Predicting toddler performance on a novel word learning task through engagement

Aaron Richard Glick

Master of Science
School of Communication Sciences and Disorders
McGill University
Montreal, Canada

November 2021

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Master of Science.

ABSTRACT

Young children engage with screens multiple times a day, so it is important to study the factors impacting children's ability to learn from screens. Engagement is considered a necessary condition for learning (Aguiar & McWilliam, 2013), yet it has typically been reported only as a secondary result in studies assessing word learning in online contexts (see Myers, LeWitt, Gallo, & Maselli, 2017; Roseberry, Hirsh-Pasek, & Golinkoff, 2014; Strouse, Troseth, O'Doherty, & Saylor, 2018; Troseth, Saylor, & Archer, 2006). Social contingency and co-viewing moderate learning outcomes in children, particularly in online environments (Myers et al., 2017; Roseberry et al., 2014; Strouse et al., 2018); however, these factors do not analyze how the child is engaging with the task. Here, we ask whether the *children's level of engagement* moderates learning in online contexts, predicting learning outcomes.

We conducted a word learning study remotely with children at home and collected data via webcam. Twenty-seven children (21- to 46-months, M = 37-months) saw four novel objects, two of which were labelled, via prerecorded video or live video-chat. We tested word learning by examining their responses on two types of tests: (1) label production: whether the children could produce the novel labels for the objects and (2) referent selection: whether the children could select the novel object from a set of objects when given the novel label. Because verbal production was limited, we analyzed the referent selection measure, as observed through both the child's gaze and gestural responses. We measured the child's engagement through both (a) observational coding of verbal and gestural behaviors and through (b) parent ratings and (c) experimenter ratings of holistic engagement (emotional, cognitive, and behavioral). With regression models, we analyzed if engagement measures (a, b, c) predict referent selection outcomes. Preliminary results suggest that engagement measures may moderate referent selection outcomes, yet no single measure reliably moderated outcomes;

instead, there were trends between age, observational, and parent rating measures and trends between observational and experimenter rating measures.

Since engagement is difficult to measure, we extended our analysis through an exploratory factor analysis (EFA), which is particularly useful in reducing dimensionality. The EFA allows us to analyze which combination of engagement metrics can better describe the child's engagement and word learning outcomes. Preliminary results suggest that engagement measures may be able to distinguish different word learning outcomes in younger children as compared to older children. The findings from this study contribute to the limited works theorizing how engagement moderates word learning, suggesting that child engagement may improve outcomes in learning contexts.

RÉSUMÉ

Les jeunes enfants utilisent des écrans plusieurs fois par jour, il est donc important d'étudier les facteurs qui influencent la capacité des enfants à apprendre à partir d'écrans. L'engagement est considéré comme une condition nécessaire à l'apprentissage (Aguiar & McWilliam, 2013), mais il n'a généralement été signalé que comme un résultat secondaire dans les études évaluant l'apprentissage des mots dans les contextes en ligne (voir Myers et al., 2017; Roseberry et al., 2014; Strouse et al., 2018; Troseth et al., 2006). La contingence sociale et le co-viewing modèrent les résultats de l'apprentissage chez les enfants, en particulier dans les environnements en ligne (Myers et al., 2017; Roseberry et al., 2014; Strouse et al., 2018) ; cependant, ces facteurs ne prennent pas en considération pas la façon dont l'enfant s'engage dans la tâche. Ici, nous nous demandons si le *le niveau d'engagement des enfants* modère l'apprentissage dans des contextes en ligne, en prédisant les résultats d'apprentissage.

Nous avons mené une étude sur l'apprentissage des mots à distance avec des enfants à la maison et recueilli des données par webcam. Vingt-sept enfants (âgés de 21 à 46 mois, $M = 37$ mois) ont vu quatre objets nouveaux, dont deux étaient étiquetés, par le biais d'une vidéo préenregistrée ou d'un vidéo-chat en direct. Nous avons testé l'apprentissage des mots en examinant leurs réponses à deux types de tests : (1) la production d'étiquettes : si les enfants pouvaient produire les nouvelles étiquettes pour les objets et (2) la sélection du référent : si les enfants pouvaient sélectionner le nouvel objet parmi un ensemble d'objets lorsqu'on leur donnait la nouvelle étiquette. La production verbale étant limitée, nous avons analysé la mesure de la sélection des référents, telle qu'observée par le regard des enfants et leurs réponses gestuelles. Nous avons mesuré l'engagement de l'enfant par (a) le codage observationnel des comportements verbaux et gestuels et par (b) les évaluations des parents et (c) les évaluations de l'expérimentateur de l'engagement holistique (émotionnel, cognitif et comportemental). À l'aide de modèles de régression, nous avons analysé si les mesures d'engagement (a, b, c) prédisent les résultats de la sélection des référents. Les résultats préliminaires suggèrent que les mesures d'engagement modèrent les résultats de la sélection de référents, mais aucune mesure unique ne modère de manière fiable les résultats ; au contraire, il y a eu des tendances entre les mesures d'âge, d'observation et d'évaluation des parents et des tendances entre les mesures d'observation et d'évaluation de l'expérimentateur.

L'engagement étant difficile à mesurer, nous avons étendu notre analyse par une analyse factorielle exploratoire (EFA), qui est particulièrement utile pour réduire la dimensionnalité. L'EFA nous permet d'analyser quelle combinaison de mesures d'engagement peut mieux décrire l'engagement de l'enfant et les résultats de l'apprentissage des mots. Les résultats préliminaires suggèrent que les mesures d'engagement peuvent être en mesure de distinguer différents résultats d'apprentissage des mots chez les jeunes enfants par rapport aux enfants plus âgés. Les résultats de cette étude contribuent aux travaux limités qui théorisent la façon dont l'engagement modère l'apprentissage des mots ainsi que suggérant que l'engagement de l'enfant peut améliorer les résultats dans les contextes d'apprentissage.

ACKNOWLEDGMENTS

I would like to extend my thanks to my advisor, Dr. Aparna Nadig, for her guidance and support during my research at McGill University. The long discussions we had were all insightful, inspiring, and joyful. She was reliable, rock steady, and resilient despite the many challenges we faced transitioning this project online because of the COVID-19 pandemic, and I am so appreciative of her guidance and mentorship. I am also thankful for the financial support she provided. My special thanks to my MSc committee member Dr. Susan Rvachew and my external reviewer Dr. Adam Dubé. Their thoughtful comments on my work are greatly appreciated. I would also like to thank Dr. Kristine Onishi for her thoughtful, insightful mentorship and countless contributions to the SCREEN project. Many undergraduate students made it possible to collect the data presented below. In particular, I would like to thank Ekaterina Tsoradis, Ella Vezina, Alicia Martineau, Fauzia Saiyed, and Gabriella Boffice. The many hours we spent designing, configuring, reconfiguring, and discussing every aspect of the SCREEN project will be cherished and remembered fondly.

And last but not least, my deepest appreciation goes to my wife Kaylin Land, for her endless love and patience, and to my family whose unconditional love and support was a great source of motivation throughout challenging times.

Contents

List of Figures

List of Tables

Introduction, background, and research aims

In word learning studies, a child is taught a novel word and shown the object or action it corresponds to. Later the child is tested to observe if they can produce the novel word or identify the object when provided the novel word. By manipulating factors such as how the object is presented by the teacher (e.g., with/without attentional bids like pointing or gesturing) or the type of presentation media (e.g., in-person, video chat, or pre-recorded video) during this simple teaching demonstration, these word learning paradigms help to investigate what factors help children learn.

In the context of an adult teaching a new word to a child, researchers distinguish social factors like the adult's accuracy, timing, and relevance of social responses (a package of behaviors that has been termed "social contingency", see Baldwin, 1994; Myers et al., 2017; Nielsen, Simcock, & Jenkins, 2008; Roseberry et al., 2014; Troseth et al., 2006) as a primary moderator of learning in this context. In a similar vein, other social partners who watch the teaching interaction with the child (i.e., co-viewing, see Myers, Crawford, Murphy, Aka-Ezoua, & Felix, 2018; O'Doherty et al., 2011; Strouse et al., 2018) may also increase word learning outcomes. When an adult (a responsive co-viewer) watches the teaching session with the child, the child shows significantly better word learning (compared to having an unresponsive co-viewer). Social cues from both the teacher and other viewers matter; children's responsiveness is sensitive to whether a co-viewer had contingent gaze during the session (Myers et al., 2018). That is, social cues made an impact on the child's learning only when the co-viewer looked at the screen and responded at appropriate times, suggesting that both social contingency and co-viewing are important but individually neither explains the full story (for children 24-30 months old; Myers et al., 2018). While both social contingency and social cues from an adult have been shown to moderate word learning, these factors do not ensure learning, leaving the door open for other factors that may affect how children learn (Smith & Yu, 2013; Troseth, Strouse, Verdine, & Saylor, 2018). The wide range of learning outcomes may reflect different types of learning (Bannard & Tomasello, 2012).

1.1 Background

There may be alternative moderators to word learning beyond social contingency and co-viewing.[1] For example, the child's engagement during the session could impact word learning outcomes. While there are many context-specific definitions of engagement (see Appleton, Christenson, & Furlong, 2008; Busselle & Bilandzic, 2009; Richardson et al., 2020; Yazzie-Mintz, 2007), here we consider engagement as the ability to stay on task or remain undistracted (Hirsh-Pasek et al., 2015) including the effort and willingness to take part in an experience (Lauricella, Blackwell, & Wartella, 2017).

Engagement is often considered a necessary condition for learning (Aguiar & McWilliam, 2013), yet is typically reported as a secondary result in word learning studies (see Myers et al., 2017; Roseberry et al., 2014; Strouse et al., 2018; Troseth et al., 2006). However, the teacher's social contingency and the child's engagement may be linked but distinct sides of

[1] A moderating factor is variable that affects the strength and/or direction between an independent variable and a dependent variable (Baron & Kenny, 1986).

the same interaction. The impact of engagement in these contexts has not been explored as a unique factor like social contingency or social cues, and to my knowledge has not been explicitly modeled as a moderating factor of performance on word learning tasks. It has, however, been modeled for language classrooms (Language Task Engagement model, Egbert, Shahrokni, Abobaker, & Borysenko, 2021). The Language Task Engagement model, shown in Figure 1.1, was proposed by Joy Egbert with children learning English as a second language, and children's interactions with technology while learning the fundamentals of coding (Egbert, 2020a, 2020b). Notably, in 2nd grade classroom lessons teaching coding, through the Language Task Engagement model, there was a significant increase between pre- and post-test scores ($d = 2.43$; Egbert et al., 2021). Correlations between engagement and individual learning outcomes were not reported. However, self-report surveys indicated that children were focused and engaged on the task because they were interested and had autonomy while completing the task, but also had relevant learning supports to complete the challenging task. The model suggests three main elements that may impact the child's engagement and ultimately predicts learning outcomes.

First, there is the *language task engagement facilitator* element, which focuses primarily on how the teacher frames the learning task in the context of the lesson and how the teacher adapts it for the social setting. This is closely integrated with the *language task engagement* element which focus more on the specific learner. These two elements of the language task are the primary focus of most word learning studies, which focus on factors like social contingency (the authenticity, social interaction, and learning strategies of the learning session) and co-viewing (the social interaction and support given by others). These two elements are shown in yellow in Figure 1.1. Importantly, this model also includes the *level of language task engagement* element which adds in the behavioral, cognitive, and emotional aspects of engagement that are important factors for predicting learning outcomes (Appleton et

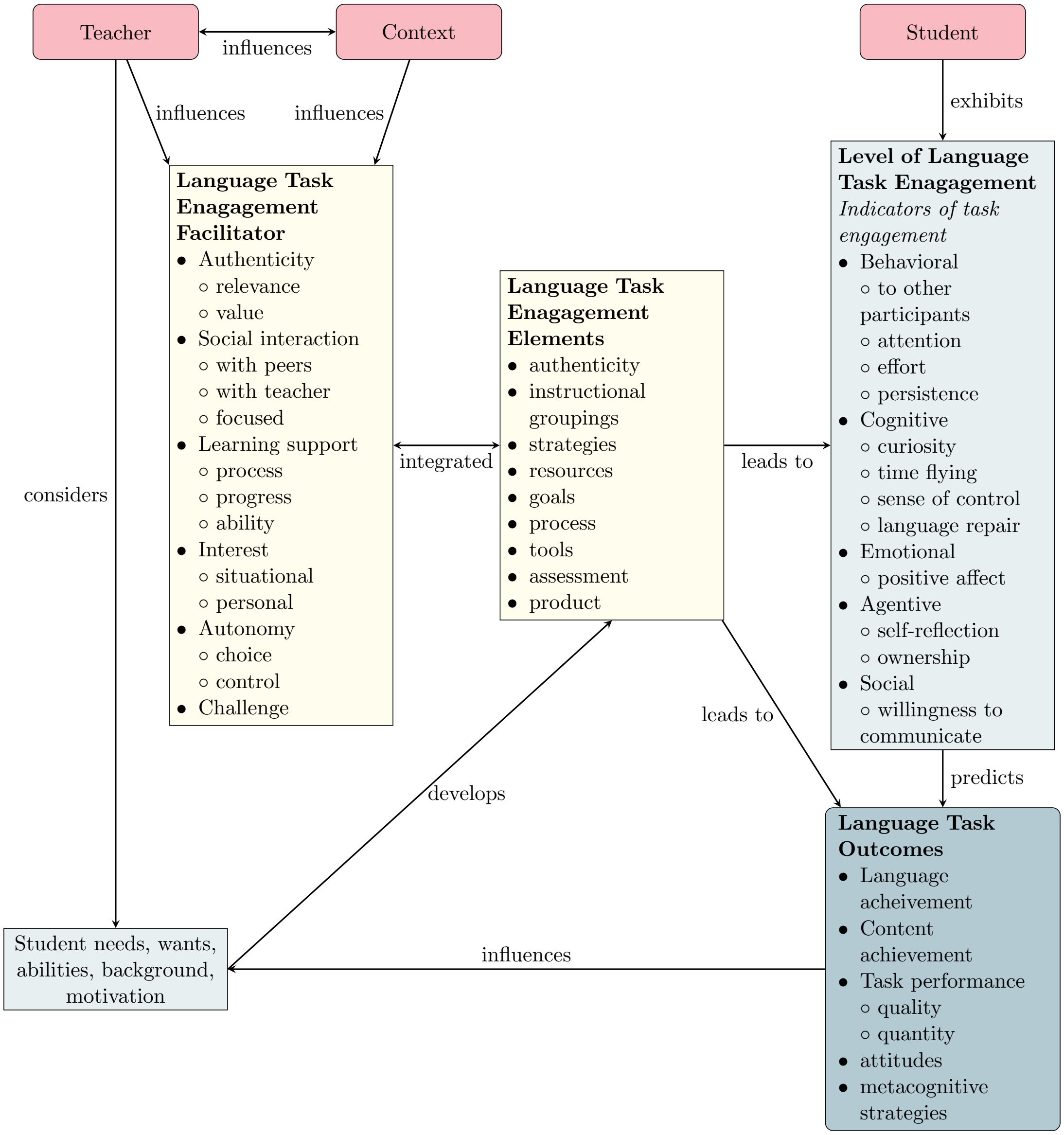

Fig. 1.1 Adapted language task engagement model.

Note: Adapted by Aaron Glick from Egbert et al. (2021), which is licensed under Creative Commons Attribution-Non Commerical 4.0 International License.
The construct of social contingency, not explored in this thesis or explicitly in this model, overlaps with aspects of Language Task Engagement Facilitator and Language Task Engagement Elements, in yellow. This thesis explores how the student's engagement (here the Level of Language Task Engagement, in blue) moderates Language Task Outcomes.

al., 2008; Hirsh-Pasek et al., 2015; Thijs & Verkuyten, 2009). Together these three primary elements of the language task interaction lead to the *language task outcomes*, or word learning in our case, that we were interested in predicting.

The model suggests there is a critical interaction between the teacher's active role and the engagement elements of the lesson being presented; these two elements capture the effects of social contingency seen in word learning studies. Teacher responses need to be relevant and accurate in regard to the social interaction and child's interest, while maintaining authentic and supportive prompts to scaffold the child's learning; these aspects lead to language task outcomes. Importantly, the model adds the additional layer of **how the child is engaging with the task**. It suggests that the child's effort, willingness, and curiosity also predict language task outcomes. Similarly, in word learning studies, engagement should be considered a unique factor in predicting word learning outcomes. We anticipate that engagement will have a positive impact on learning outcomes. The primary aim of my project was a moderator analysis of engagement on word learning performance.

1.1.1 Prior work on engagement from word learning studies

As established, most word learning studies do not separate a child's engagement explicitly from an adult's social contingency, co-viewing, or joint attention. Therefore, I analyzed these study results from a different perspective, a child-focused perspective, and examined engagement during the learning interaction.

Infants are sensitive to contingent interactions, as described below, and the behaviors and emotional reactions they display in reaction to social contingency can be used for measuring engagement. For example, one-month-old infants will show less positive affect, increased negative affect, and higher rates of averted gaze when a social partner suddenly becomes unresponsive (Bertin & Striano, 2006 for reviews see Adamson & Frick, 2003; Mesman, van

IJzendoorn, & Bakermans-Kranenburg, 2009), and two-month-old infants will grimace or avert their gaze when their mother's behavior is interrupted with non-contingent interactions (Soussignan, Nadel, Canet, & Gerardin, 2006). Three- and four-month-old infants will begin noticing when caregivers do not respond contingently and will change their own vocalizations in response (Bloom, Russell, & Wassenberg, 1987; Masataka, 1993; Toda, 1993). These gestural and vocal responses are some of the first signs an infant is engaging with their environment socially.

Later in their first year, infants start to share and direct the attention of people around them. Striano and Stahl (2005) found that 3-, 6-, and 9-month-olds smiled more when an adult shared joint attention over a toy, that is, when the adult looked back and forth between the child and object compared to when the adult only looked toward the toy. The child also gazed longer at the object when the adult looked back and forth and had the longest gaze when the adult was also talking about the object while looking back and forth. This suggests that infants are sensitive to these gaze and verbal cues for joint engagement (i.e., alternating gaze and verbal cues) and will hold attention longer and smile more when they are present. These results have recently been supported and extended to video chat and pre-recorded video (McClure, Chentsova-Dutton, Holochwost, Parrott, & Barr, 2020). From 6-18 months, infants continue responding more to joint attention, which is thought to reflect the child's engagement during these social transactions (Carpenter, Nagell, Tomasello, Butterworth, & Moore, 1998; Mundy et al., 2007). This evidence suggested social contingency and engagement are important from an early age.

Word learning studies show this continues to be true through the early years of childhood (e.g., 12-25 months see Myers et al., 2017; 24-months see Nielsen et al., 2008; 24-30 months see Myers et al., 2017; 24-72 months see Nielsen et al., 2008). These results suggested social contingency and engagement may be intertwined, but that they focused on different aspects of the same social interaction—in the context of word learning, social contingency focuses more on the actions and behaviors of the teacher whereas engagement focus more on the actions and behaviors of the child. Existing word learning studies heavily focused on the adult's behavior (the social contingency, the co-viewing, and joint attention) since these factors are easy to manipulate; however, engagement which is driven by the child's behavior could also moderate word learning outcomes.

Word learning studies are particularly useful in examining the role of engagement and social contingency on learning, since word-object mapping indicates a child's awareness of a relationship between the word and an object (Bannard & Tomasello, 2012). As mentioned above, most word learning studies looked at a child's engagement primarily through measures of joint attention (Myers et al., 2017; Roseberry et al., 2014; Strouse et al., 2018) or joint engagement (Myers, Keyser, & Cors, 2019; O'Doherty et al., 2011). Joint attention measures, quantified by the child's gaze, sometimes using eye tracking, whereas joint engagement measures typically include other behavioral observations like utterances and gestures and participation. Recent studies have started to directly measure infant's engagement beyond observational measures by adding physiological measures (skin conductance, heart rate, and respiration rate, see McClure et al., 2020). Though initially planned for the present study, my project had to be conducted completely remotely, due to the COVID-19 pandemic, preventing the collection of physiological measurements. Unfortunately, in the context of a completely online study, there were not many studies which set the precedents and standards for infant word learning studies. Studies have looked explicitly at engagement in the context of word learning on tablet devices (Russo-Johnson, Troseth, Duncan, & Mesghina, 2017), which measure engagement observationally through the child's physical interactions with the activity on the tablet. This is a substantially different context from our word learning environment, where the child's interactions do not trigger their progression through the activity, and the teacher's lesson progresses even when a child watches passively and does not engage. To my knowledge, no word learning studies in toddlers have looked at the moderating effects of engagement on word learning performance in conventional word learning paradigms in the lab where stimuli are delivered irrespective of the child's response (e.g., the child could be passive or responsive).

