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Weitz, M., Pahl, S., Mattsson, A. F., Buyl, A., & Kalbe, E. (2010). The input quality observation scheme (IQOS): The nature of L2 input and its influence on L2 development in bilingual preschools. In K. Kersten, A. Rohde, C. Schelleter & A. K. Steinlen, (Eds.) *Bilingual preschools volume 1: Learning and development* (pp.5-44). Trier: WVT Wissenschaftlicher Verlag Trier.

Wesche, M. (1994). Input and interaction in second language acquisition. In C. Gallaway & B. Richards (Eds.), *Input and interaction in language acquisition* (pp. 219-249). Cambridge: Cambridge University Press.

Wong, D. (2018). Scaling up supply of Mother Tongue pre-school teachers. *The Straits Times.* Retrieved from: https://www.channelnewsasia.com/news/singapore/scaling-up-supply-of-mother-tongue-pre-school-teachers-9880148.

Wright, W. E., Boun, S., & Garcia, O. (2015). Introduction: Key Concepts and issues in bilingual and multilingual education. In W. E. Wright, S. Boun & O. Garcia (Eds.), *Handbook of Bilingual and Multilingual Education*. Malden, MA: Wiley-Blackwell.

Yussof, N. T., & Sun, H. (2020). Mismatches between teacher perceptions, practices and reasons for English use in preschool Malay language classrooms. *Language & Education*. doi: 10.1080/09500782.2020.1720230.

Zellman, G. L., & Fiene, R. (2012). *Validation of quality rating and improvement systems for early care and education and school-age care* (Research-to-Policy, Research-to-Practice Brief OPRE 2012-29). Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Appendix A

Comparison of a set of selected observational rating tools and their contributions to the development of the Mother Tongue Adapted Coding Scheme (MACS)

Observation Tool	Aligned features that were adopted	Adaptations	Limitations
Classroom Assessment Scoring System (CLASS; Pianta, La Paro & Hamre, 2008)	• Application to preschool context • Relevant dimensions incorporated into MACS: -Concept development, positive climate, instructional learning formats, language modelling, quality of feedback • Scoring system: - 1-7 scale - Including depth, duration, and frequency of observed behavior -'Within one' reliability scoring	• Relevant dimensions included in different MACS domains (input, output, strategies) • Positive climate made specific to language learning • Modelling and scaffolding strategies modified to be appropriate for second or MT language learners • Modelling/scaffolding strategies extended to include effectiveness in terms of elicited child talk • Overall productive learning activities replaced with specific, independently scored varied teaching strategies • Low inference items scored on 1-3 scale of frequency	• A more holistic observation tool for preschool classroom experience, not specific to language teaching and learning • Behaviors related to bilingual contexts, such as language switching, not included
Input Quality Observation Scheme (IQOS; Weitz et al., 2010)	• Application to bilingual preschool context • Subset of appropriate domains incorporated into MACS (input, output, strategies) • Specific strategies for second language teaching included in MACS:	• Total number of domains reduced • Different scoring scales used for low and high inference items • Different observation and scoring procedures	• Emphasizes a target language-only input • Consideration of language switching as negative instance • Does not consider both negative and positive instances for use of children's stronger language

<table>
<tr><td></td><td>-Ritualized language; facial expression, gestures; pictures, objects, realia; adapted speech; variety & richness of input</td><td></td><td></td></tr>
<tr><td>Language Interaction Snapshot (LISn; Atkins-Burnett, Sprachman, & Caspe, 2010)</td><td>• Application to bilingual preschool context with children from differing home language proficiency levels
• Subset of domains incorporated into MACS</td><td>• Number of domains reduced
• Types of talk and activity types used as examples to observe for input and output</td><td>• Observation focus on individual child ‘snapshots’ rather than classroom level
• More detailed set of types of talk by children and which language used for each
• Language switching scored at a more detailed scale of types of talk</td></tr>
<tr><td colspan="4">Additions unique to MACS</td></tr>
<tr><td colspan="2">Additional aspects</td><td colspan="2">Purpose for the additions</td></tr>
<tr><td>Language correctness</td><td>Gauging quality through correctness of pronunciation, language choice, syntax etc.</td><td colspan="2">Correct and fluent input considered key for language teaching and learning. Considered impact of colloquial language use (Singlish) that may impact teacher input.</td></tr>
<tr><td>Language appropriateness</td><td>Addressing concerns regarding diglossia in the classroom
• Formal language for literacy instruction
• Informal language for developing communicative skills</td><td colspan="2">Considered impact of diglossia in local languages that may impact children’s language resources, and what kinds of language children need to be exposed to for school preparation, but also to facilitate their communicative skills within the relevant language communities</td></tr>
<tr><td>Teachers’ English use</td><td>Attempts at optimizing English use in MTL classrooms</td><td colspan="2">Given the dominance of English as the societal language, and studies that reflect how target-language only classrooms are rare, instructional purposes for English use were observed</td></tr>
<tr><td>Two scoring scales</td><td>Two different scales were used to facilitate scoring of low and high inference items</td><td colspan="2">Different evaluation approaches different types of variables:
• 1-3 point scale for observing behavioural frequency
• 1-7 point scale for observing the level of teacher facilitation</td></tr>
</table>

Appendix B

Details and sensitivity of child mother tongue language (MTL) outcome measures.

