Caregivers use gesture contingently to support word learning

Rachael W. Cheung[1], Calum Hartley[1], and Padraic Monaghan[1,2]

[1]Lancaster University, UK

[2]University of Amsterdam, The Netherlands

Conflict of Interest Statement: No conflicts of interest exist regarding this work for any of the authors.

Data Availability Statement: Data for this submission is available on the Open Science Framework (https://osf.io/6frcw/?view_only=72344789a6294aa19d63a8bd93a628f3).

Acknowledgements: This work was supported by the Leverhulme Trust (RWC, Leverhulme Trust Doctoral Scholar [DS-2014-014]) and the International Centre for Language and Communicative Development (LuCiD) at Lancaster University funded by the Economic and Social Research Council UK (PM and CH) [ES/L008955/1]. Our thanks also to Delyth Piper for aiding video coding.

Corresponding author: Rachael W Cheung, Department of Psychology, Fylde College, Lancaster University, Lancaster, LA1 4YF, UK

Title: Caregivers use gesture contingently to support word learning

Research Highlights

- We tested word learning with varying referential ambiguity in a computational model and experimental study of caregiver-child dyads (18–24-month-olds)
- The model predicted gesture would be more useful with more referential ambiguity, and that learning might be more robust in the presence of variability
- The behavioural study confirmed that the use of deictic gesture by caregivers increased in the presence of referential ambiguity
- Both the computational model and child participants learnt best according to the presence, rather than degree, of referential ambiguity

Abstract

Children learn words in environments where there is considerable variability, both in terms of the number of possible referents for novel words, and the availability of cues to support word-referent mappings. How caregivers adapt their gestural cues to referential uncertainty has not yet been explored. We tested a computational model of cross-situational word learning that examined the value of a variable gesture cue during training across conditions of varying referential uncertainty. We found that gesture had a greater benefit for referential uncertainty, but unexpectedly also found that learning was best when there was variability in both the environment (number of referents) and gestural cue use. We demonstrated that these results are reflected behaviourally in an experimental word learning study involving children aged 18-24-month-olds and their caregivers. Under similar conditions to the computational model, caregivers not only used gesture more when there were more potential referents for novel words, but children also learned best when there was some referential ambiguity for words. Thus, caregivers are sensitive to referential uncertainty in the environment and adapt their gestures accordingly, and children are able to respond to environmental variability to learn more robustly. These results imply that training under variable circumstances may actually benefit learning, rather than hinder it.

Keywords: gestures, word learning, variability, child language acquisition, computational modelling

Introduction

Word learning is a complex process, requiring children to individuate words from continuous speech and pair them with intended referents in the environment. However, there are multiple possible references within multiword utterances (Monaghan & Mattock, 2012; Yu & Ballard, 2007) and multiple potential referents in the environment for each word (Quine, 1960; Siskind, 1996; L.B. Smith & Yu, 2008). Although internal constraints may aid special cases of language acquisition (Carey, 1988; Golinkoff et al., 1992; Markman & Wachtel, 1988; Mervis, 1987), alternative accounts have explored how constraints present in the environment can be utilised by more general purpose learning mechanisms.

The environment contains multiple sources of information that can help to constrain word-object mappings. This includes cross-situational statistics, where possible links between words and referents may be resolved by tracking co-occurrences between them across multiple situations (Siskind, 1996; L.B. Smith & Yu, 2008). Other cues include prosody, such as the referring word having the highest amplitude (Fernald & Mazzie, 1991), and distributional information from syntax, such as nouns and verbs being preceded by frequently-occurring articles (Fries, 1952; Mintz, 2003; Monaghan et al., 2007). Gestural cues also contribute vital information, forming an integral part of communication from early infancy (Iverson & Goldin-Meadow, 2005; Southgate et al., 2007), and helping caregivers delineate referents during word learning (Cartmill et al., 2013; Iverson et al., 1999).

Despite huge environmental variation across learning situations, word learning studies generally assume a relatively stable environment for children (McMurray et al., 2012; Yu et al., 2012). Importantly, this variability may actually be useful. In a computational model of word learning, Monaghan (2017) developed the multimodal integration model (MIM; A.C. Smith et al. 2017) to explore the role of multiple cues – distributional, prosodic, and gestural – in supporting language acquisition. The model was trained to learn word-object pairings when words and objects were presented among multiple possibilities and when cues were present or absent. Although learning benefited from all cues, learning was more efficient and more accurate when cues occurred 75% of the time, rather than when they were present 100% of the time (Monaghan, 2017). This was confirmed in behavioural studies with adults (Monaghan, Brand, Frost, & Taylor, 2017). The MIM showed that multiple cues support learning over single cues, and that the model learnt most robustly when the cues were individually variable. This prevented the model from relying too heavily on single cues in the environment, akin to dropout training, in which input units are stochastically dropped to improve model generalisation and avoid overfitting (Srivastava et al., 2014). In the MIM, the existence of variability within the environment itself circumvents the requirement for this to be incorporated into the learner, providing the necessary degree of dropout to maintain the learner's sensitivity to multiple cues in the environment. These results indicate that although word learning occurs in noisy contexts with multiple, variable cues, learners are able to make use of this variability to benefit learning.