1.1.2 Behavior driven paradigms of engagement

There is a separate body of literature that focuses on engagement in the classroom, focusing primarily on school-age students. This work typically defines engagement metrics in three categories: behavioral, emotional, and cognitive (Appleton et al., 2008; Hirsh-Pasek et al., 2015; Thijs & Verkuyten, 2009). Each is critical for learning, so we incorporated metrics of each category in this study. Typically, behavioral engagement is indicated through a child's conduct (following rules, adhering to norms, lack of disruptive behaviors), involvement in a task (effort, persistence, concentration, attention, asking questions, contributing to conversations), and participation (Finn & Zimmer, 2012; Fredricks, Blumenfeld, & Paris, 2004). Social skills play a large role in measuring behavioral engagement, particularly in terms of conduct and involvement (Grassmann, 2014). Indicators of behavioral engagement may be outward, i.e., a child initiating an interaction towards an object or another person, a child's responsiveness to prompts, or simply the number of verbal/gestural responses. Word learning studies have looked at social imitation in children (24 months old) and shown that they are more likely to copy behaviors with a social contingent partner than in a non-interactive context (Nielsen et al., 2008). Similar results have been extended to children aged 12-25 months, and show children have more synced behavior with an adult in video chat than pre-recorded video (Myers et al., 2017). There are also internal measures of behavioral engagement like interest in activity and focus on a task, typically self-reported by students

(Appleton et al., 2008; Yazzie-Mintz, 2007).

There are very few models that describe engagement during a word learning task, and none that describe the interaction specifically in young children. However, we can extend aspects of the language task engagement model (Figure 1.1) to the context of word learning studies to collect data to analyze the child's engagement. As previously motivated, this model combines the aspects of social contingency that have been shown important to learning outcomes in word learning studies (Baldwin, 1994; Myers et al., 2017; Nielsen et al., 2008; Roseberry et al., 2014; Troseth et al., 2006) with the three categories of engagement (behavioral, emotional, and cognitive) from the student literature (Appleton et al., 2008; Hirsh-Pasek et al., 2015; Thijs & Verkuyten, 2009). This provides a theoretical framework to bridge these two separate, but related, bodies of literature together.

1.1.3 Measuring engagement for remote word learning studies

In order to test whether engagement moderates word learning outcomes, it was important to define the measures that indicate a child's level of engagement. Most of the existing literature is on student engagement in the classroom and typically focused on the behavioral, emotional, and cognitive aspects of engagement (Appleton et al., 2008; Egbert, 2020b; Thijs & Verkuyten, 2009); however, these categories were too broad and provide little utility for the simple teaching demonstration and assessment in word learning tasks for toddlers because they do not link the learning outcome directly to the child's engagement (like the language task engagement model proposes). For our context, engagement is centered around a triadic interaction between the adult, the child, and the object they are learning about, where the child is sharing attention to an object with another person. Most studies classify this situation of engagement using the term joint attention (Myers et al., 2017; Roseberry et al., 2014; Strouse et al., 2018), joint engagement (Myers et al., 2019; O'Doherty et al., 2011), symbol-infused joint engagement (Adamson, Bakeman, Deckner, & Nelson, 2013), or mutual engagement which are all measure engagement in a triadic interaction between child, adult, and object. While each term may highlight certain aspects of the child's engagement, the definitions are highly interrelated and often interchangeable. These terms are not used consistently across the literature, despite the overlapping definitions, making it difficult to compare measures of engagement between studies. Therefore, we propose two approaches of measuring engagement that can be operationalized across studies: observational measures and holistic measures. These two approaches help to classify the overall engagement state of the child during the word learning task in two ways. First, they distinguish the method of data collection: observational measures are applied to countable physical behaviors and holistic measures are survey-based ratings of gestalt perception/holistic impression. Second, they distinguish what kind of judgement is inherent in the data; observational measures are assigned by independent/objective coders and holistic measures are subjective responses reported from two perspectives (self-report from the parent and from the experimenter). Table 2.2 outlines the observational measures and holistic measures of engagement that we used, which are described in Section 2.6.2. Observational measures of engagement were collected through specific observations of the child's behavior and responses and are similar to the behavioral measures used in word learning studies (O'Doherty et al., 2011; Strouse et al., 2018; Troseth et al., 2006), with the important distinction that our video of the child during the task was obtained via webcam in our fully remote study, versus from video obtained in the lab.

During teaching phases (consisting of a warm up subphase and teaching subphase), there were specific bids that encourage the child to engage in the activity (exact wording of bids detailed in Table B.2). During the testing phases we ask the participants specific questions to test their word learning. As we were interested in the role of engagement on word learning, we coded for engagement for the entire experiment, including both teaching and testing phases. Most studies focus only on engagement during the interactive teaching session and describe them as measures of the child's participation (see Myers et al., 2017; O'Doherty et al., 2011; Roseberry et al., 2014; Strouse et al., 2018). A child's interest and willingness, however, may change between the socially interactive teaching phase and the non-contingent testing phase, which is why we collected engagement data during both phases.

Collecting engagement during both teaching and testing phases also is a way to indirectly measure the cognitive processing of children during the task, and may reveal learners who are in a transitional state and are ready to learn (Goldin-Meadow, 2020). Gestures can provide a reliable signal that the child could benefit from further instruction on a task, particularly when their gestures and speech are mismatched and convey different information (Church & Goldin-Meadow, 1986; Goldin-Meadow, 2020; Pine, Lufkin, & Messer, 2004). For instance, children 3- to 5-years old who had gesture-speech mismatches when labeling sets (counting or returning objects) were shown to benefit more from enriched instruction (Gibson, Gunderson, Spaepen, & Goldin-meadow, 2020). For our novel word learning task, the teacher may respond to the child's gestures contingently during teaching; however, no feedback will be given during the testing phase. In the cases where children's verbal responses do not reflect the child learning the novel words, gestures during testing may indicate that the child is still processing the task and has acquired some understanding of the novel objects presented.

From these two approaches of engagement measures, we collected data that investigated both the child's specific interactions, e.g., observational measures like the number of verbal responses they produced during the teaching episode (obtained from recording their behaviors), and an overall perception of the child's affect and attitude during the session, e.g., holistic measures like interest in activity reported from the parent (since children were too young to self-report) and from the experimenter. These approaches conveniently allowed measures from different sources to be aggregated based on similar collection methodologies and then combined into an overall score of engagement.

1.2 Research aims and predictions

Research Aim 1 The first research goal was to determine whether engagement moderates performance on a novel word learning task. We expected that children who were more engaged during the teaching and testing phases would also have higher performance on the word learning tasks. With any moderator analysis, it is best practice to base the overall effect on an established trend; therefore, we used the established positive correlation that older children show higher performance on word learning tasks. Evidence shows that older toddlers (24- to 36-months) have higher performance on word learning tasks than younger toddlers (12- to 24-months) (Myers et al., 2017; Nielsen et al., 2008). We tested the moderating effects of engagement on word learning, to see if higher engagement impacts learning outcomes for toddlers (20- to 48-months). As we are concerned with the moderating effects engagement, even in the absence of age-related effect on word learning, we can see if the trend between age and word learning performance is different also considering the child's level of engagement. We used aggregated measures of engagement that characterize the child's overall observational and holistic engagement. In other words, does the relationship between age and word learning performance change when children have higher engagement during the session?

Research Aim 2 The second research goal was to analyze all of our individual measurements of engagement, putting method of data collection aside, and investigate which collectively show the most common variance when predicting word learning outcomes in this context. An exploratory factor analysis (EFA) identifies similar trends among groups of engagement metrics and describes underlying latent constructs of engagement (Suhr, 2005). Thus, the second analysis had two parts. First, engagement measures were analyzed via EFA which help explain which measures account for the variance among the data. Second, these latent constructs were added to a linear regression to see if they stronger moderators of word learning than the model in Aim 1. We hypothesized that these latent constructs would better predict learning outcomes compared to models from Aim 1, which use engagement measures aggregated by method of data collection.

Methodology

2.1 Participants

Data collection for this study is ongoing, this thesis presents the data collected and coded by August 2021. Children 20- to 48-months old were recruited for this study, though our current sample with analyzable data ranged from 25- to 46-months. Participants were recruited through the McGill Childcare Center; the Center for Research on Brain, Learning, and Music (CRBLM) network; International Laboratory for Brain, Music, and Sound Research (BRAMS) network; and remotely through Children Helping Science and via Facebook. Additionally, children were included in our study if they heard English or French at least one-third of the time, and if they had a computer with a working webcam and stable internet connection for video conferencing. Ethics approval for this remote research study was obtained from the McGill Faculty of Medicine IRB. Prior to participation, caregivers completed an informed consent form on behalf of themselves and their child, through moderated consent process with a researcher on a video call.

41 children were recruited for this thesis, 4 of whom did not meet all inclusion criteria. Of the 37 children enrolled in the study, 10 children were dropped: 1 child due to experimenter's

technology error (session not recorded), 6 due to participant's technology error (i.e. inaudible for portions of test, video recording uncodable, spotty internet connection), and 3 children due to fussiness. This resulted in 27 children who completed the study, detailed in Table 2.1.

Table 2.1 Sample characteristics

	Children completing study procedures	*Children who passed inclusion criteria for each dependent variable*		
	Current Sample	**Referent Selection**	**Gaze**	**Label Production**
n	27	21	26	5
Age[a]	37m15 (7m04) 21m09–46m29	38m13 (6m06) 25m11–46m29	37m09 (7m06) 21m09–46m29	38m17 (8m22) 29m14–43m18
Sex	M: 16 F: 11	M: 13 F: 8	M: 16 F: 10	M: 4 F: 1
Testing language	EN: 22 FR: 5	EN: 16 FR: 5	EN: 21 FR: 5	EN: 5 FR: 0
Teaching condition	Chair: 8 Other: 8 Participant: 11	Chair: 8 Other: 6 Participant: 7	Chair: 8 Other: 7 Participant: 11	Chair: 2 Other: 1 Participant: 2
Mean words produced (MCDI)[a,b]	72 (19) 32-100	70 (20) 32-99	71 (19) 32-99	70 (14) 53-86

[a] Values reported as $M(SD)$ and ranges are $min - max$. Age reported as ${}_{\text{months}}\text{M}_{\text{days}}$.

[b] Expressive vocabulary scores from MCDI matching the language the child was tested in.

Children with known hearing loss, developmental delays, or language problems (determined through parent report) were not actively recruited to participate. After intake, 6 parents reported that their child had either a diagnosed (1) or undiagnosed (5) speech delay. These children were included for participant in the study nonetheless, as none were reported having processing delays and all children could still participate with the test.

Information on age, sex, and expressive vocabulary in the language of testing were collected through a survey before participating in the study. Children were on average 2.5 to 3.5 years old (M = 37m09, SD = 7m06). There were 10 females and 16 males in the study. Language of testing was English (n = 21) or French (n = 5). For English-French bilingual children ($n = 23$), the child's strongest verbal language, as indicated by the parent, determined whether we tested in English or French. Parents completed the short form MacArthur-Bates Communicative Development Inventory (MCDI; Fenson et al., 2000). For bilingual children, parents completed the MCDI in both English and French, but only results from the MCDI that matched the language the child was tested in are reported here. Participants were also randomly assigned to a teaching condition (described in Section 2.3).[1]

To be included in analysis of our dependent variables (referent selection and gaze), participants had to show the ability to meet task demands (e.g., gaze to or indicate the object tested) for at least one out of four trials with known filler objects. 26 out of 27 met this criterion at a greater than chance rate for the gaze measure, while only 21 out of 27 met this minimum criteria for the referent selection measure. Only 5 out of 27 children met inclusion criteria for label production trials (for details, see Appendix D), so we will exclude this data from our analyses. Note, 20 out of 27 children produced a codable answer for at least one filler trial, but only 5 children produced a codable answer for at least one target trial, so we do not have a large enough sample to look at word learning outcomes via label production. For this thesis, we are limited to the data collected for the 21 children who met the minimum criteria for the referent selection measure. We recognize we do not have adequate statistical power to investigate the moderating effect of engagement through

[1] Age distribution between conditions were slightly uneven. While condition assignment was randomly distributed, the final age distribution between conditions was slightly uneven. Due to small sample in each condition and this uneven age distribution, the comparison of performance between conditions was not considered in our primary analyses.

interaction terms in a regression model at this time; therefore interpretation will focus on trends in our current sample. Table 2.1 shows full demographic information for our sample overall and for participants included in our analysis of referent selection and gaze.

2.2 Design

This study was centered around how children engage during a word learning task. A teacher presents a novel object to the child and provides a novel name for the object. During this lesson, the teacher presented four novel objects, two which were given novel names (for teaching script, see Table B.2). Later, word learning is tested through two different tests: label production (child produces the name when shown a picture of a novel object) and referent identification (child identifies a novel object from a set of objects). Data was collected in the context of a larger study that has considerably more complex between-subjects experimental design where participants are randomly assigned to one of three conditions that manipulate social contingency and directed social cues (described below). For my purposes, participants from all three conditions will be included in the same moderator analysis.[2]

2.3 Experimental procedures

The experiment has two distinct phases, teaching and testing, and is shown below in Figure 2.1. First the participant was presented with the teaching phase, which lasted approximately 3 minutes. There were three conditions in the larger study (participant-directed, other-directed, and chair-directed). Depending on the condition, the participant saw a video

[2] While it is expected that children may have different levels of engagement in each condition, the primary goal is to see if, in general, engagement improves the overall trend that older children perform better word learning tasks.

recording or a live video chat of the teacher presenting the four novel objects. The teacher script was similar in each condition (different warm up for other-directed, shown in Appendix B) and has been designed to be as similar as possible for both English and French versions.

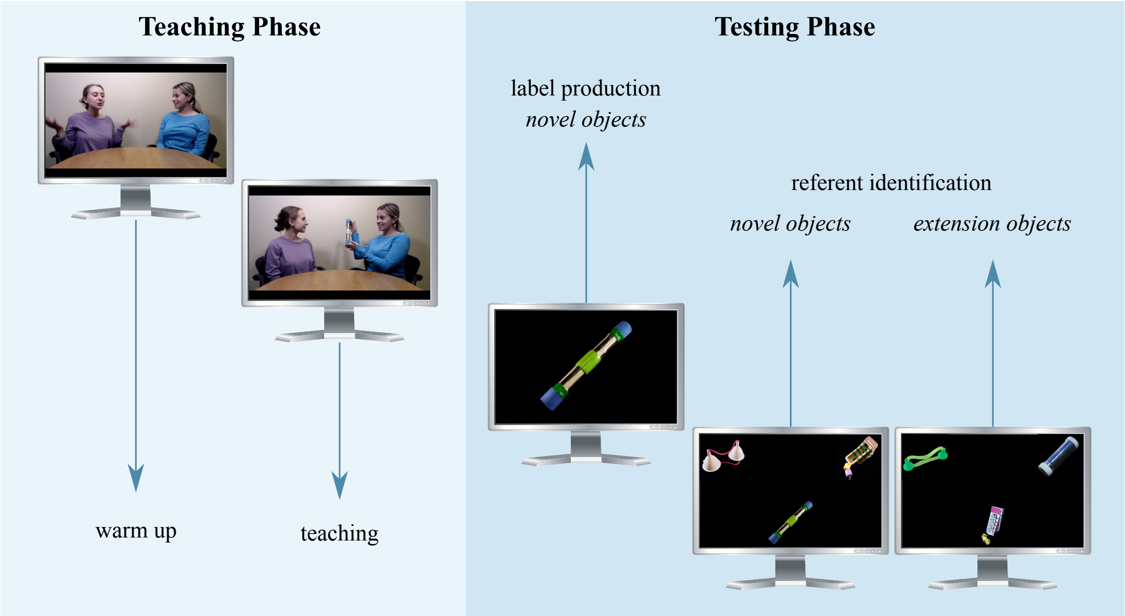

Fig. 2.1 Experiment procedure

Teaching phase The teaching phase had two distinct subphases: warm up and teaching. During the warm up subphase the teacher introduces herself and makes 7-10 bids for the child (or other, as outlined in Table B.1) to respond to. After each bid, the teacher waited approximately 2 seconds for the child to respond, before continuing the script. The teacher then transitioned to the teaching phase by saying, "Now I'm going to show you how to use some new objects." During the teaching subphase the teacher sequentially showed each of the four novel objects, prompting the child to look at each object four times. Two of the novel objects were given labels, *fopam* and *mimole*. The novel label was repeated four times for each object (for details see Table B.2). Engagement was coded separately for each novel object for the full duration that the object is being presented. The teacher concluded the session with “I’ll see you later. Bye!”

Testing phase The testing phase had three subphases that evaluated different aspects of word learning: target label production, target referent identification, and extension object identification. The target objects were the same objects used in the teaching phase. With these objects we first tested the child’s ability to produce the novel label of the target object (label production) and then tested if the child could identify the novel object when given the name from a set of three novel objects (referent identification). We then repeated the same referent identification test with a set of extension objects to see if the child could generalize the label to an object that was similar but not identical.

For the label production trials, the child was first shown a single image of an object taking up the majority of the screen, which was visible for the full duration of the trial, as shown in Figure 2.1. After 750 ms, the child was prompted to say the name of the object (i.e., “What’s this called?”). To compare label production for the novel objects to known objects, we interspersed trials with filler objects throughout the task. We always began testing with a filler trial so that the child can warm up for the test using known objects. There were seven label production trials (1 instruction, 4 novel, 2 filler), ordered as [instruction, filler, novel, novel, filler, novel, novel]. Since only the fopam and mimole were labeled, each target object was tested twice for label production. The child had 7 seconds to produce the novel label for each label production trial.

For the referent identification trials, the child was shown images of three objects on the screen, simultaneously displayed in the top left, top right, and bottom center of the screen as shown in Figure 2.1. The objects displayed on the screen were a selection of 3 of the novel objects (*mimole*, *fopam*, and either *distractor A*/*distractor B*) or 3 of the filler objects (*car*, *bottle*, and *fork*). After 1000 ms, the child was prompted to point to one of the objects (i.e., "Mimole, point to the mimole."). The onset audio delay was longer for referent identification trials to allow the child to scan all three objects on the screen before hearing the prompt. There were 14 referent identification trials (1 instruction, 6 novel, 3 filler, 4 extension), ordered as [instruction, filler, novel, novel, novel, filler, novel, novel, novel, filler, extension, extension, extension, extension]. The extension objects were designed to have the same features as the target and distractor objects but vary in color and size. It is important to note that during the novel object trials, both target objects were always displayed on the screen. Again, both the fopam and mimole were tested twice for referent identification. The child had 10 seconds to find the novel object for each referent identification trial.

After the test trials were complete, the parent completed a brief engagement survey immediately after the child was finished testing. This survey is how we collected the parent/caregiver perspective on the holistic measures of engagement. The experimenter also completed the same engagement survey from the experimenter's perspective and could not see the parent's responses to the survey.

2.4 Experimental objects

Four novel objects (two target objects with labels and two distractor objects without labels) as well as four filler objects were used for testing. The novel objects were created in our lab with a prior norming study, and have been used in related studies (Bang, 2017). Images of these objects are shown in Figure 2.2, which were the same images presented on-screen during the testing phase. Below we describe the objects and the specific action associated with each novel object.

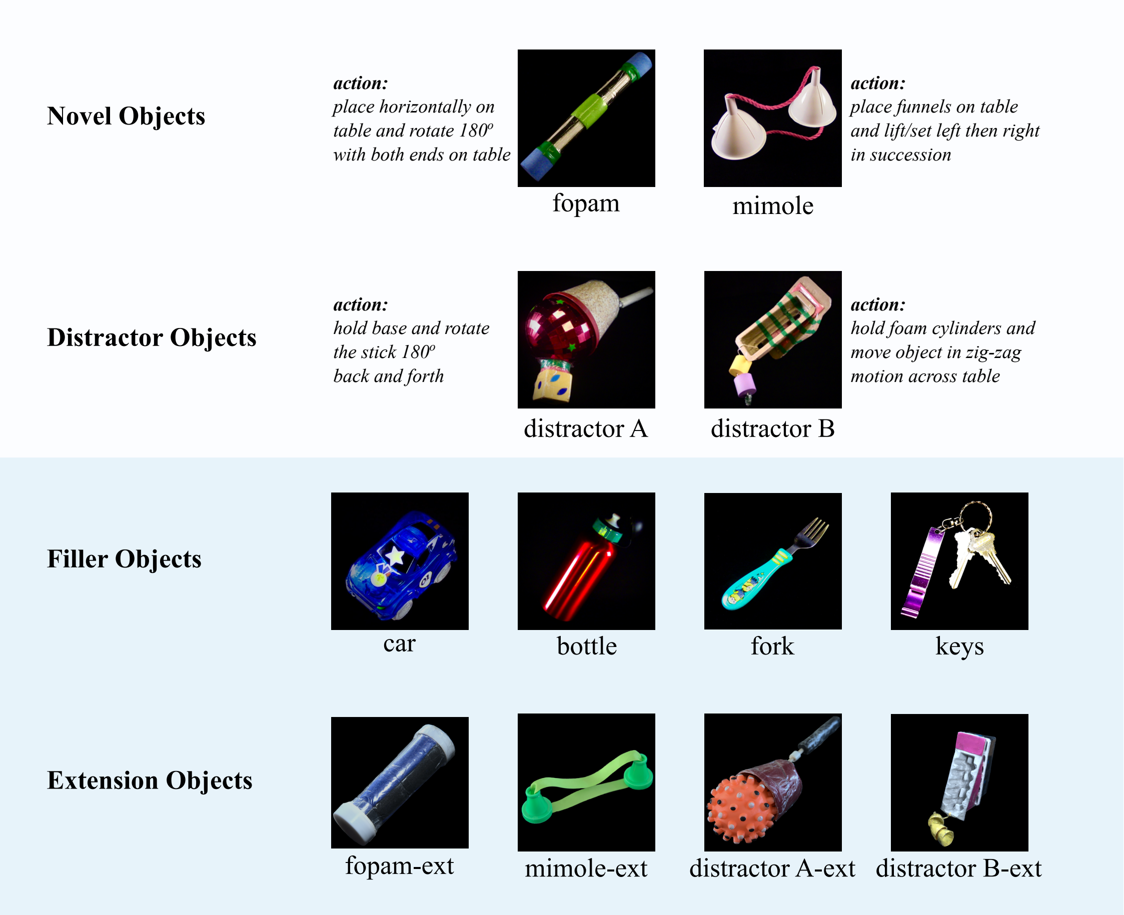
Novel Objects
action:
place horizontally on table and rotate 180º with both ends on table
fopam
mimole
action:
place funnels on table and lift/set left then right in succession
Distractor Objects
action:
hold base and rotate the stick 180º back and forth
distractor A
distractor B
action:
hold foam cylinders and move object in zig-zag motion across table
Filler Objects
car
bottle
fork
keys
Extension Objects
fopam-ext
mimole-ext
distractor A-ext
distractor B-ext

Fig. 2.2 Experimental objects

Target objects The *fopam* was a cylindrical tube with three prominent features (a green foam center and two blue foam ends). The cylinder had white and black stripes running along its length and had screws inside to add weight and made noise when moved. The novel action was to place the fopam horizontally on the table (parallel to the child) and rotate it 180° while both ends remained on the table, removing and twisting your hand back after each rotation.