Vocabulary. The locally developed measure was sensitive to change over time, according to the mean and range of scores for the whole sample: Means for Chinese at K1 = 32.3, K2 = 38.6, and P1 = 43.6; for Malay at K1 = 30.8, K2 = 37.7, and P1 = 41.9; and for Tamil at K1 = 25.6, K2= 31.1, and P1 = 35.9. There was also no indication of floor or ceiling effects. Minimum-maximum scores at each time point = 13-60, 17-66, 17-71 for Chinese; 13-54, 19-55, 17-57 for Malay; and 12-47, 12-47,18-61 for Tamil. The 25th-75th percentile scores at each time point = 26-38, 32-45, 37-51 for Chinese; 25-37, 32-43, 37-48 for Malay; and 21-30, 25-37, 29-43 for Tamil.

Reading. For Chinese, grapheme recognition was based on the character discrimination test of the Preschool and Primary Chinese Literacy Scale (Li, H. (2015). *Teaching Chinese literacy in the early years*. New York, NY: Routledge) and included 19 items where the child had to match one of four characters to the spoken word. Children then completed a naming task where they gave the names of 15 strokes and radicals, and a character reading task, where they read up to 100 characters. The word list was administered in blocks of 25 characters, with words decreasing in frequency over the blocks (according to the corpus of Loo, 1989). The task was discontinued when the child could not read any words in a block. Total correct responses were summed over the three tasks.

For Malay, children completed a letter naming task with 11 items from the Latin script, and a word reading task where they read up to 100 words. The word list was administered in blocks of 25, with words decreasing in frequency over the blocks according to the corpus of Lee and Low (Lee, L. W., & Low, H. M. (2011). Developing an online Malay language word corpus for primary schools. *International Journal of Education and Development using Information and*

Communication Technology, 7(3), 96-101), and with increasing word complexity within each block (in terms of word length and syllable structure). The task was discontinued when the child was not able to read any words in a block. Total correct responses were summed over the two tasks.

For Tamil, grapheme recognition included a letter discrimination task with 18 items, where the child was shown four glyphs and had to match one to the spoken letter for each item. Children then completed a naming task where they gave the names of 12 Tamil letters (aksharas), and a word reading task, where they read up to 100 words. The word list was administered in blocks of 20, with words increasing in difficulty (according to grade level literacy resource materials, from the National Institute of Education and public libraries). The task was discontinued when the child was not able to read any words in a block. Total correct responses were totalled across the three tasks.

These locally developed reading measures showed sensitive to change over time. Mean total raw scores over time for the whole sample are as follows: For Chinese at K1 = 28.3, K2 = 35.5, and P1 = 40.7; for Malay at K1 = 11.6, K2 = 19.1, and P1 = 57.9; and for Tamil at K1 = 20.4, K2 = 32.9, and P1 = 59.5. The scores did not yield floor or ceiling effects. The range (minimum to maximum) at each time point was 0-96, 0-99, 0-111 for Chinese; 0-80, 0-108, 0-108 for Malay; and 1-88, 2-118, 0-122 for Tamil, respectively. The 25th-75th percentile scores at each time point = 14.3-40, 20-49, 23.5-57 for Chinese; 7-11,10-14, 28-90 for Malay; and 6.5-29, 9-50.8, 39-85 for Tamil.

Morphological awareness. For each language, compound production requires the child to to generate as many compound words as they could from a base word (e.g. the word ‘book’ could be used to make the words ‘bookstore’ and ‘bookshelf’). There were 15 items in Chinese, 10 items in Malay, and 8 items in the Tamil tasks, and total correct responses across all items were scored. For the compound structure task, the experimenter would describe a scenario in a sentence, followed by multiple choices for the child to choose to best describe the scenario. For example, "A house that is built in a tree: It is a treehouse or a housetree?" There were 13 items for Chinese, 11 items for Malay, and 14 items for Tamil. Total correct items were summed and the total score was calculated with a correction for guessing. The sum of compound production and compound structure task scores was taken as the total raw score.