However, the MIM did not test the extent to which variability in cues may be contingent on the informational content of situations. For instance, when there is only one possible referent in the environment, gesture may be redundant. Alternatively, when there are many possible referents, gesture may be crucial. Thus, during learning situations, if the speaker is sensitive to this environmental ambiguity, we may see cues deployed differently according to the situation.

Speakers adjust their prosody, syntax, word selection, and phonology according to context and the listener's perspective (Brown-Schmidt & Duff, 2016; Gorman et al., 2013), and children also adapt speech and gesture according to the perspective of adults (Bahtiyar & Küntay, 2009; Bannard et al., 2017; Nadig & Sedivy, 2002; Nilsen & Graham, 2009; O'Neill, 1996). In contrast, how caregivers adapt to the environment is less established. Caregivers demonstrate patterns of behaviour when labelling objects that align with children's internal constraints, such as naming whole objects rather than parts (Masur, 1997), or using one label per object, encouraging mutually exclusive labelling (Callanan & Sabbagh, 2004). Caregivers also adjust how they use labels according to their child's knowledge (Luce & Callanan, 2010; Masur, 1997); for example, by placing unfamiliar nouns and verbs saliently in an utterance and physically presenting unfamiliar objects more clearly (Cleave & Bird, 2006). However, these adaptations depend on perceived levels of familiarity in the child, rather than perceived uncertainty in the environment when the level of familiarity is consistent (such as when all objects are novel). These studies show that caregivers are sensitive to the informational content of cues relative to their child, but whether this sensitivity exists when environmental variability itself is manipulated has not yet been tested.

Gesture offers a prime candidate for further exploration of how caregivers might adapt contingently during word learning. Not only is gesture facilitative of vocabulary development, with increased early child gesture use predicting larger future vocabulary size (Brooks & Meltzoff, 2008; Fenson et al., 1994; Kuhn et al., 2014), but caregiver gesture use can predict early child gesture use (Rowe et al., 2008) and offer highly valuable information for word-referent mapping (Cartmill et al. 2013). Caregivers also alter gestures according to whether an object is familiar to their child as well as present or absent (Vigliocco et al., 2019), and in response to increased task complexity when communicating with children with delayed language development (Wray & Norbury, 2018).

The types of gestures produced by caregivers and children are rich and varied (Capone & McGregor, 2004; Özçalışkan & Dimitrova, 2013). They may occur in isolation or combined with speech, providing information that may overlap, complement, or even mismatch speech content – all of which offer valuable communicative insight (Goldin-Meadow & Wagner, 2005). Yet, when faced with high referential ambiguity during word learning, the most informative caregiver gestures may be those that clearly delineate the target of a novel label. Children follow deictic gestures such as pointing from approximately 12-months-old (Carpenter et al., 1998), and caregivers also use deictic gestures more than other gestures with children under 22-months-old (Özçalişkan & Goldin-Meadow, 2005). Whether caregivers alter these useful gestures based on the presence of environmental referential ambiguity remains unexplored.

In this paper, we examined how environmental variability might affect word learning by testing the contingency of caregiver gesture use to support word learning under referential uncertainty. We first adapted an established computational model of word learning (MIM; Monaghan, 2017) to test the benefit of contingent gestural cues for word learning when the number of possible referents for speech varies. We then conducted a behavioural study to determine whether caregivers varied in their gesture use when teaching novel words under different degrees of referential uncertainty, and whether the predictions of the computational model for optimal behaviour are exhibited in naturalistic exchanges. We thus considered the presence and interaction of two distinct aspects of variability: *referential uncertainty*, conferred by differing numbers of potential referents for a given word, and the *availability of gestural cues*, with their role determined firstly by altering the occurrence of such cues systematically in a computational model, and then by examination of naturally-occurring differences in caregiver cue use during a behavioural study.

Computational model

We adapted Monaghan's (2017) implementation of the MIM by varying the number of possible referents in the visual field during training to test the effect of environmental indeterminacy on cue influence. Monaghan's (2017) implementation is an adaptation of A.C. Smith et al. (2017), and simulates word learning via acquiring the correspondence between one of several words heard in an utterance and one of several objects in the environment. The model is a neural network that learns through backpropagation, operating on principles of acquiring associations between representations. The MIM is similar in principle to other associative models of word learning (e.g. McMurray et al., 2012; Yu & L.B. Smith, 2012), but extends these to test multiple cues in the child's immediate environment that provide information about the intended reference of speech. Our aim in this paper is to examine how such a simple associative learning system might respond to variation in environmental cues in terms of how associations between words and objects cohere.