The *mimole* had two white funnel-like features which were held together with a red rope through the funnel centers. The novel action for the mimole was to grab both funnels with your fingertips and then lift each funnel up and then return to the tabletop (first left then right shortly after, in succession), which created a distinct sound when hitting the table.

Distractor objects Distractor A had a large globe (with a cube affixed to the surface) connected through a base to a stick that will allow it to rotate. The novel action associated with distractor A required the person to hold the base and rotate the stick 180°.

Distractor B had two wedge-like pieces, and a rope with two foam cylinders attached to the short end of the wedges. The novel action associated with distractor B required the person to hold the foam cylinders and move the object in a zig-zag motion.

Filler objects We had four known objects, which were referred to as the filler objects. The four filler objects were a blue toy car, a fork, a red water bottle, and a white hat. We never specified a specific action associated with the filler objects at any point during the study and were only shown the images during the testing phase.

Extension objects Additionally, we had one set of extension objects which are variations of the mimole, fopam, distractor A, and distractor B. We only showed these objects to the child during referent identification as an additional test to see if they could extend the properties of a mimole or fopam onto a new object that looks similar but was not identical.

2.5 Remote delivery of study

Due to the ongoing COVID-19 pandemic and McGill's prioritization of remote human participant research, this study was conducted remotely in an online format. The consent form and all demographic questionnaires were adapted for LimeSurvey (version 3, LimeSurvey Project Team & Schmitz, 2012), according to the guidelines of the McGill

Cloud Directive. The demographics information and language questionnaires collected in the larger project are outlined in Appendix C. In the current study, only the child's age were included.

WebEx (version 41.5.12; Cisco Systems, 2021), a video-conferencing platform approved by the McGill Cloud Directive for remote research, was used to interact and record the participants during the study. WebEx was also used to conduct the live participant-directed condition.

Pre-recorded video for other-directed and chair-directed teaching sessions were recorded using a Sony HDR-CX900, recorded at high quality resolution of 1440 x 1080/60i. All videos were recorded in the same room with consistent backdrops and lighting.

Pre-recorded videos and the testing stimuli presentation were shown to the participant through jsPsych (version 6.1.0, de Leeuw, 2015). This software standardized the display and timing of all videos and stimuli, which was displayed remotely on a computer in the family's home, via a web browser (typically Chrome or Firefox).

Recruitment links, access, and test data was managed through a server-based version of JATOS (version 3.6.1, Lange, Kühn, & Filevich, 2015) hosted in Montreal, Quebec on the BRAMS-OTP servers.

2.6 Data coding methods

There were 5 separate data coding methods to process each child's session data. Four of these methods used data collected from the video recordings was processed using BORIS, an open-source event-logging software (version 7.10.2, Friard & Gamba, 2016). These methods were for precoding, observational engagement, referent selection, gaze, and label production. Each method had a team of dedicated coders (research project students and undergraduate research assistants) to complete data coding. For each method (besides precoding) videos were double-coded, and interrater reliability was calculated using Cohen's kappa (κ). Values of κ can be interpreted as follows: ≤ 0 indicates no agreement, $0.01 - 0.2$ indicates none to slight agreement, $0.1 - 0.40$ indicates fair agreement, $0.41 - 0.60$ indicates moderate agreement, $0.61 - 0.80$ indicates substantial agreement, and $0.81 - 1.00$ indicates as almost perfect agreement (McHugh, 2012). The fifth method was collected from the engagement surveys for holistic engagement. These surveys were collected using jsPsych (version 6.1.0, de Leeuw, 2015) and extracted and processed using R (R Core Team, 2020). Below each data coding method is explained in detail.

2.6.1 Precoding

Before coding engagement, label production, gaze, and referent selection, we standardized videos through precoding. Video recordings were flipped horizontally to ensure when the child points to the left of their screen, the video also reflects them pointing left; which was occasionally reversed if the participant's webcam was inverted. The correct direction was determined from the calibration and post-calibration phases of the experiment where we asked the child to point to a character in each region of interest (top left, bottom center, and top right of the screen). The primary goal of the precoding methods was to find start/end timestamps for engagement bids during teaching and trial boundaries during testing to align the child's responses with what was being said or presented. The start of the engagement bids was found through waveform of the audio and ended with the start of the next engagement bid. Each test trial was separated into the pre-prompt (which started with a fixation and noise, via the waveform) and the prompt (the start of the trial prompt, via the waveform) which ended with the start of the following trial. While engagement bid scripts were standardized and trial times were consistent for all participants, we manually assigned the pre-coding timestamps for consistency across conditions, teachers, and the variety of hardware and network speeds on the participant's computer for two primary reasons. First, all engagement bids for the participant-directed condition needed to be assigned manually. During the live participant-directed condition, the teacher's responses were contingent with participant's responses, meaning conversational pauses varied between engagement bids as the teacher followed the script (see Appendix B); thus, manual assignment made pre-coding consistent for all teaching conditions. Second, we tried automatically assigning these timestamps based on our standardized trial durations and found manual assignment to be more efficient and accurate for the test trials. Occasionally, participant's internet connection loaded images and audio slowly (upwards of 50 msec), and caused a short delay before the start of the trial, before the images and prompt were displayed. Thus, manually assigning timestamps for test trials standardized procedures across all participants. The timestamps recorded during precoding were used to group measurements for observational engagement, label production, gaze, and referent selection coding.

2.6.2 Engagement

Similar to the aspects of engagement highlighted in current models of language task engagement (Figure 1.1, Egbert et al., 2021) and the engagement measures used in other word learning studies, we created the engagement coding scheme listed in Table 2.2. This scheme measures engagement from two approaches: behaviors we can observe and holistic assessments. Observational measures of engagement were grouped into verbal measures and gestural measures. Holistic measures of engagement were collected separately from two perspectives: from the parent and from the experimenter. Below we review how observational and holistic engagement measures were coded and analyzed.

Table 2.2 Measurements of observational and holistic engagement

Observational engagement measures[a]	**Holistic engagement measures**
Verbal measures	*Survey responses*
(i) immediate repetition	(i) interest in the activity
(ii) direct verbal response to on-screen question[b]	(ii) responsiveness to prompts
(iii) verbal comment to teacher or activity	(iii) perceived engagement
(iv) verbal comment to parent	(iv) focus during the task
(v) verbal comment directed at instrument/object	(v) emotional state
(vi) other vocalization	(vi) social engagement
Gestural measures	
(i) direct gestural response to on-screen question	
(ii) point, reach, or other hand movement[c]	
(iii) communicative/conventional gesture	
(iv) change in facial expression	
(v) shift in body position	
Collected via video recording from experimenter	Collected via survey from parent/caregiver and experimenter

[a] Observational measures were assigned in the hierarchy listed, i.e., if the child said, "I like this game." since it was not an immediate repetition, and it was not a direct response to an engagement bid or prompt, but it was related to the activity, it would be categorized as a *verbal comment to teacher or activity*

[b] Direct verbal responses overlap with the measures we used to show word learning during label production, so direct verbal responses were excluded for label production trials.

[c] Pointing gestures overlap with the measures we used to show word learning during referent identification, so pointing gestures were excluded for referent identification trials.

2.6.2.1 Observational measures of engagement

Commonly used measures of toddler engagement formed the basis of the system for coding child engagement in this study (McClure et al., 2020; Striano & Stahl, 2005), yielding a list of countable behaviors as shown in Table 2.2. Using BORIS software for event coding, the frequency of these behaviors was counted in discrete time intervals delimited by adult engagement bids (i.e., prompts from the teacher as shown in Appendix B). Child engagement was represented as number of verbal and nonverbal behaviors per adult bid in each phase of the experiment (teaching or testing, shown in Figure 2.3).

A team of two coders conducted the coding for observational engagement. Coders were

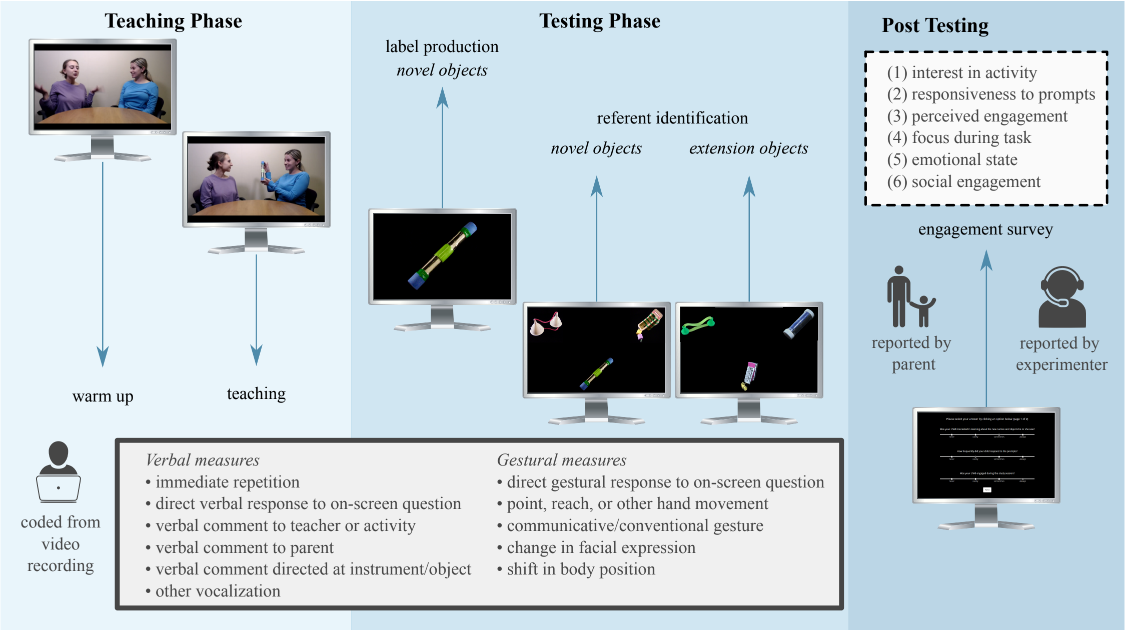

Fig. 2.3 Engagement measures collected in all phases of the experiment. Observational measures (solid box) coded from video recording of participant. Holistic measures (dotted box) will be collected separately from the parent and from the experimenter perspectives

blind to the information that was shown on-screen to the child, and (where possible) were not involved with the data collection of the child. Video recordings were played back at half speed, with both video and audio enabled. At half speed, video was slow enough to capture gestures and audio was fast enough to comprehend verbalizations. Coders were able to view each video one time to assign codes for both verbal measures and gestural measures. When a child produced a verbalization and gestured at the same time, coders were able to jump back 1 second to ensure both observations were recorded properly. In these cases, the ordering did not matter, as observation counts were grouped into bids/trials.

We double coded 43% percent of the videos. Interrater reliability between coders was established for most observational engagement codes used (substantial agreement, between $\kappa = 0.60$ and $\kappa = 0.73$, for most codes; 95% agreement overall). Two measures, change in facial expression and body position, did not have high enough reliability between coders ($\kappa = 0.19$ and $\kappa = 0.29$ respectively) and therefore were excluded in our analysis. For all double coded videos, coders reviewed the video and came to agreement on all codes that did not agree.

Measures of observational engagement were totaled for the teaching and testing phases separately. For the primary analysis we only used the sum of observational engagement during the teaching phase. This, in part, was to reduce the number of variables in the regression analysis; however, this also follows how engagement is used in other novel word learning studies. For the exploratory factor analysis, the totaled scores for the teaching phase will be used, and the individual observational measures from each testing trial will be used and examined. Verbal and gestural measures will be analyzed separately for the exploratory factor analysis.

Verbal measures of engagement The verbal measures of engagement aim to capture how the child is responding to the activity through the verbalization and reactionary utterances, like “ooh” or “ahh”. Namely, we aimed to capture the nature of the utterance and where the child directed the utterance. During the warm up and teaching phases children were asked specific questions like “What’s your name?” or “Isn’t this one fun?” which were designed to elicit verbal responses. During the label production trials children were asked to produce the name of the object on the screen. To standardize how responses were coded in these different phased we implemented a hierarchy of verbal measures of engagement (also listed in Table 2.2): (i) immediate repetition (ii) direct verbal response to on-screen question (iii) verbal comment to teacher or activity (iv) verbal comment to parent (v) verbal comment directed at instrument/object (vi) other vocalization.

We were most interested in direct verbal responses to on-screen questions. There were on-screen questions that elicited verbal responses during both teaching and testing. During testing, many children simply repeated what was last said; for example, when asked, “Fopam. Point to the fopam.” the child would respond, “Uhh, fopam.” Therefore, in practice, it was more consistent to categorize this type of response as an immediate repetition instead of a direct verbal response, since the child’s utterance did not indicate a novel production. This is why immediate repetition is at the top hierarchy. Direct verbal responses were novel utterances that answered the question asked on screen, such as “What’s your favorite color?” ⟶ “Blue.” during teaching or “What’s this called?” ⟶ “I don’t know.” during testing. The correctness of the response did not matter for verbal engagement coding, but the child’s response needed to be relevant to the on-screen question.

We were also interested in verbal comments, where unprompted, the child made a verbalization. We separated verbal comments into three codes: directed at the teacher or activity, at the parent, or at another instrument/object unrelated to the study. Common comments directed at the teacher or activity took the form, “This one.” or “I wanna play another game.” Common comments directed at the parent were “I’m all done.” and “I like this one!” While there is some subjectivity between these codes for some responses (for example, “Is that the fopam?”) if the child was attending or gesturing towards the screen it was considered towards the teacher or activity, and if the child shifted their body towards their parent it was considered directed towards the parent. Comments directed at an instrument/object then followed as other verbal comments unrelated to the activity and not directed towards the parent, such as, “I’m hungry.” Finally, all other non-verbal utterances, like “ooh” or a laugh were coded as other vocalizations.

It should be noted that compound utterances, such as, multiple rapid repetitions (“fopam, fopam”) or a vocalization and direct response (“Hmmm, car. It’s a car!”) were counted as a single verbal response, and were coded with the highest relevant verbal measure.

For consecutive codes (multiple repetitions or two direct responses, “I don’t know...[*pause*] Mimole!”) there needed to be either a long pause between the utterances or significantly different content in the response for it to be counted as two separate verbal responses.

There was one exception when combining verbal measures of engagement together for the overall sum of observational engagement. We excluded direct verbal responses to on-screen questions for the label production trials. This is because the dependent variable used to analyze label production are the child’s verbal responses to the label production questions. As to not confound verbal engagement with the label production measure, the direct verbal responses to on-screen questions for label production trials were excluded from each child’s total count.

Gestural measures of engagement The gestural measures of engagement aim to capture how the child responds to the activity through the non-verbal cues. We attempted to capture gestures that reflected the cognitive and emotional state of the child. During the warm up and teaching phase, some engagement bids asked the child to respond with a gesture, “Can you point to your nose, like me?” Similarly, during the referent identification trials, children were asked to point to an object on screen. To standardize how responses were coded in these different phases, we implemented a hierarchy of gestural measures of engagement (also listed in Table 2.2): (i) direct gestural response to on-screen question (ii) point, reach, or other hand movement (iii) communicative/conventional gesture (iv) change in facial expression (v) shift in body position.

We were most interested in direct gestural responses to on-screen questions and pointing, reaching, and other hand movements. Direct gestural responses were primarily for engagement bids during warm up, such as, “Can you touch your ears like this?” ⟶ [*child touches ears*]. Point, reaching, and other hand movements were grouped together for two reasons. First, it was not important for us to try and distinguish a point versus a reach; both were hand gestures indicating something in a specified direction. Second, for the referent selection coding we did not distinguish pointing from reaching, and we wanted to reflect similar information across all phases of the study. Thus pointing, reaching, or other hand movements that directed focus toward a person/location were coded together. This included the direct responses during the referent identification trials, i.e., "Mimole. Show me the mimole." ⟶ [*child points to any object on-screen*]. Again, coders did not have access to what information was presented on the child's screen at the time, so correctness of this pointing gesture did not matter. To standardize when the pointing gesture temporally occurred, we coded when the pointing gesture landed on a position. If a child slowly raised their hand, then extended their hand and finger, we coded the end of the observation when the gesture stabilized.

We were also interested in other gestures that reflected the child's cognitive and emotional state. Communicative/conventional gestures included gestural responses that had common intentions to communicate, such as waving hello, clapping, and nodding or shaking the head in agreement/disagreement. This commonly occurred during the Teaching phase when the child was prompted, "Isn't this one cool?" ⟶ [*child nods*], or at the beginning/end of warm up when the child waved hello or goodbye. We attempted to capture changes in facial expression, like raising eyebrows or pursing lips. And finally attempted to capture changes in body position, like shifting in the chair or turning the head sideways.[3]

Procedures on how to code consecutive gestures were similar to the verbal engagement

[3] Changes in body position are not typically categorized as gestures. However, in our remotely acquired webcam data, children frequently moved away from the computer screen or shifted their body when they were either very excited (getting very close to the screen to respond) or when they were disengaged (and moved away from the activity). While these responses are not strictly gestural, they made up a significant proportion of children's observable engagement behaviors (see Table 3.2 for details), and therefore included in this category.

measures. If a child pointed to all three objects, and paused and hovered towards each region of interest, we coded this as three separate gestures. If a child pointed to all three objects, but waved their hand a circle/triangle quickly, we coded this as one gesture since it was one continuous motion.

After coding and calculating interrater reliability on the gestural engagement measures, we found that changes in facial expressions ($\kappa = 0.19$) and changes in body position ($\kappa = 0.29$) had low agreement between coders. These behaviors were difficult to count and typically did not match between coders because of the number of observations recorded (i.e., 2 versus 3 changes in facial expressions). While Cohen's kappa is not sensitive to these cases, as it measures rating agreement between coders, we decided to exclude these gestural measures from the sum of observational engagement. It is possible that such holistic coding of expression change is not reliable, as emotion research employs coding these behaviors with very specific physiological definitions and movement of specific muscular groups (Ekman & Rosenberg, 2012).

2.6.2.2 Holistic measures of engagement

The holistic measures of were collected via a brief 6-question survey at the end of the experiment (detailed in Table A.1). The survey aims to capture holistic ratings on six different components of engagement measures (also listed in Table 2.2): (i) interest in the activity (ii) responsiveness to prompts (iii) perceived engagement (iv) focus during the task (v) emotional state (vi) social engagement

Each of these measures were recorded on a 4-point Likert scale (never, rarely, sometimes, always) focused on the frequency of the child's overall engagement during the session. The 4-point Likert scale allows, at a minimum, to split the results into two nominal categories such as high and low levels of engagement. This enabled us to consider children as more or less engaged for each holistic measure. The responses were recoded on a numeric scale as scores from $0 - 3$. An overall score of holistic engagement were calculated by summing all six survey responses. The question pertaining to the child's focus during the task (see Table A.1), was phrased with reversed coding, so the scores for this question were inverted, so a higher score indicates more focus and a lower score indicates more distraction.

Holistic measures were reported from two perspectives: the parent/caregiver and the experimenter running the session. For the primary analysis, the overall holistic engagement score were totaled separately from each perspective, into a single score for holistic parent rating and holistic experimenter rating. For the exploratory factor analysis, we explore how survey responses from both perspectives correlate with other variables independently.