Mean total scores for the whole sample indicate that the task was sensitive to change over time: for Chinese at K1 = 10.2, K2 = 14.0, and P1 = 18.1; for Malay at K1= 5.4, K2 = 7.4, and P1 = 9.3; and for Tamil at K1 = 3.4, K2 = 7.2, and P1 = 13.5. The range of scores suggested no floor or ceiling effects: Minimum-maximum scores at each time point = 0-36, 0-35, 0-43 for Chinese; 0-15, 0-22, 2-23 for Malay; and 0-14, 0-15, 4-32 for Tamil. The percent of 0 scores = 3.1, 0.6, 0.3 % for Chinese at K1, K2, P1; 2.2, 4.4, 0.7% for Malay; 25, 0.6, 0% for Tamil. The 25th-75th percentile scores at each time point = 5-14, 10-18, 12-23 for Chinese; 4-7, 5-9, 6.3-11 for Malay; and 0-5, 6-9, 9-16 for Tamil.

Appendix C

Growth mixture models to identify latent classes of MTL learners

Growth mixture models (GMMs) with 1 to 3 classes were fit to data over 3 waves (K1, K2, P1) using Mplus v.8 Software with an MLR estimator (Muthén & Muthén, 2017). Students were clustered in classrooms for all models (although intraclass correlation coefficients indicated a design effect for classroom level only for Chinese reading and Chinese and Tamil morphological awareness). For each Mother Tongue Language (MTL) group, model fit was compared for a 2-class model compared to 1-class or 3-class models. The best fitting model in each case was considered across several criteria, including model fit, likelihood ratio tests (Vuong-Lo-Mendell-Rubin Likelihood Ratio Test, Lo-Mendell-Rubin Adjusted Likelihood Ratio Test, p-values < 0.05), entropy (> 0.8), and the proportion of students within the smallest latent class (> 1%). In addition, plots were inspected to determine interpretability of the pattern of results (e.g., Fu, R., Chen, X., Wang, L., & Yang, F. (2016). Developmental trajectories of academic achievement in Chinese children: Contributions of early social-behavioral functioning. *Journal of Educational Psychology*, *108*(7), 1001-1012). Fit indices are shown in Table C1 below. Further, the number of children with incoming scores below the low progress group's intercept, above the high progress group's intercept, and in between intercepts is also shown in Table C2.

Table C1.

Growth mixture model fit indices for 2-class versus 1-class model, per variable for each language group

		Likelihood Ratio Tests (p-values)				Classification Accuracy						Comparison to 3-class model	
	n	VLMR-LRT	LMR-LRT	Smallest Class (%)	Entropy	Min	Max	Class 1 intercept	Class 1 slope	Class 2 intercept	Class 2 slope	2- vs. 3-class model	3-class entropy
Vocabulary													
CL	*1074*	0.000	0.000	37.9%	0.747	0.910	0.935	27.6	5.34	40.4	5.89	**	.690
ML	*194*	0.179	0.193	39.8%	0.570	0.821	0.887	27.1	3.87	33.7	6.70	*ns*	
TL	*236*	0.000	0.000	40.2%	0.733	0.905	0.928	29.6	7.67	22.6	3.67	*ns*	
Reading													
CL	*323*	0.003	0.003	38.4%	0.822	0.933	0.955	17.8	5.13	45.4	8.41	*ns*	
ML	*137*	0.017	0.021	12.4%	0.995	0.996	1.000	6.25	7.84	35.7	38.19	*ns*	
TL	*160*	0.267	0.280	19.3%	0.928	0.949	0.981	53.1	23.1	11.5	15.02	*ns*	
Morphological Awareness													
CL	*319*	0.000	0.000	19.9%	0.777	0.867	0.956	19.5	1.16	7.8	4.64	*ns*	
ML	*137*	0.016	0.020	13.6%	0.798	0.885	0.949	5.6	1.29	4.5	6.00	*ns*	
TL	*161*	0.178	0.194	15.9%	0.744	0.857	0.946	2.3	4.69	7.3	3.16	*ns*	

Note. Likelihood ratio tests compared 1-class to 2-class models (VLMR-LRT = Vuong-Lo-Mendell-Rubin likelihood ratio test; LMR-LRT = Lo-Mendell-Rubin likelihood ratio test). Smallest class percent, entropy, classification accuracy, and class intercepts and slopes are from the 2-class models. Comparing 2-class model vs. 3-class model fits with LRT: *ns* = 3-class was not a better fit; **p* < 0.05, ** *p* < 0.01, meaning 3-class model yielded a better fit, but had lower entropy than the 2-class model. CL= Chinese language learners; ML= Malay language learners; TL= Tamil language learners.

Table C2.

Number of children scoring above, below or between the intercepts per latent class for Mother Tongue Language (MTL) measures

	Below latent "low progress" class intercept	Between latent class intercepts	Above latent "high progress" class intercept
Vocabulary	157	183	107
Reading	92	175	74
Morphological Awareness	48	143	146

Note. CL= Chinese language learners; ML= Malay language learners; TL= Tamil language learners.