2.6.3 Word learning performance

We had three ways of knowing if a child learned the novel word presented, via referent selection, gaze, and label production. These are common measures used (but analyzed independently) in word learning studies (Kirkorian, Choi, & Pempek, 2016; Roseberry et al., 2014), which we adapted for our remote study. Referent selection and gaze relate to the referent identification trials, where the child must point to the object tested when shown a set of three objects on the screen. Since we had a large age range in our final sample (see Table 2.1), the children may have different developmental abilities when it comes to processing and responding to the referent identification task. Therefore, we also analyze the child's gaze during this task. We have not included the label production task in this analysis due to lack of data (only 5 out of 27 children produced target labels during testing), but have outlined our procedures and analysis in Appendix D.

2.6.3.1 Referent selection

Referent selection coding was coded from the participant video recordings. We were primarily interested in any gesture or indication the child made that was directed towards one of the three objects on screen: (i) top left (ii) bottom center (iii) top right (iv) other. Ultimately, we wanted to know whether the child indicated the correct object to most often. Every gesture was coded for the region of interest the child selected with their gesture, as well as a subcode of how they indicated their selection, which were (a) extension of arm (b) body movement. For purposes of our analysis we did not consider the difference between these subcodes. Overall, there were very few body gestures recorded and nearly all were coded as arm extension ($M_{\text{arm}} = 15.38, \sigma_{\text{arm}} = 7.82, M_{\text{body}} = 0.14, \sigma_{\text{body}} = 0.36$). This decision is further supported by existing literature that suggests index-finger pointing has been identified as the most important pre-linguistic pointing behavior (theoretically by Butterworth, 2003 and empirically by Colonnesi, Stams, Koster, & Noom, 2010).

Regions of interest were coded when the gesture landed, similar to how pointing gestures were coded for gestural measures of engagement (see Section 2.6.2.1). This referent selection coding scheme follows establish methods from similar studies (Allen & Scofield, 2010; Matthews, Behne, Lieven, & Tomasello, 2012; McGillion et al., 2017). It is, however, one of the few completely online word learning studies to implement referent selection coding from a video recorded via webcam.

A team of three coders conducted the coding for referent selection. Coders were blind to the information that was shown on the screen. Additionally, videos were coded without sound, so all gestures were coded solely based off of visual cues.

We double coded 50% of the videos, randomly assigning the coders for each video (5 by $\text{coder}_1 \cap \text{coder}_2$, 4 by $\text{coder}_1 \cap \text{coder}_3$, 5 by $\text{coder}_2 \cap \text{coder}_3$). Interrater reliability was established for referent selection coding ($\kappa_{\text{TL}} = 0.73, \kappa_{\text{BC}} = 0.71, \kappa_{\text{TR}} = 0.84, \kappa_{\text{other}} = 0.66$, 95.5% agreement overall). Because agreement was high, and 81.4% of bids had only 0 or 1 codes assigned, referent selection data was not coded to consensus. Instead, codes from coder$_1$ and coder$_2$ were used by default for all double coded videos.

The relative proportion of referent selection was calculated for each region of interest for each trial. We did not include the other category in these calculations, since we specifically wanted to know out of the selections of the three objects shown on the screen, which did the child point to most often. This was paired with data on which region of interest the correct object was located for each trial, to calculate the main value used to indicate word learning via referent selection (Equation 2.1). This measure, relative proportion of referent selection to word tested, can range from $0.0 - 1.0$. A value of 0.33 would be the equivalent of random chance to select the correct object.

$$\text{proportion referent selection of target} = \frac{n_{\text{correct ROI}}}{n_{\text{TL}} + n_{\text{BC}} + n_{\text{TR}}} \tag{2.1}$$

2.6.3.2 Gaze

We were planning on following established methods for gaze coding, using an eye tracking device (ASL EYE-TRAC 7) to capture gaze data (both visual attention and arousal). However, the COVID-19 pandemic required the study to be run completely online and the established methods using the eye tracker were no longer possible. Instead, gaze coding was coded from the participant video recordings. We were primarily interested in the direction of the child's gaze, particularly towards the three objects on screen: (i) top left (ii) bottom center (iii) top right (iv) other (v) uncodable. Ultimately, we wanted to know whether the child gazed towards the correct object to most often. This direction was coded visually from the position of the child's eyes. Participants were required to be seated in a well lit room close enough to the webcam so that the whites of their eyes could clearly be seen. This standardized how we assigned direction of the gaze. Direction was coded at the onset of a gaze shift. Occasionally, children shifted such that part of their body was off the screen and their eyes were not clearly visible in the recording. During these times, coders were instructed to mark the interval as uncodable.

A team of three coders conducted the coding for gaze. Coders were blind to the information that was shown on the screen. Gaze coding was done without sound with frame-by-frame playback, which was the smallest temporal resolution available in the recordings (typically 60 fps). Gaze shifts are on a much smaller timescale than the engagement and referent selection gestures, so observations were then binned into 33 ms bins, to compare the temporal reliability of our gaze coding, which is a standard process in other studies (Fernald, Perfors, & Marchman, 2006; Loi, Marchman, Fernald, & Feldman, 2017; Marchman et al., 2019; Pel, Manders, & van der Steen, 2010). We double coded 54% of the videos, randomly assigning the coders for each video (4 by $\text{coder}_1 \cap \text{coder}_2$, 5 by $\text{coder}_1 \cap \text{coder}_3$, 6 by $\text{coder}_2 \cap \text{coder}_3$). Temporal reliability was high, with 83.5% frame agreement and 84.9% agreement within one frame, as was interrater reliability ($\kappa = 0.78$).

Similar to referent selection, the relative proportion of gaze towards each object (i.e., $t_{\text{TL}}/(t_{\text{TL}} + t_{\text{BC}} + t_{\text{TR}})$) was calculated for each region of interest for each trial. We did not include the other category in these calculations, since we specifically wanted to know out of the selections of the three objects shown on the screen, which did the child look at longest. Since gaze is a continuous measure, minor difference between the gaze measure and the referent selection measure is that gaze describes the proportion of time relative to the time spent staring at the screen. This normalizes the measure for all trials across all participants. The relative proportion of gaze was paired with data on which region of interest the correct object was located for each trial, to calculate the main value used to indicate word learning via gaze (Equation 2.2). This measure, relative proportion of time gazing towards word tested, can range from $0.0 - 1.0$. A value of 0.33 would be the equivalent of random chance for this measure.

$$\text{proportion gaze to target} = \frac{t_{\text{correct ROI}}}{t_{\text{TL}} + t_{\text{BC}} + t_{\text{TR}}} \tag{2.2}$$

Results

3.1 Descriptive statistics, trends, and correlations

All statistical analyses were carried out using R (version 4.0.4, R Core Team, 2020). First, we analyzed the primary relationship, the effect of age on our word learning measures, which were proportion referent selection of target and proportion gaze to target. We also analyzed the correlations between the observational engagement measures and the holistic engagement measures.

The trends and correlations between observational engagement measures and the holistic engagement measures shown below will motivate the regression analysis and exploratory factor analysis to address our two research aims: (1) Does engagement moderate word learning outcomes? and (2) Which engagement measures best explain the variance in word learning outcomes?

Overall, the means reported in Table 3.1 show on average children were slightly above chance (0.33) for both referent selection and word learning gaze outcomes. For observational engagement measures, the number of gestural measures were very low for referent identification trials. This matches our expectations, since we removed the point, reach, and hand movement measure (since it overlaps with how we measured proportion referent selection of target) as well as the two gestural codes with low reliability. Thus, gestural measures were only counting direct gestural responses to on-screen questions and communicative/conventional gestures. There were only three participants who made communicative/conventional gestures; so, gestural measures are primarily counting direct gestural responses to on-screen questions.

Table 3.1 Descriptive statistics on word learning outcomes

	n	M	SD	min	max
Referent selection					
proportion referent selection of correct object	21	0.37	0.17	0.11	0.75
number of gestures or indications	21	11.48	6.31	6.00	29.00
Gaze					
proportion gaze to correct object	26	0.35	0.06	0.27	0.55

3.1.1 Referent Selection

We expected older children to perform better when prompted to point or show, indicating the region of the screen where the filler or target object appeared. Yet, the relationship between age and referent selection, shown in Figure 3.1, did not completely follow our hypothesis. Over all trials, participants performed only slightly above chance ($M = 0.37$). As we expected, there was a clear difference between performance on filler objects compared to novel objects. All participants performed at or above chance on the filler trials ($M \approx 0.7$) meaning children selected the correct object well above chance. These are used to ensure the child can complete the task, as filler objects were familiar objects children would know the name of. However, for the target trials (the novel objects) taught during the teaching phase of this study, children were far less successful ($M \approx 0.27$). Seven children, spread across all ages, had an average of 0.0 for all novel object trials. This type of behavior could happen because the child simply did not attempt to select any objects or alternatively because they selected a single incorrect object, and therefore never selected the word tested. There is a positive relationship between age and relative proportion of referent selection, suggesting that older children had higher word learning performance. Yet, nearly half of the children tested performed below chance.

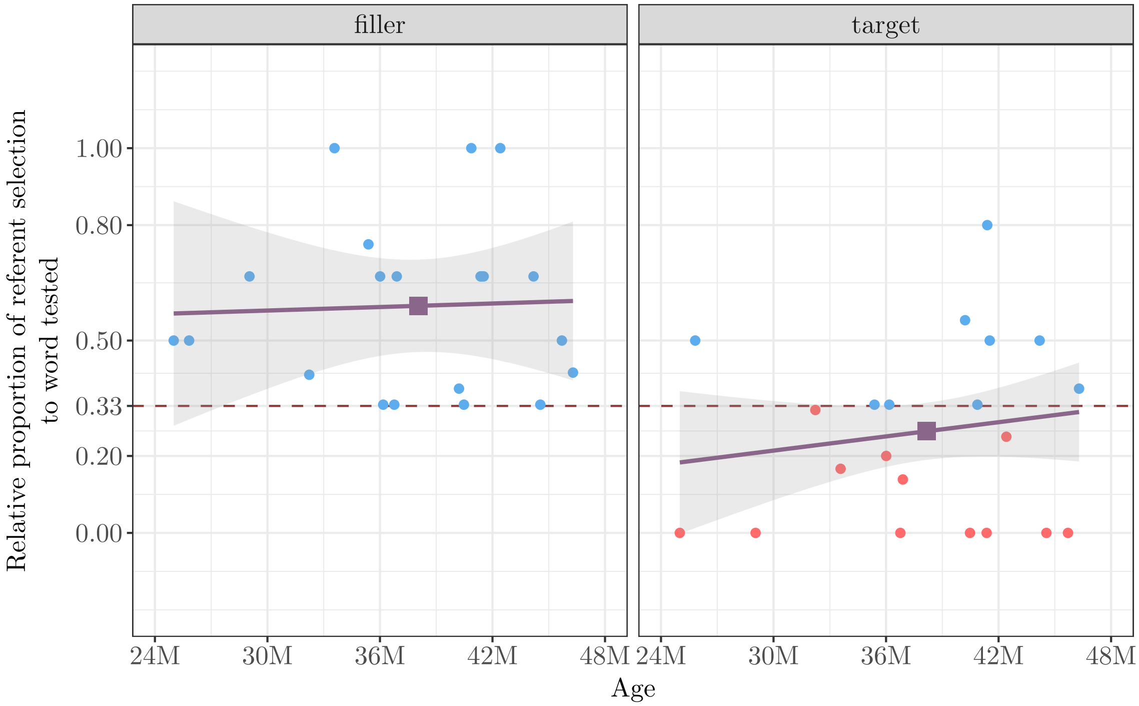

Fig. 3.1 Word learning performance via referent selection by child's age.

Note: Averages for each participant are colored red when below chance (dotted red line). Linear fit and grand mean shown in purple.

3.1.2 Gaze

The relationship between age and referent selection, shown in Figure 3.2, was not consistent with expectations. As with referent selection, children performed better on filler trials ($M \approx 0.5$) compared to target trials ($M \approx 0.3$). Yet with gaze, there did not appear to be a clear relationship between age and relative proportion of time gazing towards the word tested. All ages performed similarly on target trials. Unlike referent selection, only six children performed above chance for gaze, and their performance on the word learning task was only slightly above 0.33.

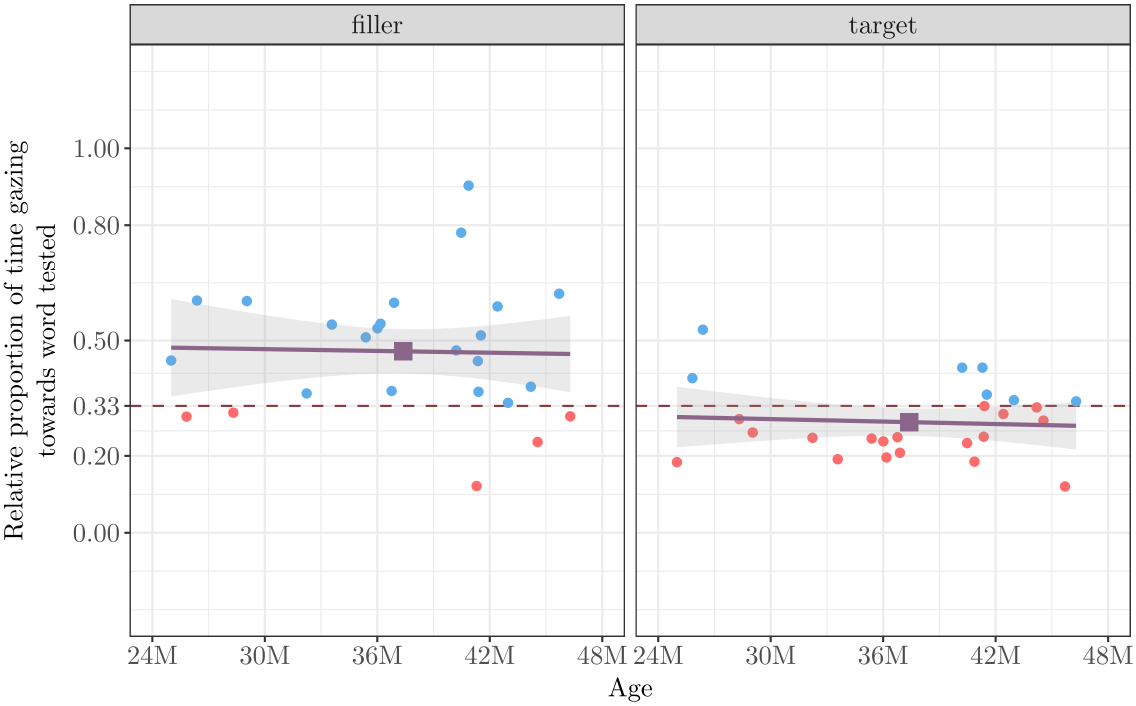

Fig. 3.2 Word learning performance via gaze by child's age.

Note: Averages for each participant are colored red when below chance (dotted red line). Linear fit and grand mean shown in purple.

There is not a clear relationship between age and gaze, which did not follow our hypothesis that there would be a positive age effect on word learning via gaze. Yet the gaze data in Figure 3.2 shows, in comparison with referent selection, that gaze may be a better indicator of word learning for younger children; on this task, their performance based on gaze was higher than based on referent selection. Until this study, gaze has not been analyzed in a fully remote implementation of a novel word learning study that we know of, so these results suggest that the analysis methods used for in-person word learning studies may not directly translate to remote word learning studies. Due to the lack of a clear age effect and because the majority of children are not performing above chance on the task, we decided to not create the engagement moderator models using gaze.

3.1.3 Engagement

Below we describe the distribution and correlations between engagement measures. Engagement was measured through two approaches, observational measures (the child's verbal and gestural responses during the session) and holistic measures (reported from parents and experimenter perspectives). Descriptive statistics are provided for observational and holistic engagement in Table 3.2. Observational engagement was measured during both the teaching and testing phases.

In terms of observational engagement in each phase, children on average responded more frequently (relative to duration) during the teaching phase (3.67 verbal or gestural responses per minute) compared to testing (2.15 verbal or gestural responses per minute). There was large variation, based on standard deviation values, in verbal and gestural responses during both phases.

In terms of holistic engagement, on average both parents and experimenters rated children overall more engaged for all measures (value of 1.5 middle of our scale) except for experimenter ratings of social engagement. Social engagement also had the largest standard deviation, nearly double compared to other measures, suggesting that values had a larger range than other measures. These general trends for both observational measures and holistic measures are investigated further below.

Table 3.2 Descriptive statistics of engagement measures ($n = 21$)

	M	SD	min	max
Observational engagement				
overall score	25.95	18.73	5	82
overall verbal measures	17.90	17.01	2	72
overall gestural measures	8.05	5.25	0	20
Teaching				
teaching score	10.86	11.46	0	46
verbal measures	5.24	9.03	0	39
gestural measures	5.62	5.20	0	19
Testing				
testing score	15.10	9.75	2	36
verbal measures	12.67	9.61	2	33
gestural measures	2.43	1.78	0	7
Holistic engagement from parent				
activity interest	1.71	0.72	0	3
responsiveness	2.10	0.77	0	3
perceived engagement	2.29	0.72	1	3
activity focus	2.62	0.80	1	4
emotional state	2.29	0.85	0	3
social engagement	1.52	1.12	0	3
Holistic engagement from experimenter				
activity interest	2.33	0.48	2	3
responsiveness	1.90	0.62	1	3
perceived engagement	2.38	0.50	2	3
activity focus	2.81	0.51	2	4
emotional state	2.48	0.51	2	3
social engagement	1.19	1.08	0	3

Observational measures of engagement Figure 3.3 shows the trend between observational engagement and referent selection did not align with predictions. There is a difference in trends between performance on filler trials and target trials. On filler trials, observational measures of engagement and children's selection of referents positively correlated (slope $= 0.0028, R^2 = 0.013; r = 0.116, p = .387$). Conversely, on target trials, children appeared very engaged even though most children did not learn to identify the correct referents for the novel words (slope $= -0.0006, R^2 = 0.00072; r = -0.027, p = .780$).

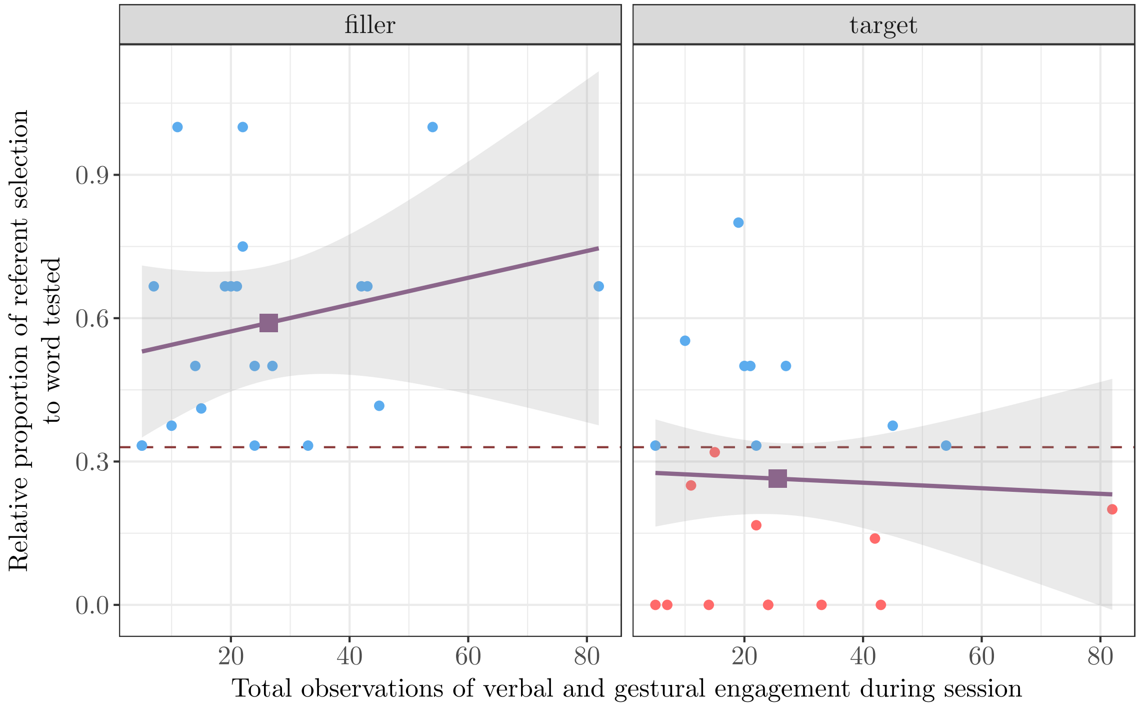

Fig. 3.3 Performance on referent selection by observational engagement

Note: Averages for each participant are colored red when below chance (dotted red line). Linear fit and grand mean shown in purple.

Additionally, Figure 3.4 shows that engagement scores and referent selection scores are widely distributed for filler trials and target trials across the age range; therefore, the trend lines shown in the plots do not strongly fit either set of data.

Further, separating observational engagement from the teaching and testing phases, we analyzed our two observational approaches, verbal and gestural measures. Overall, there were positive correlations between both approaches in teaching and the verbal measures in testing. There was one significant correlation between the verbal scores during the teaching phase and during the testing phase ($r = 0.667, p = .001$). There was little to no correlation

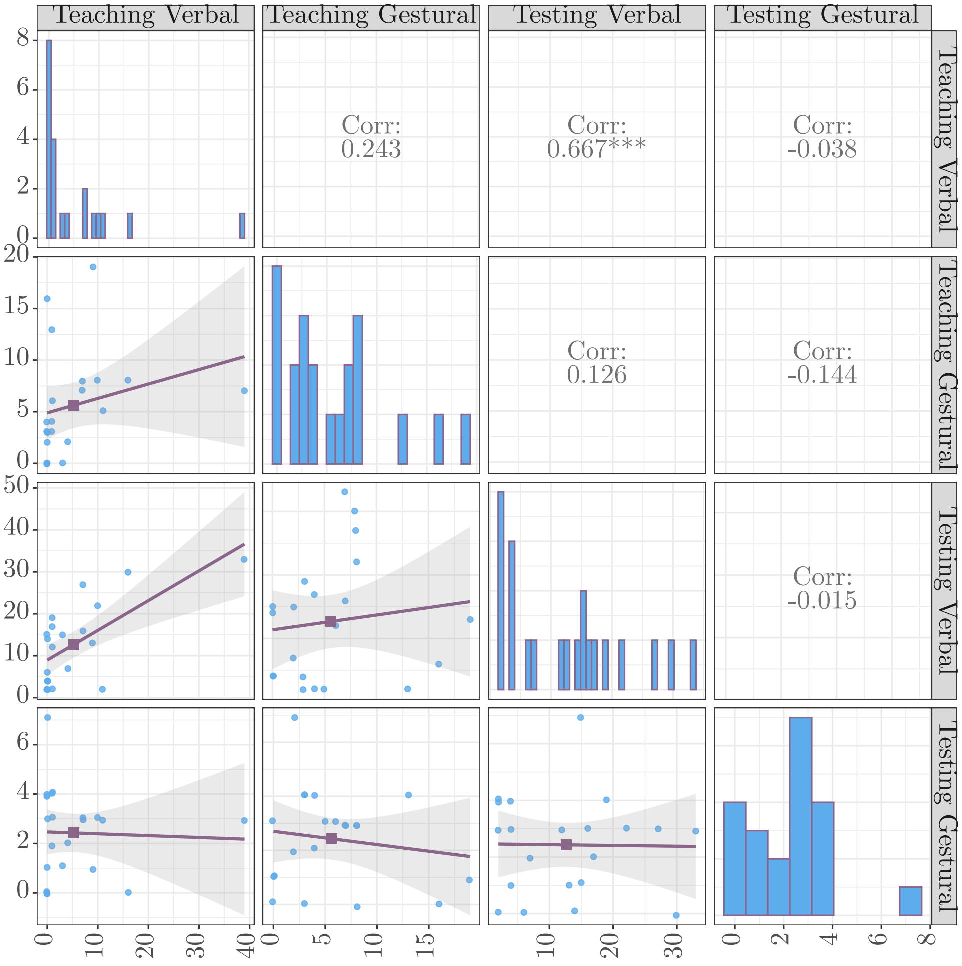

Fig. 3.4 Correlations of observational measures of engagement between teaching and testing phases

Note: p-values shown as $*** < .001$, $** < .01$, $* < .05$, $. < .1$

between gestural measures during testing and the other measures. This is unsurprising as only direct gestural responses and communicative/conventional gestures were included in the gestural measures during testing, since the other measures were removed due to reliability or overlap with the measure of word learning for referent selection; which also explains why there were very few gestural observations for each participant (max 7, in the bottom row in Figure 3.4) compared to the verbal observations. We also note that these measures were not normally distributed, and there appears to be a few outliers. We did not remove these outliers for our analysis, in order to retain the individual variability for each child's engagement.

Similar to the trend in the sum of all observational measures shown in Figure 3.3, we saw slightly negative trends from the overall observational engagement measures collected from the testing phase primarily (right target panel, Figure 3.5). Additionally, we saw that children who had higher verbal and gestural responses during teaching did not always have more responses during testing. These data from children with this response profile can be found in the top-left and bottom-right corners of each correlation plot in Figure 3.4.

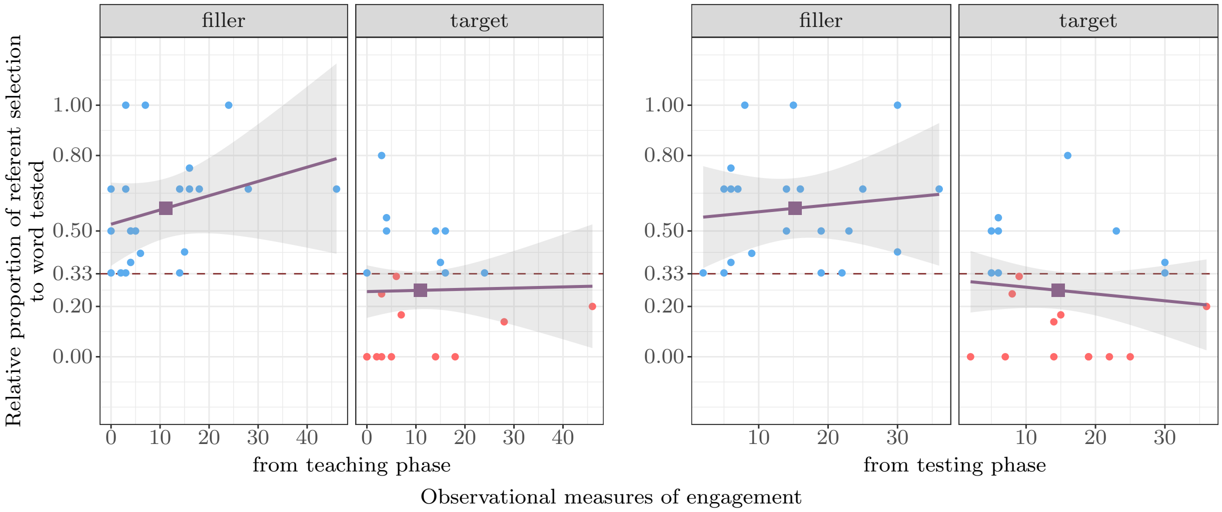

Fig. 3.5 Performance on referent selection by sum of observational engagement in teaching and testing phases

Note: Averages for each participant are colored red when below chance (dotted red line). Linear fit and grand mean shown in purple.

Holistic measures of engagement To analyze the holistic measures of engagement, we first observed the correlations between survey questions. From Figure 3.6, parent responses showed that nearly all holistic engagement measures were positively correlated with one another. The exceptions were the correlations between activity focus and activity interest

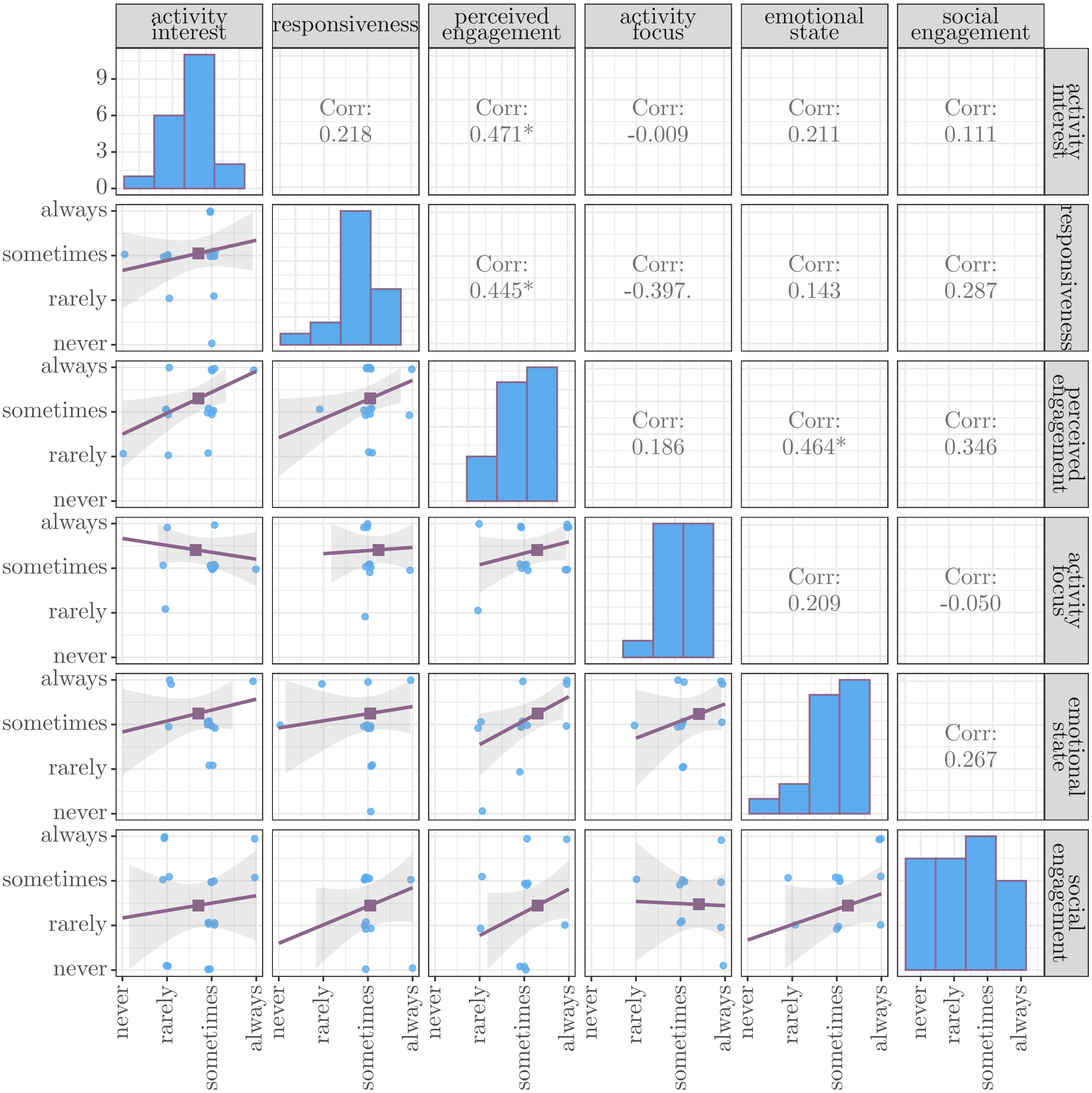

Fig. 3.6 Holistic engagement correlations from parent survey

Note: p-values shown as $*** < .001$, $** < .01$, $* < .05$, $. < .1$

($r = -0.009$) and between activity focus and responsiveness ($r = -0.397$); both slightly negative. The correlation between activity focus and social engagement was marginally significant ($p = .083$). The most significant correlations, which were only moderately significant, were between perceived engagement and activity interest ($r = 0.471, p = .036$) and between perceived engagement and emotional state ($r = 0.464, p = .039$), and between perceived engagement and responsiveness ($r = 0.445, p = .049$).

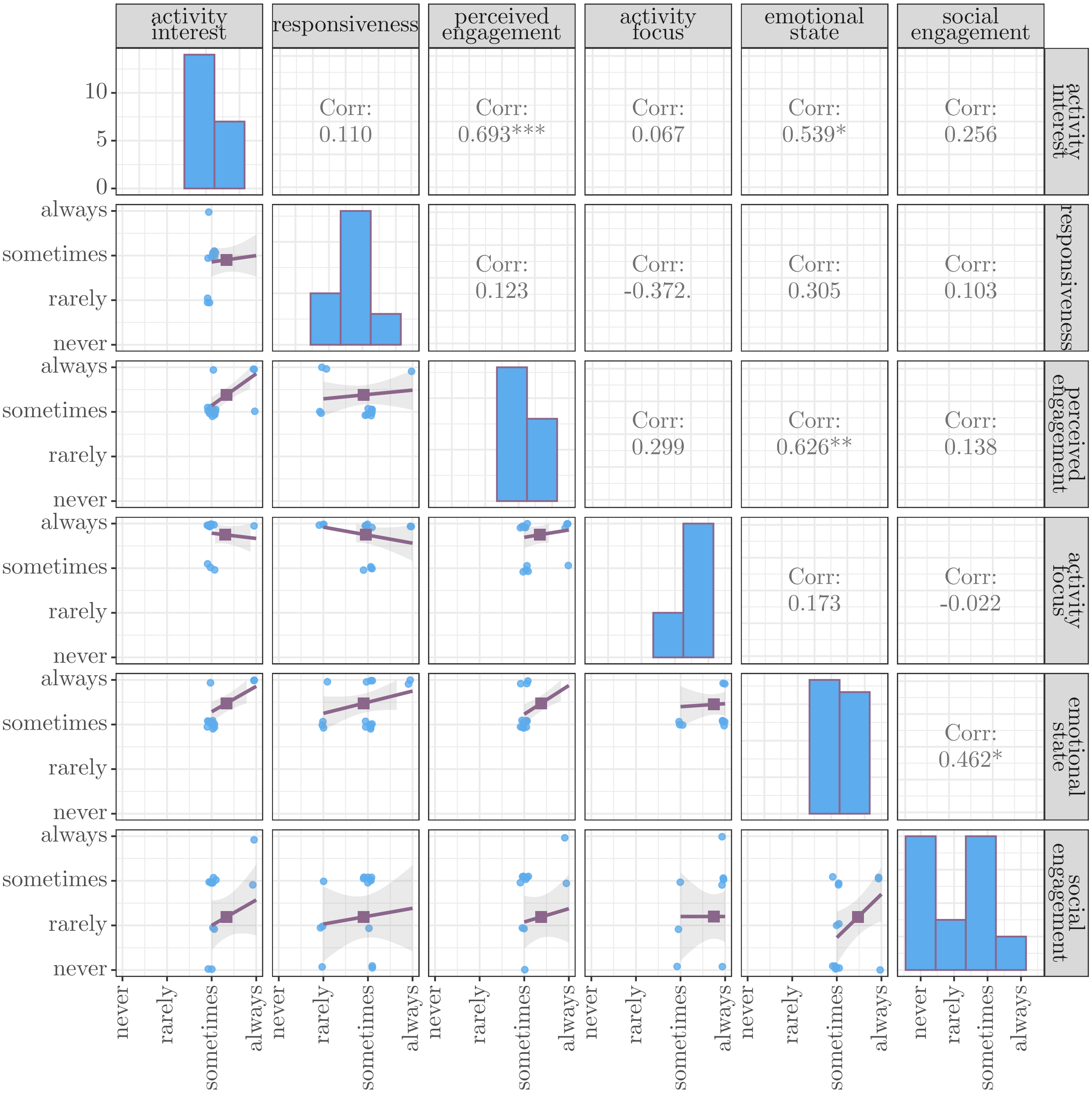

Fig. 3.7 Holistic engagement correlations from experimenter survey

Note: p-values shown as $*** < 0.001$, $** < 0.01$, $* < 0.05$, $. < 0.1$

Similarly, we analyzed the correlations among the experimenter's responses to the survey (Figure 3.7). All correlations were positive, except for activity focus and responsiveness and between activity focus and social engagement. One notable difference from parent surveys was that the experimenter surveys had a much narrower range, often being reported as "sometimes" or "always." There were three significant correlations, between perceived engagement and activity interest ($r = 0.693, p < .001$), between perceived engagement and emotional state ($r = 0.626, p = .002$), and between social engagement and emotional state ($r = 0.462, p = .035$).

Finally, we analyzed correlations between parent and experimenter ratings. Figure 3.8 shows the results for each holistic measure. The strongest correlations were for social engagement ($r = 0.66, p = .001$) and responsiveness ($r = 0.44, p = .048$), suggesting that

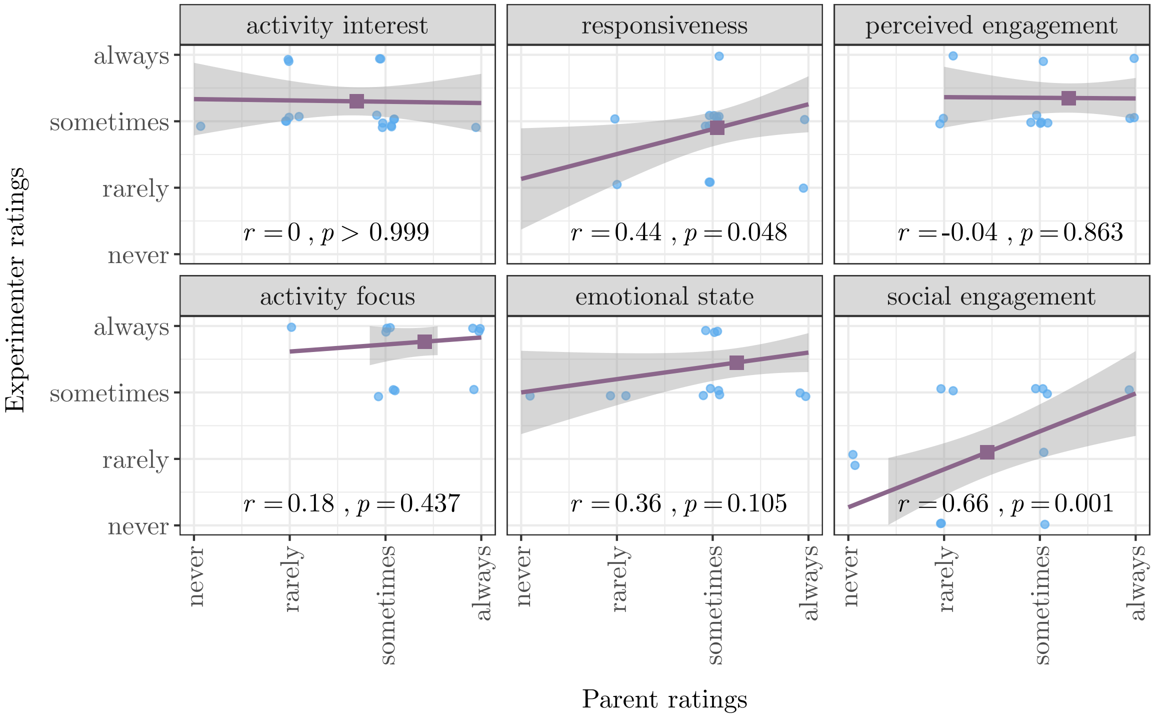

Fig. 3.8 Comparison of holistic engagement between parent and experimenter survey

parents and experimenters rated these two qualities of engagement most similarly. There was no correlation between perspectives for ratings of activity interest ($r = 0, p > .999$) and for perceived engagement ($r = -0.04, p = .863$). Measures of emotional state and activity focus had mild correlations and were not significant.

Again, we investigated the trends between holistic engagement and referent selection. Figure 3.9 shows the trends from both perspectives. Here survey responses have been aggregated into a single holistic engagement score (range from 0-24). We see very similar trends across both perspectives, where higher survey ratings correspond to higher referent selection performance for both filler and target trials. Notably, it appears that for target trials the experimenter perspective has a slightly steeper slope than the parent perspective. However, there is a wide variation in proportion referent selection of target for the holistic measures from the experimenter, so the linear fit may not be the best indicator of the trend.

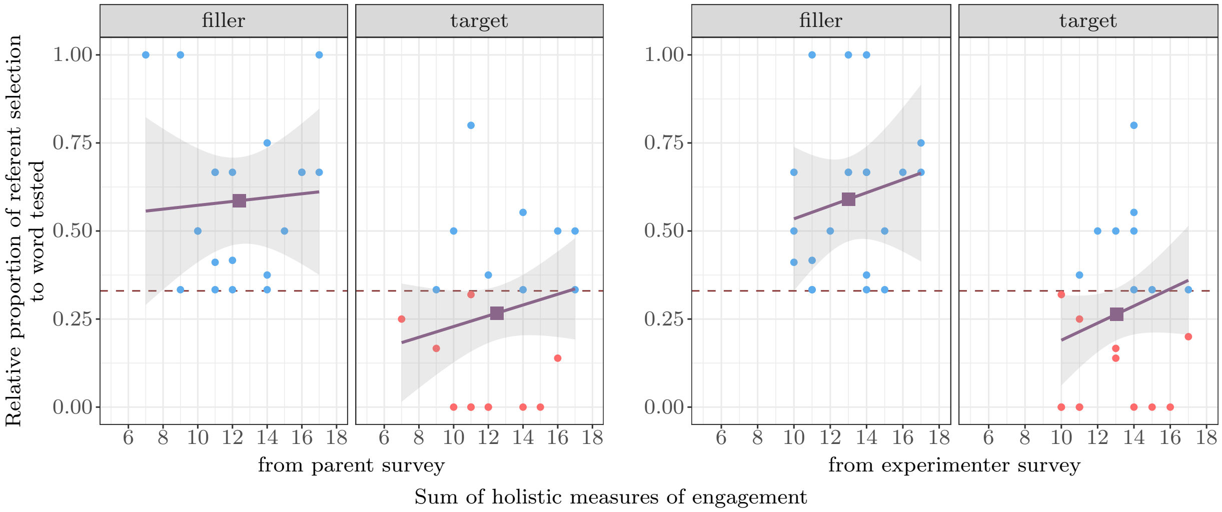

Fig. 3.9 Performance on referent selection by holistic engagement

Note: Averages for each participant are colored red when below chance (dotted red line). Linear fit and grand mean shown in purple.

3.2 Analysis 1: Engagement moderating word learning

To answer the two research goals we had two main analyses. The primary analysis was the moderator analysis to investigate the effect of engagement on the relationship between age and word learning outcomes using the linear model (`lm`) function in R (R Core Team, 2020). This package calculates p-values using the Satterwaite's approximation. All predictor variables were transformed using `arm::rescale` to standardize and normalize variables. This recenters around the grand mean and scales each variable relative to the standard deviation of the distribution, and enables comparison of the model estimates.

As indicated in the Introduction, background, and research aims section, we do not have prior studies employing engagement as a moderator for word learning outcomes to use as a benchmark for *a priori* power analysis. Clearly our current, preliminary sample of only 21 participants is statistically under-powered to evaluate a complex regression model with 12 predictors total (including 2 and 3 way interactions between age and our 3 aggregate measures of engagement, described in detail below), (e.g., Wilson VanVoorhis & Morgan, 2007). One way method to address our limited statistical power was to expand the amount of data collected from each participant, so we analyzed learning outcomes from each trial ($n = 106$). This assumes that each trial is independent (which in fact, in our study these were repeated measures presented in sequence) and that engagement measures could vary by participant by trial (which was true for some but not all of our engagement measures). Results from a post hoc power analysis run on our regression model for Aim 1, with $n = 106$ trials ($f^2 = 0.142, \alpha = 0.05, n_{\text{parameters}} = 12$) support that this model does have sufficient power, 0.986, to rule out Type I error. However, the assumptions made in this power analysis with regard to trial independence are likely not met due to our study design. Given this fact, and that our current sample only includes 21 children, rather than statistical significance, we focus on effect sizes (Cumming, 2014). To obtain comprehensive experience with data analysis and interpretation through this master's thesis, we interpret the trends in our preliminary data based on current effect sizes with plans to confirm these effects with a larger sample.

We wanted a model that captured the relevant interactions between age and each measure of engagement to look at how engagement moderated word learning (via referent selection). As discussed above, verbal and gestural measures were summed across the entire teaching phases into an overall score of observational engagement and survey scores from the parent and experimenter perspectives were summed into two overall scores of holistic engagement. This provided three primary measures of engagement for our moderator analysis: observational engagement (during teaching phase only), holistic engagement ratings from the parent, and holistic engagement ratings from the experimenter.

The model used for the moderator analysis includes 4 first-order predictors of age and each of the three measures of engagement above; 3 second-order interaction terms between age and each total score of engagement (Age:observational engagement, Age:holistic engagement parent, Age:holistic engagement experimenter), 1 second-order interaction term between observational engagement:holistic engagement experimenter; and 3 third-order interaction terms with age (Age:observational engagement:holistic engagement parent, Age:observational engagement:holistic engagement experimenter, Age:holistic engagement parent:holistic engagement experimenter). This model analyzes all the interactions between age and our engagement metrics that we were interested in.

The regression model (shown in Table 3.3) was selected because it specifically looks at the interaction term to see if our measures of engagement moderate the relationship to age. Other models were considered; however, the regression model used to predict word learning retains all age related interaction terms and reduced non-significant terms to avoid overfitting.

The regression model ($R^2 = 0.224, AIC = 106, BIC = 143, p = 0.0155$) passed an omnibus test compared to the model with only 1 first-order predictor of age ($R^2 = 0.00795, AIC = 115, BIC = 124, p = 0.352$). The omnibus model was a slightly better fit based off of BIC values but had a $p > 0.05$ suggesting that the predictor of age alone fails to reject the null hypothesis for our data set; however, we can reject the null hypothesis with the regression model results.

Table 3.3 Regression model of engagement predicting word learning

	$\hat{\beta}$	$SE(\hat{\beta})$	t	p	
Intercept	0.45	0.09	5.20	0.000	***
Age	-0.13	0.20	-0.63	0.53	
Observational teaching	-0.13	0.30	-0.44	0.66	
Holistic parent	0.08	0.26	0.32	0.75	
Holistic experimenter	0.10	0.14	0.71	0.48	
Age:Observational teaching	0.62	0.52	1.19	0.24	
Age:Holistic parent	-0.43	0.38	-1.14	0.26	
Age:Holistic experimenter	0.16	0.31	0.53	0.60	
Observational teaching:Holistic parent	-0.36	0.37	-0.97	0.34	
Observational teaching:Holistic experimenter	-1.14	0.43	-2.68	0.009	**
Age:Observational teaching:Holistic parent	2.82	0.94	2.99	0.004	**
Age:Observational teaching:Holistic experimenter	0.27	0.81	0.33	0.74	
Age:Holistic parent:Holistic experimenter	-1.49	1.05	-1.42	0.16	

$R^2_m = 0.224, R^2_c = 0.124, \sigma = 0.373, df = 12, n = 106, AIC = 106$

Continuous predictors are mean-centered and scaled by 1 s.d.

Results from the model suggest that, within target trials only, age alone does not predict word learning performance ($\hat{\beta} = -0.13, SE(\hat{\beta}) = 0.20, p = 0.53$). We can see that the confidence intervals on the linear fit in the right panel of Figure 3.1 are relatively large, which may explain why age alone was not a predictor. Similarly, no first-order term of our engagement measures was a strong predictor.

Second-order interactions between age and the engagement terms had relatively small influence on the overall model. One second-order interaction between observational engagement during teaching and holistic engagement ratings from the experimenter did have an effect ($\hat{\beta} = -1.14, SE(\hat{\beta}) = 0.43, p = 0.009$). This suggests that higher holistic ratings of the child's engagement from experimenters are important when predicting learning outcomes, particularly for children who showed lower observational engagement during teaching (as compared to high observational engagement during teaching). It is important to note that the experimenter was not aware of the child's performance on the task when complete the engagement survey, since they could not see which object the child indicated during the session.

The third-order interaction between age, observational engagement during teaching, and holistic engagement ratings from the parent was the largest predictor in the model ($\hat{\beta} = 2.82, SE(\hat{\beta}) = 0.94, p = 0.004$). This third-order interaction is best understood by how the model predicts learning outcomes for a younger and older child with high/low scores for observational engagement and high/low scores for holistic ratings from their parent. This interaction term suggests older children who have both high observational engagement and high holistic ratings from their parent perform significantly better than younger children with high observational engagement and high holistic ratings. Conversely, older children with high observational engagement but low holistic ratings from their parent perform worse than younger children high observational engagement but low holistic ratings. Moreover, older children with high observational engagement but low holistic ratings from their parent perform comparatively worse than older children with low observational engagement and low holistic ratings from their parent; similarly but to a much smaller degree, younger children with high observational engagement but low holistic ratings from their parent perform slightly worse than younger children with low observational engagement and low holistic ratings from their parent. This interaction term suggests that alone high observational engagement scores do not correlate to higher word learning performance, and there is a non-trivial age effect where high observational engagement actually predicts lower performance for older children; interestingly, the third-order interaction term suggests parent ratings are able to distinguish this non-trivial effect.

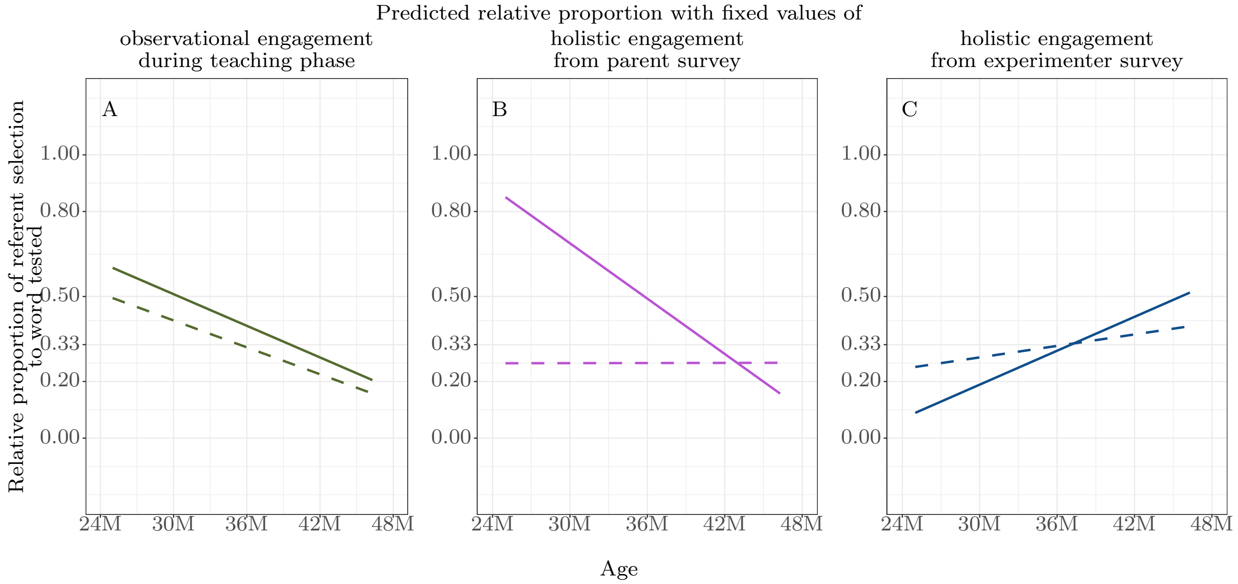

Fig. 3.10 Predicted performance on referent selection with fixed scores for engagement measures.

Note: Fixed values for each respective engagement measures was approximated as low scores with the first quartile value (dotted line) and high scores with the third quartile value (solid line) for the dataset. Predicted values based on our sample; mean age was 38M13.

There is an additional third-order term between age, holistic parent ratings, and holistic experimenter ratings which has an effect on overall word learning outcomes $(\hat{\beta} = -1.48, SE(\hat{\beta}) = 1.05, p = 0.16)$. This term relates the trends shown in Figure 3.9.

The interaction terms with the largest magnitudes in the model are difficult to interpret from Table 3.3. Figure 3.10 shows the model predictions with fixed values of engagement (high/low scores). For each respective engagement measure high scores were approximated with the third quartile values from the dataset, and low scores (shown as dotted lines) where approximated with the first quartile values. Panel A shows that our observational engagement measure moderates word learning (simply, higher scores relate to higher word learning performance) but the overall trend with age does not really change. Panel B, instead suggests that holistic ratings from parents may be better predictors for young children, whereas Panel C suggests that holistic ratings from experimenters may be stronger predictors for older children.

3.3 Analysis 2: Exploratory factor analysis and updated moderator analysis

We used exploratory factor analysis (EFA) to further understand how the engagement measures collected explained the variance in our data. Using the `psych` package in R, we used the factor analysis function (`fa`) and the correlation function (`cor`, Revelle, 2021). For EFA, we refer to the input variables, our engagement measures, as factors. The correlation function computes the correlation matrix which will provide eigenvalues for factor loadings for the EFA. Typically, low factor loadings (i.e., below 0.3) indicate that too many input factors are used, and some factors should be combined before rerunning the analysis. Factor ratings below 0.5 are sufficient for analysis but typically indicate poor prediction power. We used ordinary least squares to calculate the minimum residual solution for the EFA.

Table 3.4 shows the factor loadings of the EFA. We found 3 latent factors for the EFA, and based on the loadings characterized these as *MR1:* social/emotional, *MR2:* responsiveness, *MR3:* enthusiasm. We characterized the first as social/emotional interest, because it primarily is derived from social/emotional scores (perceived engagement, social engagement, emotional) from the experimenter survey and as well as scores for direct verbal responses and activity focus. The second latent factor was defined

Table 3.4 Factor loadings from exploratory factor analysis

	MR1	MR2	MR3
	social/emotional	responsiveness	enthusiasm
Observational engagement during teaching			
all verbal engagement		0.68	
all gestural engagement	0.98		
Observational engagement during testing			
repetition			
direct response verbal			0.49
vocalization			
comment to parent			
comment to teacher		0.49	
conventional/communicative gesture			
Holistic engagement from parent			
activity interest		0.79	
responsiveness		0.63	
perceived engagement		0.70	0.35
activity focus			0.45
emotional state	0.33		0.39
social engagement	0.86		
Holistic engagement from experimenter			
activity interest			0.76
responsiveness			
perceived engagement			0.85
activity focus		0.39	
emotional state	0.42		0.53
social engagement	0.78		

Proportional variance: social/emotional (15%), responsiveness (14%), enthusiasm (13%)
Three latent factors explain 42% of variance among the measures of engagement

Factor correlations: social/emotional and responsiveness (0.30), social/emotional and enthusiasm (0.14), responsiveness and enthusiasm (-0.04)

as responsiveness, since all measures relate to the child's responsiveness, social engagement, and comments to the teacher. Finally, the third latent factor was called enthusiasm as this factor related the child's perceived engagement and interest with their verbal responses and emotional state. It should be noted that these characterizations of the latent factors are not strictly defined by factors and weights in our dataset. Results should not be interpreted as definitions of the underlying factors. However, our characterizations aim to generalize the results specific to our data to similar contexts, and suggest further analysis into possible aspects of engagement.

Together, each latent factor explains around 14% of the variance in our dataset, totaling 42% cumulative variance overall.[1] We settled on three latent factors for two reasons. First, the Velicer Minimum Average Partial (MAP) criterion achieves a minimum of 0.051 with three factors, which is a common criterion to select the number of factors to use (Courtney & Gordon, 2013). Second, three factors allows us to have a regression model structure that mirrors the regression model from Analysis 1.

The EFA regression model ($R^2 = 0.21, AIC = 105, BIC = 148, p = 0.203$) again passed an omnibus test compared to the model with only 1 first-order predictor of age ($R^2 = 0.00795, AIC = 115, BIC = 124, p = 0.352$). However, we fail to reject the null hypothesis with this model (as $p > 0.05$). As for the prior regression model, we plan to reassess this model with a larger sample, and in the current thesis we focus on the effect size of preliminary results, if they were to continue to pattern this way in a larger sample.

Table 3.5 shows the results for the EFA regression model. Results from the EFA model suggest that these latent factors of engagement may moderate the age effects of word learning. Again, age alone was again not a strong predictor of word learning performance ($\hat{\beta} = 0.21, SE(\hat{\beta}) = 0.14, p = 0.13$). Similarly, each first-order term of our latent factors alone had relatively small effects in the model.

The second-order interaction between age and the responsiveness latent factor was a

[1] Often the rule of thumb for EFA is to keep enough factors until 50% variance is explained (Peterson, 2000). However, in our case, additional factors have low loadings ($< 10\%$) and do not significantly improve explained variance. Further investigation showed that additional factors were composed of different combinations of the same underlying variables. Thus, to avoid overfitting (with our limited sample) and overestimating the number of factors needed, we proceed with analysis despite our cumulative variance being only 42%.

Table 3.5 Regression model of latent factors of engagement predicting word learning

	$\hat{\beta}$	$SE(\hat{\beta})$	t	p
Intercept	0.21	0.14	1.54	0.13
Age	0.36	0.37	0.98	0.33
social/emotional	0.41	0.40	1.02	0.31
responsiveness	0.00	0.41	0.01	0.99
enthusiasm	0.05	0.26	0.19	0.85
Age:social/emotional	-1.52	1.16	-1.31	0.20
Age:responsiveness	1.54	0.55	2.79	0.0067 **
Age:enthusiasm	1.05	0.65	1.62	0.11
social/emotional:responsiveness	-0.19	0.98	-0.19	0.85
social/emotional:enthusiasm	0.21	0.76	0.28	0.78
responsiveness:enthusiasm	0.74	0.94	0.78	0.44
Age:social/emotional:responsiveness	-0.79	3.10	-0.26	0.80
Age:social/emotional:enthusiasm	-4.39	2.20	-2.00	0.05 *
Age:responsiveness:enthusiasm	2.54	1.33	1.90	0.06 .
social/emotional:responsiveness:enthusiasm	1.62	2.34	0.69	0.49
Age:social/emotional:responsiveness:enthusiasm	-11.38	5.06	-2.25	0.03 *

$R_m^2 = 0.211, R_c^2 = 0.053, \sigma = 0.395, df = 15, n = 91, AIC = 105$

Continuous predictors are mean-centered and scaled by 1 s.d.

Note: number of observations were decreased due to some missing survey data from two pilot participants.

predictor ($\hat{\beta} = 1.54, SE(\hat{\beta}) = 0.55, p = 0.0067$). Here older children who showed high responsiveness during teaching performed significantly better on our word learning task than older children with low responsiveness; these effects were smaller in younger children (depicted in Panel B in Figure 3.11).

Furthermore, we see two third-order interactions with age, social/emotional, and enthusiasm ($\hat{\beta} = -4.39, SE(\hat{\beta}) = -2.00, p = 0.05$) and age, responsiveness, and enthusiasm ($\hat{\beta} = 2.54, SE(\hat{\beta}) = 1.33, p = 0.06$) which were both large predictors The interaction with age, social/emotional, and enthusiasm has a negative sign which suggests that younger children with higher social/emotional scores and higher enthusiasm perform better than older children with higher social/emotional scores and higher enthusiasm. This same trend

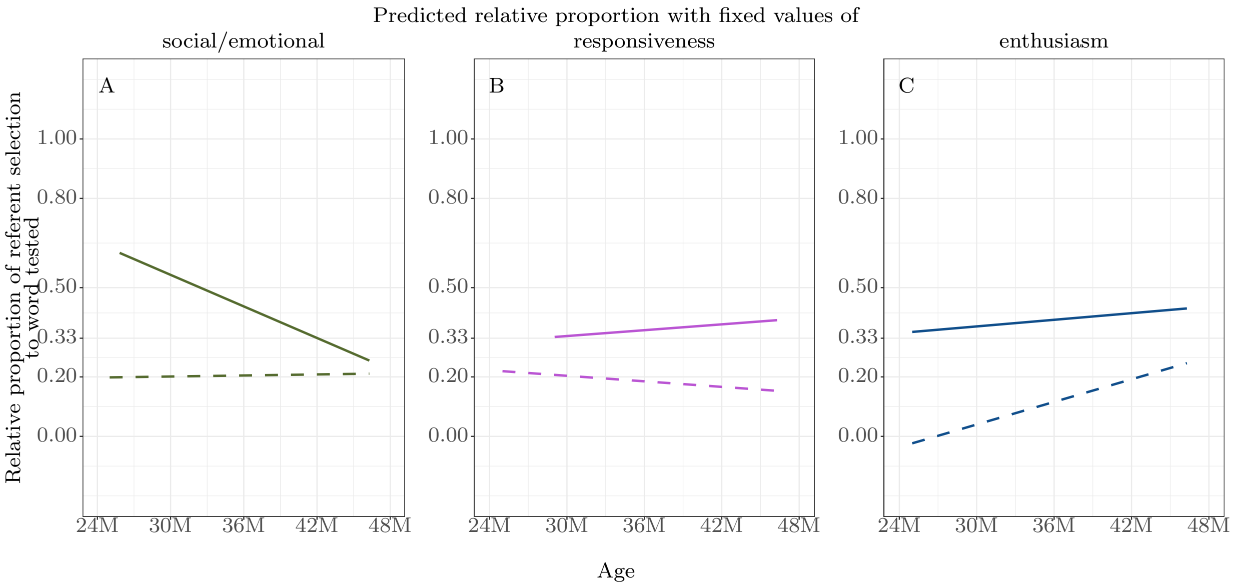

Fig. 3.11 Predicted performance on referent selection with fixed scores for EFA latent factors.

Note: Fixed values for each respective latent factor was approximated as low scores with the first quartile value (dotted line) and high scores with the third quartile value (solid line) for the dataset. Predicted values based on our sample; mean age was 38M13

occurs between younger and older children with low social/emotional scores but higher enthusiasm, as well as both low social/emotional scores and low enthusiasm. However, the model predicts higher performance in the case where older children have low enthusiasm scores but high social/emotional scores (compared to young children with the same profile). The interaction with age, responsiveness, and enthusiasm suggests older children with high scores of responsiveness and enthusiasm perform better than young children with similar scores. Yet, this is only true when both scores are high, otherwise older children perform worse when there is a mismatch between high/low scores for responsiveness and enthusiasm than younger children. Interestingly, the EFA model predicts here that high enthusiasm with low responsiveness actually leads to lower word learning outcomes.

Finally, the fourth-order interaction term with age, social/emotional, responsiveness, and enthusiasm is also predictive ($\hat{\beta} = -11.38, SE(\hat{\beta}) = 5.06, p = 0.03$). This term is significantly more complex to decompose. The main effects suggest that high social/emotional scores predict word learning outcomes better for younger children (regardless of scores on responsiveness and enthusiasm). There is a non-trivial age effect that predicts older children with low enthusiasm scores but higher social/emotional and responsiveness scores will perform better overall. The exception to this is when older children have high scores for all three latent factors of engagement, suggesting that the opposing trends shown in Figure 3.11 moderate word learning performance differently for younger children versus older children.

Results from the EFA model suggest that there is a complex interaction between the latent factors of engagement, and that the impact of each factor contributes to word learning outcomes differently as children get older. We do note that from the EFA model, responsiveness appears to be one of the key factors, as three out of the four significant predictors include responsiveness. When taken independently, we see in Panel B in Figure 3.11 that responsiveness is the only latent factor that predicts older children perform better particularly with higher responsiveness scores. Social/emotional and enthusiasm scores appear to better distinguish word learning outcomes for younger children.

Discussion

This study was designed to investigate whether engagement moderates young children's performance on word learning tasks. We expected for all ages, higher engagement would improve performance on the word learning task. While social factors like social contingency and co-viewing have been shown to improve learning outcomes for children, these factors do not consider how the child is interacting with the activity. From the student engagement literature, it is clear that engagement incorporates many cognitive, emotional, and behavioral aspects of the child's interactions. Models like Language Task Engagement propose factors beyond social contingency have some role in children's language task outcomes (see Figure 1.1; Egbert et al., 2021). We wanted to investigate the role that the child's level of engagement had on learning outcomes, which has not been explicitly modeled in novel word learning tasks.

We had three ways of measuring word learning: referent selection, gaze, and label production. Unfortunately, children performed above chance with only one of these measures. Very few children produced any novel labels during label production. Gaze, which is frequently used in word learning studies, showed very few children gazed at the correct object above chance on the target trials. This study is the first to our knowledge to collect gaze data in a completely remote word learning study. Thus, we were left with only a single measure of word learning, referent selection via gestures such as pointing or reaching toward the picture of the target object. Conducting a word learning study remotely with toddlers was challenging and, without prior studies to base our results, our conclusions serve as starting point to guide future research on remote wording learning methodology. Our preliminary results suggest that traditional word learning measures may not work in remote settings, which we discuss in more detail in the 4.2 section below.

Referent selection outcomes suggest this was an age-appropriate task for the range of children tested, while performance was lower than expected. Generally, older children performed better on the task (Figure 3.1). This motivates the research aims for this study, which were to answer (1) Does engagement moderate word learning outcomes? and (2) Which engagement measures best explain the variance in word learning outcomes? More generally, this project explored how we measure engagement in word learning tasks. We analyzed how our engagement measures correlate and predict word learning outcomes. Our analysis of the selected measures via exploratory factor analysis suggest aspects of engagement that may be more predictive of word learning outcomes at various ages. This study provides preliminary information which may inform future efforts to understand the role of engagement in toddler's word learning in an experimental paradigm, particularly for remote settings.

Does engagement moderate word learning? We chose two approaches to measure engagement which can be operationalized to other word learning studies: observational engagement, a count of the child's verbal and gestural responses to the task, and holistic engagement, derived from survey responses. Holistic engagement measures were collected from two perspectives, the parent and experimenter. Generally, holistic measures from both perspectives correlated with each other; yet, the two perspectives had different predictive power when predicting word learning outcomes.

Overall, preliminary regression models with a small sample (Table 3.3) suggest that observational and holistic measures of engagement may moderate word learning outcomes. However, a single measure of engagement failed to moderate word learning outcomes exclusively; instead, a complex combination of the engagement measures predicted the word learning outcomes observed in our sample. Furthermore, only looking at holistic ratings from only one perspective would not capture this complexity.

Two important trends from the regression model results showed that high holistic engagement ratings from experimenters are important when predicting learning outcomes, particularly for children who showed lower observational engagement during teaching. The second-order interaction between observational engagement during teaching and holistic engagement ratings from the experimenter suggests that high holistic ratings from experimenters are important when predicting learning outcomes, particularly for children who showed lower observational engagement during teaching (as compared to high observational engagement during teaching). This suggests that experimenter's holistic ratings were useful in distinguishing outcomes when children were not interacting during the teaching session. This highlights a common case in toddler studies, where a child is focused and attending to an activity but not actively participating though verbal or gestural responses.

Secondly, our model suggests a potential interaction between age, observational engagement during teaching, and holistic ratings from the parent. This interaction suggests older children that have both high observational engagement and high holistic ratings from their parent perform better than younger children with high observational engagement and high holistic ratings. Simply, high parent ratings and lots of engagement suggest higher word learning performance. Yet, this result suggests that parent ratings may provide insight when observational results do not correlate with the child's performance as we expect.

These results suggest different word learning outcomes for different profiles of children. Broadly, older children perform better when they have high scores for both observational engagement and parent ratings, compared to younger children with similar scores. Yet, parent ratings are able to distinguish when older children have low word learning outcomes (even when the child had high observational engagement scores). This may be, for example, an older child who is intentionally giving wrong answers, or an older child who completes the task but continues to respond to the other objects in the task. While our objective measures of observational engagement did not distinguish these children, parent ratings were sensitive to children with this profile.

Latent factors of engagement Our second aim was to further investigate how our engagement measures were interacting and which measures described the variance in our data. We added our engagement measures to an exploratory factor analysis, which suggest three latent factors of engagement that our measures capture: social/emotional, responsiveness, and enthusiasm. These results (Table 3.4) suggest that engagement measures do not cleanly separate into the theoretical categories like behavioral/emotional/cognitive or even into methodological approaches like observational/holistic; instead aspects of each category and approach contributed to the latent factors which described the variability in the children's engagement. Specifically, all three latent factors included both observational and holistic measures, as well as holistic measures from both perspectives.

Using these latent factors in a preliminary regression model suggests that there maybe different moderating factors on the age effects of word learning outcomes. While the responsiveness factor appears to be most predictive for older children, the social/emotional and enthusiasm factors are better predictors for young children. However, similar to the other regression results, there is not a single latent factor of engagement that exclusively predicts word learning outcomes.

The second-order interactions between age and the responsiveness suggests that, for our task, high responsiveness is the best moderator of the age effects of word learning. The interaction with age, responsiveness, and enthusiasm suggests older children with high scores of responsiveness and enthusiasm perform better than young children with similar scores. Yet, this is only true when both scores are high, otherwise older children perform worse when there is a mismatch between high/low scores for responsiveness and enthusiasm than younger children. The EFA model predicts that high enthusiasm with low responsiveness actually leads to lower word learning outcomes. The fourth-order interaction suggests that high social/emotional scores predict word learning outcomes better for younger children (regardless of scores on responsiveness and enthusiasm). A non-trivial age effect predicts older children with low enthusiasm scores but higher social/emotional and responsiveness scores will perform better overall. The exception to this is when older children have high scores for all three latent factors of engagement, suggesting that the opposing trends shown in Figure 3.11 moderate word learning performance differently for younger children versus older children.

Importantly, this methodology proves useful as it reduces dimensionality over the many possible factors of engagement but retains how each factor is related to the latent factor. Thus, there is a direct data transformation that can relate each individual measure to a more generalized factor suggested by engagement models.

4.1 Implications and recommendations for measuring engagement

Our preliminary results imply that the impact of child engagement on word learning outcomes is not simple or straightforward. There are many aspects that our two approaches to measures of engagement tried to access; however, no single measure fully predicted word learning outcomes. Instead, our results suggest that specific measures may be more accurate predictors depending on context, i.e., depending on the age or the child's response to the activity. If this pattern of results remains in a larger sample, it would suggest that the important measures of engagement for young children (from our sample 25- to 38-months-old) may be different from the important measures for older children (from our sample 38- to 46-months-old).

Furthermore, it is important to collect holistic ratings from different perspectives, because even while asking the exact same questions, holistic responses from the parent predicted word learning outcomes differently than holistic responses from the experimenter. Both were useful in predicting and child's word learning responses. There was no correlation between perspectives for ratings of activity interest ($r = 0$) and for perceived engagement ($r = -0.04$). This may be because these measures are difficult to judge, or simply because one perspective (likely the experimenter) did not have a good read of the child's interest or level of engagement. Notably, it appears that for target trials the experimenter perspective has a slightly steeper slope than the parent perspective. This suggests that the experimenter perspective may be a stronger predictor of word learning, as indicated by the steeper slope in experimenter ratings versus word learning performance compared to parent ratings (Figure 3.8).

Results from the primary regression model suggest simply, that in broad strokes, higher engagement does lead to higher performance on our word learning task; however, there are many exceptions to this rule. For example, older children may show many engagement behaviors that suggest they are actively participating, but still have low performance on the task; parent ratings may explain this discrepancy. Perhaps the child is making a game out of the task or joking around, even if they know the answer. Secondly, while parent and experimenter ratings generally correlate, these two perspectives are sensitive to different behavioral profiles. While experimenter ratings may explain discrepancies in performance based in children with similar behavioral engagement, parent ratings may be more sensitive to their child's ability to perform the task relative their baseline behavior.

The preliminary EFA model results suggest there may be a complex relationship between engagement and word learning. Enthusiasm, social/emotional engagement, and responsiveness are all factors that help explain how a child is interacting and experiencing the word learning task. Responsiveness was the largest predictor of word learning in our model. Yet, responsiveness and enthusiasm may be stronger predictors for older children, compared to younger children. And importantly, when responsiveness and enthusiasm are mismatched, word learning outcomes do not follow our expectations.

From both models, this suggests a spectrum of learning outcomes for children with different engagement profiles. Our expectations of word learning outcomes are straightforward for children who are not engaged or highly engaged, and when their behavioral and holistic ratings align: children who have many engagement behaviors and high holistic ratings are more likely to have higher word learning scores. However, outcomes vary when there is a mismatch between engagement measures. Our preliminary results suggest that these mismatches lead to unexpected outcomes, and these complex interactions should be further explored.

4.2 Study limitations

While some limitations have been mentioned above, this section will discuss limitations that may have impacted study results overall such as (1) small sample size, (2) lackluster performance on the word learning task, (3) naturalness of online learning interaction, and (4) additional considerations regarding COVID-19.

Critically, as discussed above in the Participants section and regression results section (3.2), this thesis presents preliminary data from 21 children in a study where data collection is ongoing. We do not currently have the power to discuss statistical significance of findings, and therefore focus on effect size and trends in the models tested, which will be reassessed with a larger sample.

This analysis was done in the context of a larger study whose primary manipulation was the effects of social contingency and directed social cues on the learning outcomes. Participants were randomly assigned into three conditions, which were explained in Section 2.3. While we had on average 7 children in each condition, we did not have a large enough sample size to effectively analyze the role of this critical manipulation as a parallel potential moderator of word learning outcomes.

Second, the overall performance on the word learning tasks were lackluster with respect to the label production measure. While children were responding to the filler trials, only 5 children responded to the target trials with codable responses. Similarly, gaze data suggested that children were not selecting the correct object above chance. Gaze has been show to be a good indicator of selective visual attention in young children (e.g., 16- to 24-month-olds Loi et al., 2017), yet our results do not support this. This suggests that either children were simply not performing on our task, our measure was somehow inaccurate, or our temporal window to collect responses was too large. This study is the first to our knowledge to collect gaze data in a completely remote word learning study, and therefore is a novel contribution to the field. Using the same approach as in-person word learning studies (using proportion of gaze across the full duration of the trial) did not yield useable results; however, we can use other definitions of gaze (i.e. first look, or proportion of gaze over the first half of the trial) to reanalyze word learning outcomes for this variable. Finally, we had only moderate retention of our referent selection outcome measure (65% of all participants tested).

Third, we wanted to include engagement measures from the testing phase to collect more detailed information on the child's responses. Gestures in particular, may indicate that the child is still processing the task. However, we had low reliability on our observational measures of engagement. Furthermore, our engagement scores for testing only included two gestural measures (direct gestural responses and communicative/conventional gestures) to ensure our engagement measures did not overlap with word learning measures. To look at the effects we anticipated, we would need a more nuanced approach to collecting gestural responses.

Lastly, children may not have been interested in the task or may not recognize the value or relevance of what was being taught. Many parents commented after completing the study that their child was confused if they needed to respond to the prerecorded video. We need to consider the naturalness of the learning interaction in an online setting. On one hand, the child is in their own home which they likely would be more comfortable in, and likely perform better. On the other, there are more distractions in this environment, and the expectations to learn new words via video chat may not be clear. Based on parent surveys, children often use video chat to communication with family and loved ones, and not necessarily to learn (McClure, Chentsova-Dutton, Barr, Holochwost, & Parrott, 2015). While COVID-19 has forced most classrooms online, there have been devastating effect on test scores generally (reading scores dropping 3-6 percentile points and math scores dropping 8-12 percentile points) which was exacerbated in younger students and economically disadvantaged schools (Lewis, Kuhfeld, Ruzek, & McEachin, 2021). Trends for pre-primary school age children are not established, since they are not actively participating in online classes.

4.3 Future directions

Expanding on our study design, we could consider other measures of engagement. Since gestural measures for observational engagement were restricted, it would be interesting to recode data with stricter definitions for facial expressions and shift in body position. We could also add measures for persistence (repeated attempts at answering) that could further elucidate the child's state of engagement. Perhaps other factors in the language task engagement model, like curiosity or effort, are more salient here than the measures we collected. Additionally, our observational measures did not consider gesture-speech mismatches, which may identify children in a transitional learning state and could explain the lower than expected word learning scores (Gibson et al., 2020).

Moreover, we could recast our measure of gaze to focus on the child's first look, rather than proportion of time gaze during the entire trial, to see if the child's first response was a better indicator of completing the task. This may result in another measure of word learning, which could further support the impact of engagement on word learning.

We could also analyze the effects of teaching condition, sex, and expressive vocabulary (via MCDI results) on learning outcomes. We intended to do this analysis with our data; however, demographic data (age and expressive vocabulary) was not normally distributed within our sample by teaching condition. Collecting more data, increasing our sample size, and balancing these factors for each condition will make these analyses will be possible. Thus, we are continuing to collect data for the larger study.

As an extensive moderator analysis for engagement has not been done in child word learning studies, it would be appropriate to reproduce results from an in-person study and add engagement measures to validate moderating effects within current frameworks. It is important to contextualize any finding within existing results, so the comparison between engagement in remote settings versus in-persons settings is critical to understand. Since this is one of the few fully remote word learning studies (collected in the child's home), it would be interesting to conduct the same study in-person and see if children's engagement would be different and how the remote setting would impact word learning outcomes.

4.4 Summary

This study was designed to approach word learning studies from a different perspective, through child engagement. Our results suggest that engagement may moderate the age effects of word learning performance. Engagement is difficult to measure. It also does not impact word learning outcomes in a simple way. Thus, we applied novel techniques of using EFA to reduce dimensionality of the various engagement measures used. This method forced important contextualization and scrutiny of the engagement measures collected. My hope is that these results can inspire alternative approaches to measuring engagement that could prove useful in better understanding the role of engagement in learning contexts.

Appendix A: Engagement Survey

Table A.1 Engagement Survey

	Question[1]	never	rarely	mostly	always
1.	Was your child interested in learning about the new names and objects he or she saw?	1	2	3	4
2.	How frequently did your child respond to the prompts?	1	2	3	4
3.	Was your child engaged during the study session?	1	2	3	4
4.	How frequently was your child distracted during the study session?	1	2	3	4
5.	How often did your child seem to enjoy the study session?	1	2	3	4
6.	How frequently did your child try to interact with the teacher?	1	2	3	4

[1] Engagement survey for caregiver/parent and experimenter are identical questions. The only different is for the experimenter survey ”your child” is replaced with ”the child” in each question.

Appendix B: Engagement Bids

Table B.1 Engagement bids during warm up phase

	Warm-up (Participant-Directed & Chair-Directed)	
1	"Hi there!" + *wave*	
2	"I can see you, can you see me?"	
3	"Are you excited for the game we're gonna play today?"	
4	"What's your name?"	
5	"What's your favorite color?"	
6	"I am touching my nose. Can you touch your nose too?"	
7	"Can you touch your ears like this?"	
8	"I am pointing at my mouth. Can you show me where your mouth is?"	
9	"Can you make your cheeks big like this?"	
10	"Can you show me all of your teeth, like this?"	
	Warm-up (Other-Directed)	
1	"Hi there!" + *wave*	
2	"Do you want to try charades?"	
3	"Guess which animal I am"	*Charade 1: Elephant*
4	"I want to try- ssss"	*Charade 2: Snake*
5	"Nice! Okay- tweet tweet"	*Charade 3: Bird*
6	"Now me"	*Charade 4: Monkey*
7	"Wanna see?/Wanna see a cool object?"	

Table B.2 Engagement bids during teaching phase

	Object	**Teaching**
1	Object 1	"Oh wow! Let's look at this." "Check this one out!" "Isn't this one fun?" "Isn't this one cool?" "Do you see what this one does?"
2	Object 2	"Oh wow! Let's look at this" "Look at the fopam!" "Isn't the fopam fun?" "Isn't the fopam cool?" "Do you see what the fopam does?"
3	Object 3	"Oh wow! Let's look at this" "Isn't this one cool?" "See how this one works?" "Isn't this one fun?" "Do you see what this one does?"
4	Object 4	"Oh wow! Let's look at this" "Look at the mimole!" "Isn't the mimole cool?" "Do you see what the mimole does?" "Isn't the mimole fun?"
5	Outro	"I'll see you later. Bye!" + *wave*

Appendix C: Additional Questionnaires

Additional data and demographics were collected through four questionnaires. Since linguistic contexts can dramatically change as a child gets older or attends childcare, it is important to gather information about the child's language exposure from birth to present day. To gather information on these linguistic contexts we had a Language Information Questionnaire which estimated how often a child heard different languages is various contexts. There were three periods we ask about, (a) from birth to the start of childcare (b) after starting childcare and (c) the current situation if it is significantly different from the first two phases. For each period we asked about three specific contexts (1) home, (2) childcare, and (3) other settings, like relatives, friends, or activities. For each of the languages the child was exposed to, we asked parents to rate the child's proficiency in both hearing and speaking each language.

We also collected information on the child's daily interactions to get an idea of the social settings the child encounters during a typical day, and if those interactions were with a single adult, multiple adults, or with a mixed group of peers and adults (Shneidman, Buresh, Shimpi, Knight-Schwarz, & Woodward, 2009). The Daily Interaction Questionnaire also provided information for how much time the child is in each of these settings.

Similarly, we also collected information on the child's use of various screen media since the experiment was conducted through video chat. Through the Screen Media Questionnaire we gathered information on how often the child used video chat; watched videos, movies, or television; played games on mobile devices, game consoles, or computers; used social networking; or used creative apps for activities like drawing, music, or video recording. These categories were adapted from other video chat experiments (McClure et al., 2015) and surveys on typical media usage for children (Rideout & Saphir, 2013).

Additionally, we used the full version of the MacArthur-Bates Communicative Development Inventories (MB-CDI) to gather information on the child's ability to produce and use language (Fenson et al., 2000). If the child had more than 30% exposure for both English and French, the parent was asked to complete the MB-CDI for both languages.

All questionnaires were completed by a parent of the child after providing consent for their child to participate in the study. Since the questionnaires and the data collection were asynchronous, we reviewed each questionnaire with the parent before beginning the test.

Appendix D: Label Production

For label production coding, we were concerned with a number of different things. Simply, we wanted to know if the child produced the name of the object we presented. However, we did not want to reduce responses down to an all-or-nothing coding scheme, since a child may produce a word like *fopan* instead of *fopam*. Phonetically, the novel labels we used were relatively simple, with five phonemes each. However, since we ran study in English and French, the filler object names (EN: keys, car, hat; FR: clé, voiture, touque) had different numbers of phonemes. Based on other studies, we decided to use phonological mean length utterance (pMLU) to help standardize our label production coding (Babatsouli, Ingram, & Sotiropoulos, 2014; Saaristo-Helin, Savinainen-Makkonen, & Kunnari, 2006).

We created a coding scheme with five rules (oulined in Table D.1), that incorporated many of the aspects of pMLU, and additionally gave children credit for any type of production. First each production was recorded in BORIS. Transcriptions of productions were recorded in Klattese, a keyboard friendly version of International Phonetic Alphabet. Then coders when back through each production and followed the set of rules to assign a score.

After scoring the child's production (via Equation D.1), the score was normalized against the maximum pMLU score from the target word. This provided a standardized proportion

Table D.1 Rules for label production coding

Label production rules

(1) **Triage rule:** If the production that the child says is scorable if it is in response to the target or if you are not sure. If the production the child says is clearly something else, then it is unscorable. If the response is unscorable, stop coding and assign and overall score of zero.

(2) **Variability rule:** If the child produces more than one relevant label, we'll take the last word that the child says.

Label production scores:

(3) **Maximum phonemes rule:** A point is given for every phoneme the child says unless the number of phonemes said exceeds those in the target. If the number of phonemes said exceeds the target, then the points received will equal to the number of phonemes in the target.

(4) **Consonants from target rule:** A point is given for every consonants said, if that consonant is in the correct order.

(5) **Extra phonemes rule:** A point is given if the child set greater/fewer phonemes than the target. If the child said something with the correct number of phonemes, they get a zero points. This score is subtracted from the overall score, and acts as a penalty for additional production.

score that can be used to compare trials.

$$y_{\text{label production}} = \frac{x_{\text{max phonemes}} + x_{\text{consonants from target}} - x_{\text{extra phonemes}}}{x_{\text{target}}} \tag{D.1}$$

A team of three coders coducted the label production coding. Overall, there was 70.9% agreement between coders in their transcriptions. There were only 4 children who received scores of 0 on all label production trials, so children were responding to the trials. Unfortunately, only 5 children produced labels on the target object trials, which left a very small sample size to compare performance between filler and target trials. For this reason, we decided to exclude the label production data from our analysis.

References

Adamson, L. B., Bakeman, R., Deckner, D. F., & Nelson, P. B. (2013). From Interactions to Conversations: The Development of Joint Engagement During Early Childhood. *Child Development*. doi: 10.1111/cdev.12189

Adamson, L. B., & Frick, J. E. (2003). The Still Face: A History of a Shared Experimental Paradigm. *INFANCY*, *4*(4), 451–473.

Aguiar, C., & McWilliam, R. A. (2013, jan). Consistency of toddler engagement across two settings. *Early Childhood Research Quarterly*, *28*(1), 102–110. doi: 10.1016/J.ECRESQ.2012.04.003

Allen, R., & Scofield, J. (2010, nov). Word learning from videos: more evidence from 2-year-olds. *Infant and Child Development*, *19*(6), 649–661. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1002/icd.712 doi: 10.1002/ICD.712

Appleton, J. J., Christenson, S. L., & Furlong, M. J. (2008, may). Student engagement with school: Critical conceptual and methodological issues of the construct. *Psychology in the Schools*, *45*(5), 369–386. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1002/pits.20303 doi: 10.1002/PITS.20303

Babatsouli, E., Ingram, D., & Sotiropoulos, D. A. (2014). Phonological word proximity in child speech development. *Chaotic Modeling and Simulation*(3), 295–313.

Baldwin, D. A. (1994). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology*, *29*(5), 832. Retrieved from /fulltext/1994-08978-001.html doi: 10.1037/0012-1649.29.5.832

Bang, J. (2017). *The role of intention in reading referential gaze: Implications for learning in typical development and in Autism Spectrum Disorder* (Doctoral dissertation, McGill University). doi: 10.13140/RG.2.2.35725.13280

Bannard, C., & Tomasello, M. (2012, nov). Can We Dissociate Contingency Learning from Social Learning in Word Acquisition by 24-Month-Olds? *PLOS ONE*, *7*(11), e49881. Retrieved from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0049881 doi: 10.1371/JOURNAL.PONE.0049881

Baron, R. M., & Kenny, D. A. (1986, 12). The moderator-mediator variable distinction in social psychological research. conceptual, strategic, and statistical considerations.

Journal of Personality and Social Psychology, *51*, 1173-1182. Retrieved from https://psycnet-apa-org.proxy3.library.mcgill.ca/journals/psp/51/6/1173 doi: 10.1037/0022-3514.51.6.1173

Bertin, E., & Striano, T. (2006, apr). The still-face response in newborn, 1.5-, and 3-month-old infants. *Infant Behavior and Development*, *29*(2), 294–297. doi: 10.1016/J.INFBEH.2005.12.003

Bloom, K., Russell, A., & Wassenberg, K. (1987). Turn taking affects the quality of infant vocalizations*. *Journal of Child Language*, *14*(2), 211–227. Retrieved from https://www.cambridge.org/core/journals/journal-of-child-language/article/turn-taking-affects-the-quality-of-infant-vocalizations/A09C792BE5C77776DE7422B23CA85AE6 doi: 10.1017/S0305000900012897

Busselle, R., & Bilandzic, H. (2009, nov). Measuring Narrative Engagement. *Media Psychology*, *12*(4), 321–347. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/15213260903287259 doi: 10.1080/15213260903287259

Butterworth, G. (2003). Pointing is the royal road to language for babies. In *Pointing: Where language, culture, and cognition meet.* (pp. 9–33). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social Cognition, Joint Attention, and Communicative Competence from 9 to 15 Months of Age. *Monographs of the Society for Research in Child Development*, *63*(4), i. doi: 10.2307/1166214

Church, R. B., & Goldin-Meadow, S. (1986). The mismatch between gesture and speech as an index of transitional knowledge. *Cognition*, *23*(1), 43–71. Retrieved from https://pubmed.ncbi.nlm.nih.gov/3742990/ doi: 10.1016/0010-0277(86)90053-3

Cisco Systems, I. (2021). *Cisco Webex.* San Jose: Cisco Systems, Inc. Retrieved from https://www.webex.com/

Colonnesi, C., Stams, G. J. J., Koster, I., & Noom, M. J. (2010, dec). The relation between pointing and language development: A meta-analysis. *Developmental Review*, *30*(4), 352–366. doi: 10.1016/J.DR.2010.10.001

Courtney, M., & Gordon, R. (2013, 11). Determining the number of factors to retain in efa: Using the spss r-menu v2 0 to make more judicious estimations. *Practical Assessment, Research, and Evaluation*, *18*, 8. Retrieved from https://scholarworks.umass.edu/pare/vol18/iss1/8 doi: https://doi.org/10.7275/9cf5-2m72

Cumming, G. (2014, 11). The new statistics: Why and how. *Psychological Science*, *25*, 7-29. Retrieved from https://journals.sagepub.com/doi/10.1177/0956797613504966 doi: 10.1177/0956797613504966

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. doi: 10.3758/S13428-014-0458-Y

Egbert, J. (2020a). Engagement, Technology, and Language Tasks: Optimizing Student

Learning. *Article International Journal of TESOL Studies*, *2*(4), 110. doi: 10.46451/ijts.2020.12.10

Egbert, J. (2020b). The new normal?: A pandemic of task engagement in language learning. *Foreign Language Annals*, *53*, 314–319. Retrieved from https://newsela. doi: 10.1111/flan.12452

Egbert, J., Shahrokni, S. A., Abobaker, R., & Borysenko, N. (2021). "It's a chance to make mistakes": Processes and outcomes of coding in 2nd grade classrooms. *Computers & Education*, *168*, 104173. Retrieved from https://doi.org/10.1016/j.compedu.2021.104173 doi: 10.1016/j.compedu.2021.104173

Ekman, P., & Rosenberg, E. L. (2012, mar). What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS). *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*, 1–672. doi: 10.1093/ACPROF:OSO/9780195179644.001.0001

Fenson, L., Pethick, S., Renda, C., Cox, J. L., Dale, P. S., & REZNICK, J. S. (2000). Short-form versions of the MacArthur Communicative Development Inventories. *Applied Psycholinguistics*, *21*(1), 95–116. Retrieved from https://www.cambridge.org/core/journals/applied-psycholinguistics/article/shortform-versions-of-the-macarthur-communicative-development-inventories/32AEF9ECC5532BDFD53C2045F0C6C68B doi: 10.1017/S0142716400001053

Fernald, A., Perfors, A., & Marchman, V. A. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. *Developmental Psychology*, *42*(1), 98–116. doi: 10.1037/0012-1649.42.1.98

Finn, J. D., & Zimmer, K. S. (2012, jan). Student Engagement: What Is It? Why Does It Matter? *Handbook of Research on Student Engagement*, 97–131. Retrieved from https://link.springer.com/chapter/10.1007/978-1-4614-2018-7_5 doi: 10.1007/978-1-4614-2018-7_5

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59–109. doi: 10.3102/00346543074001059

Friard, O., & Gamba, M. (2016, nov). BORIS: a free, versatile open-source event-logging software for video/audio coding and live observations. *Methods in Ecology and Evolution*, *7*(11), 1325–1330. Retrieved from https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.12584 doi: 10.1111/2041-210X.12584

Gibson, D. J., Gunderson, E. A., Spaepen, E., & Goldin-meadow, S. (2020). Number gestures predict learning of number words. , *22*(3), 1–29. Retrieved from https://cpb-us-w2.wpmucdn.com/voices.uchicago.edu/dist/c/1286/files/2017/01/nihms-1002742.pdf doi: 10.1111/desc.12791.Number

Goldin-Meadow, S. (2020). Approaching Learning Hands First How Gesture Influences Thought. In *A multidisciplinary approach to embodiment* (pp. 46–50).

Routledge. Retrieved from https://cpb-us-w2.wpmucdn.com/voices.uchicago.edu/dist/c/1286/files/2017/01/8.-Chapter-7.pdf

Grassmann, S. (2014). The pragmatics of word learning. In D. Matthews (Ed.), *Pragmatic development in first language acquisition* (pp. 139–160). Amsterdam: John Benjamins Publishing Company. doi: 10.1075/tilar.10.09gra

Hirsh-Pasek, K., Zosh, J. M., Golinkoff, R. M., Gray, J. H., Robb, M. B., & Kaufman, J. (2015, apr). Putting Education in "Educational" Apps: Lessons From the Science of Learning. *Psychological Science in the Public Interest*, *16*(1), 3–34. Retrieved from https://journals.sagepub.com/doi/10.1177/1529100615569721 doi: 10.1177/1529100615569721

Kirkorian, H. L., Choi, K., & Pempek, T. A. (2016, mar). Toddlers' Word Learning From Contingent and Noncontingent Video on Touch Screens. *Child Development*, *87*(2), 405–413. Retrieved from https://srcd.onlinelibrary.wiley.com/doi/full/10.1111/cdev.12508 doi: 10.1111/CDEV.12508

Lange, K., Kühn, S., & Filevich, E. (2015, jun). "Just Another Tool for Online Studies" (JATOS): An Easy Solution for Setup and Management of Web Servers Supporting Online Studies. *PLOS ONE*, *10*(6), e0130834. Retrieved from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130834 doi: 10.1371/JOURNAL.PONE.0130834

Lauricella, A. R., Blackwell, C. K., & Wartella, E. (2017, jan). The "New" Technology Environment: The Role of Content and Context on Learning and Development from Mobile Media. *Media Exposure During Infancy and Early Childhood: The Effects of Content and Context on Learning and Development*, 1–23. Retrieved from https://link.springer.com/chapter/10.1007/978-3-319-45102-2_1 doi: 10.1007/978-3-319-45102-2_1

Lewis, K., Kuhfeld, M., Ruzek, E., & McEachin, A. (2021). *Learning during COVID-19: Reading and math acheivement in the 2020-2021 school year* (Tech. Rep.). Portland, OR: NWEA and Center for School and Student Progress.

LimeSurvey Project Team, & Schmitz, C. (2012). *LimeSurvey: An open source survey tool.* Hamburg, Germany. Retrieved from www.limesurvey.org

Loi, E. C., Marchman, V. A., Fernald, A., & Feldman, H. M. (2017). Using Eye Movements to Assess Language Comprehension in Toddlers Born Preterm and Full Term. *Journal of Pediatrics*, *180*, 124–129. Retrieved from http://dx.doi.org/10.1016/j.jpeds.2016.10.004 doi: 10.1016/j.jpeds.2016.10.004

Marchman, V. A., Ashland, M. D., Loi, E. C., Adams, K. A., Fernald, A., & Feldman, H. M. (2019). Predictors of early vocabulary growth in children born preterm and full term: A study of processing speed and medical complications. *Child Neuropsychology*, *25*(7), 943–963. Retrieved from https://doi.org/10.1080/09297049.2019.1569608 doi: 10.1080/09297049.2019.1569608

Masataka, N. (1993). Effects of contingent and noncontingent maternal stimulation on the vocal behaviour of three- to four-month-old Japanese infants*. *Journal of Child Language*, *20*(2), 303–312. Retrieved from https://www.cambridge.org/core/journals/journal-of-child-language/article/effects-of-contingent-and-noncontingent-maternal-stimulation-on-the-vocal-behaviour-of-three-to-fourmonthold-japanese-infants/986BA562D1FE9EF5446903B2E4375D89 doi: 10.1017/S0305000900008291

Matthews, D., Behne, T., Lieven, E., & Tomasello, M. (2012, nov). Origins of the human pointing gesture: a training study. *Developmental Science*, *15*(6), 817–829. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-7687.2012.01181.x doi: 10.1111/J.1467-7687.2012.01181.X

McClure, E. R., Chentsova-Dutton, Y., Holochwost, S., Parrott, W. G., & Barr, R. (2020). Infant emotional engagement in face-to-face and video chat interactions with their mothers. *Enfance*, *N°3*(3), 353. Retrieved from https://doi.org/10.3917/enf2.203.0353 doi: 10.3917/enf2.203.0353

McClure, E. R., Chentsova-Dutton, Y. E., Barr, R. F., Holochwost, S. J., & Parrott, W. G. (2015, dec). "Facetime doesn't count": Video chat as an exception to media restrictions for infants and toddlers. *International Journal of Child-Computer Interaction*, *6*, 1–6. doi: 10.1016/J.IJCCI.2016.02.002

McGillion, M., Herbert, J. S., Pine, J., Vihman, M., DePaolis, R., Keren-Portnoy, T., & Matthews, D. (2017, jan). What Paves the Way to Conventional Language? The Predictive Value of Babble, Pointing, and Socioeconomic Status. *Child Development*, *88*(1), 156–166. Retrieved from https://srcd.onlinelibrary.wiley.com/doi/full/10.1111/cdev.12671 doi: 10.1111/CDEV.12671

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276. Retrieved from /pmc/articles/PMC3900052//pmc/articles/PMC3900052/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

Mesman, J., van IJzendoorn, M. H., & Bakermans-Kranenburg, M. J. (2009, jun). The many faces of the still-face paradigm: A review and meta-analysis. *Developmental Review*, *29*(2), 120–162. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0273229709000021 doi: 10.1016/j.dr.2009.02.001

Mundy, P., Block, J., Delgado, C., Pomares, Y., Hecke, A. V. V., & Parlade, M. V. (2007, may). Individual Differences and the Development of Joint Attention in Infancy. *Child Development*, *78*(3), 938–954. Retrieved from https://srcd.onlinelibrary.wiley.com/doi/full/10.1111/j.1467-8624.2007.01042.x doi: 10.1111/J.1467-8624.2007.01042.X

Myers, L. J., Crawford, E., Murphy, C., Aka-Ezoua, E., & Felix, C. (2018, jul). Eyes in the room trump eyes on the screen: effects of a responsive co-viewer on toddlers' responses to and learning from video chat. *Journal of Children and Media*, *12*(3), 275–294. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/17482798.2018.1425889 doi: 10.1080/17482798.2018.1425889

Myers, L. J., Keyser, H., & Cors, M. (2019, oct). Co-viewers support participation in video chat interactions, but live experiences promote richer word learning for 24- to 36-

month-olds in the USA. *Journal of Children and Media*, *13*(4), 415–432. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/17482798.2019.1646294 doi: 10.1080/17482798.2019.1646294

Myers, L. J., LeWitt, R. B., Gallo, R. E., & Maselli, N. M. (2017, jul). Baby FaceTime: can toddlers learn from online video chat? *Developmental Science*, *20*(4), e12430. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1111/desc.12430 doi: 10.1111/DESC.12430

Nielsen, M., Simcock, G., & Jenkins, L. (2008, sep). The effect of social engagement on 24-month-olds' imitation from live and televised models. *Developmental Science*, *11*(5), 722–731. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-7687.2008.00722.x doi: 10.1111/J.1467-7687.2008.00722.X

O'Doherty, K., Troseth, G. L., Shimpi, P. M., Goldenberg, E., Akhtar, N., & Saylor, M. M. (2011, may). Third-Party Social Interaction and Word Learning From Video. *Child Development*, *82*(3), 902–915. Retrieved from https://srcd.onlinelibrary.wiley.com/doi/full/10.1111/j.1467-8624.2011.01579.x doi: 10.1111/J.1467-8624.2011.01579.X

Pel, J. J., Manders, J. C., & van der Steen, J. (2010). Assessment of visual orienting behaviour in young children using remote eye tracking: Methodology and reliability. *Journal of Neuroscience Methods*, *189*(2), 252–256. Retrieved from http://dx.doi.org/10.1016/j.jneumeth.2010.04.005 doi: 10.1016/j.jneumeth.2010.04.005

Peterson, R. A. (2000). A meta-analysis of variance accounted for and factor loadings in exploratory factor analysis. *Marketing Letters 2000 11:3*, *11*, 261-275. Retrieved from https://link.springer.com/article/10.1023/A:1008191211004 doi: 10.1023/A:1008191211004

Pine, K. J., Lufkin, N., & Messer, D. (2004, nov). More gestures than answers: Children learning about balance. *Developmental Psychology*, *40*(6), 1059–1067. doi: 10.1037/0012-1649.40.6.1059

R Core Team. (2020). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org

Revelle, W. (2021, jun). *psych: Procedures for Psychological, Psychometric, and Personality Research [R package psych version 2.1.6].* Evanstson, Illinois, USA: Comprehensive R Archive Network (CRAN). Retrieved from https://cran.r-project.org/package=psych

Richardson, D. C., Griffin, N. K., Zaki, L., Stephenson, A., Yan, J., Curry, T., ... Devlin, J. T. (2020, jul). Engagement in video and audio narratives: contrasting self-report and physiological measures. *Scientific Reports*, *10*(1), 1–8. Retrieved from https://www.nature.com/articles/s41598-020-68253-2 doi: 10.1038/s41598-020-68253-2

Rideout, V., & Saphir, M. (2013). *Zero to Eight: Children's Media Use in America 2013* (Tech. Rep.). Retrieved from https://www.commonsensemedia.org/research/zero-to-eight-childrens-media-use-in-america-2013

Roseberry, S., Hirsh-Pasek, K., & Golinkoff, R. M. (2014, may). Skype Me! Socially

Contingent Interactions Help Toddlers Learn Language. *Child Development*, *85*(3), 956–970. Retrieved from https://srcd.onlinelibrary.wiley.com/doi/full/10.1111/cdev.12166 doi: 10.1111/CDEV.12166

Russo-Johnson, C., Troseth, G., Duncan, C., & Mesghina, A. (2017, apr). All Tapped Out: Touchscreen Interactivity and Young Children's Word Learning. *Frontiers in Psychology*, *0*(APR), 578. doi: 10.3389/FPSYG.2017.00578

Saaristo-Helin, K., Savinainen-Makkonen, T., & Kunnari, S. (2006, feb). The Phonological Mean Length of Utterance: methodological challenges from a crosslinguistic perspective. *Journal of Child Language*, *33*(1), 179–190. Retrieved from https://www.cambridge.org/core/journals/journal-of-child-language/article/abs/phonological-mean-length-of-utterance-methodological-challenges-from-a-crosslinguistic-perspective/1F9A18D466803009D155D08B609F34AC doi: 10.1017/S0305000905007294

Shneidman, L. A., Buresh, J. S., Shimpi, P. M., Knight-Schwarz, J., & Woodward, A. L. (2009, sep). Social Experience, Social Attention and Word Learning in an Overhearing Paradigm. *Language Learning and Development*, *5*(4), 266–281. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/15475440903001115 doi: 10.1080/15475440903001115

Smith, L. B., & Yu, C. (2013). Visual Attention Is Not Enough: Individual Differences in Statistical Word-Referent Learning in Infants. *Language Learning and Development*, *9*(1), 25–49. doi: 10.1080/15475441.2012.707104

Soussignan, R., Nadel, J., Canet, P., & Gerardin, P. (2006, sep). Sensitivity to Social Contingency and Positive Emotion in 2-Month-Olds. *Infancy*, *10*(2), 123–144. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1207/s15327078in1002_2 doi: 10.1207/S15327078IN1002_2

Striano, T., & Stahl, D. (2005, jul). Sensitivity to triadic attention in early infancy. *Developmental Science*, *8*(4), 333–343. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-7687.2005.00421.x doi: 10.1111/J.1467-7687.2005.00421.X

Strouse, G. A., Troseth, G. L., O'Doherty, K. D., & Saylor, M. M. (2018, feb). Co-viewing supports toddlers' word learning from contingent and noncontingent video. *Journal of Experimental Child Psychology*, *166*, 310–326. doi: 10.1016/J.JECP.2017.09.005

Suhr, D. (2005). Principal component analysis vs. exploratory factor analysis. In *Sugi 30 proceedings* (Vol. 1, pp. 69–95). Philadelphia: SAS Users Group International. Retrieved from https://support.sas.com/en/papers/proceedings-archive/sugi2005.html

Thijs, J., & Verkuyten, M. (2009, sep). Students' Anticipated Situational Engagement: The Roles of Teacher Behavior, Personal Engagement, and Gender. *The Journal of Genetic Psychology*, *170*(3), 268–286. Retrieved from https://doi.org/10.1080/00221320903218323 doi: 10.1080/00221320903218323

Toda, S. (1993). Infant response to the still-face situation at 3 and 6 months. *Developmental Psychology*, *29*(3), 532. Retrieved from /fulltext/1993-33004-001.html doi: 10.1037/0012-1649.29.3.532

Troseth, G. L., Saylor, M. M., & Archer, A. H. (2006, may). Young Children's Use of Video as a Source of Socially Relevant Information. *Child Development*, *77*(3), 786–799. Retrieved from https://srcd.onlinelibrary.wiley.com/doi/full/10.1111/j.1467-8624.2006.00903.x doi: 10.1111/J.1467-8624.2006.00903.X

Troseth, G. L., Strouse, G. A., Verdine, B. N., & Saylor, M. M. (2018). Let's chat: On-screen social responsiveness is not sufficient to support toddlers' word learning from video. *Frontiers in Psychology*, *9*(NOV), 1–10. doi: 10.3389/fpsyg.2018.02195

Wilson VanVoorhis, C. R., & Morgan, B. L. (2007, 9). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology*, *3*, 43-50. doi: 10.20982/TQMP.03.2.P043

Yazzie-Mintz, E. (2007). Voices of students on engagement: A report on the 2006 high school survey of student engagement. *Center for Evaluation and Education Policy, Indiana University*, 12. Retrieved from http://ezproxy.lib.utexas.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED495758&site=ehost-live