We trained and tested the MIM (Monaghan, 2017) under three conditions that allowed us to investigate the effects of a gestural cue on learning during: 1) a condition with no referential ambiguity, where the object presented must be the target (one object); 2) a condition with some referential uncertainty, where one object was the target and one was the foil (two objects); and 3) a condition with a higher degree of referential uncertainty, where one object was the target and there were five foils (six objects). Enumeration tasks suggest that observers are able to rapidly report the numbers of objects in a visual display between one to four objects with ease; however, above four, they switch to slow counting of individual objects (Cowan, 2001; Xu & Chun, 2009). Thus, our aim was to crowd the visual display in the six-object condition.

An increase in potential referents for a given novel word has led to less reliable learning in behavioural studies (K. Smith et al., 2011; Trueswell et al., 2013). We therefore predicted that the model would learn more quickly from the one-object than the two-object condition, which in turn would be learned more quickly than the six-object condition. We also predicted that the effect of the gestural cue would be largest when there were two objects compared to one, and six objects compared to two: as indeterminacy of the intended referent increases, gesture may become more important to support and constrain word-referent mappings.

Method

Architecture

The model's architecture is shown in Figure 1. The model had an auditory input, comprising 80 units, where sets of spoken words were presented, and an 80-unit visual input, where sets of objects were presented. Each unit in the auditory and visual inputs was capable of representing one piece of information (i.e. a phoneme feature within a word, or a visual feature of an object). Input from these auditory and visual inputs projected to a central integrative layer of 100 units, each of which combined and processed input from the set of auditory and visual inputs. This integrative layer was self-connected, and was also connected to a semantic output layer comprising 100 units, where the model had to generate the meaning representation of the target word-object pairing.

For the current simulation, we expanded the number of objects that could appear in the visual input from two (as in the original simulation; Monaghan, 2017) to six. For the one-object condition, the object could appear in any of the six possible object locations. For the two-object condition, any two of the six locations presented the objects. For the six-object condition, one object appeared in each of the six locations. The model was otherwise identical to the original simulations.

Representations

The auditory, visual, and semantic representations for each word-object mapping were identical to Monaghan (2017).

When the gestural cue was present, the activation of the target object's location was doubled, enhancing the influence that the visual features of the object in that position had on the model's learning. The role of gesture was thus implemented as increasing the salience of one position in the visual display of the model, and the effect of gesture is akin to increasing attention to a region of visual space, as implemented of visual processing in dynamic systems models (Samuelson et al., 2017). Across simulation runs, we varied the availability of the gestural cue by altering its presence across individual trials, where the cue was present 0%, 33%, 67% or 100% of the time. For example, in the 33% gesture cue availability condition, there was a 1/3 chance for each trial that the cue was present.

For each simulation, there were 100 word-object mappings to be learned, with the auditory and visual representation of each word-object mapping randomly generated for each simulation run.

Training

The model was trained to learn correspondences between 100 spoken words and 100 visual objects through cross-situational statistics.

For each training trial, the model was presented with two auditory words – one corresponded to a visual object appearing in the visual input, and the other was randomly selected from the other 99 words. The model was required to produce the semantic representation corresponding to the overlap between the target word and target object at the output.

For the one-object condition, only the target object corresponding to one of the spoken words was presented. For the two-object condition, two objects were presented – one corresponding to one of the spoken words and the other randomly selected from the other 99 objects (but not corresponding to the other, foil word). For the six-object condition, five foil objects were selected. In all conditions the positions of objects were randomised. For the one-object condition, the target object appeared in one of the six locations, and the other five locations were empty. For the two-object condition, the target and a foil object appeared in random locations across the six possible positions. For the six-object condition, the target object appeared randomly in one location, and five other foil objects filled the five remaining locations. The gestural cue was present for either 0%, 33%, 67%, or 100% of the individual trials in each condition.

Activation in the model passed between layers for five time steps. At time 1, the auditory and visual input was presented to the model. At time 2, the activation from these input layers reached the integrative layer. At time steps 3 to 5, the model was required to produce the semantic representation for the word-object pairing, with recurrent activation cycling through the integrative layer's self-connections and from the integrative layer to the semantic output layer. At the end of each training trial, the model's error was calculated across the semantic output layer as the cross-entropy error of the difference between the model's actual activation of units and the target activations. Connections were adjusted between units in the model according to the backpropagation through time learning algorithm (Pearlmutter, 1989). The model's connections were initially randomised in the range [-0.1, 0.1], and the learning rate was set at 0.01.

After 1000 learning trials had been presented to the model, its performance on each of the 100 word-object mappings was tested. The model was judged to be accurate if it produced a semantic representation closer to the target than to any of the other 99 semantic representations. The point in training at which the model was able to identify 95% of the word-object mappings correctly in four consecutive tests was identified as reflecting the ease of the model's ability to learn the words. If the model failed to learn by the end of training, then the end of training was taken to be the length of training time. Training finished after 100,000 learning trials had been presented to the model, and then the model was tested.

We formulated 10 different versions of the training patterns. For each training pattern, we ran 12 different versions of the model, with different randomised starting weights, different gesture cue availability, and a different number of objects during training. In total, there were 120 simulation runs: 10 versions of pattern x 4 gesture cue availability (0%, 33%, 67%, and 100%) x 3 numbers of objects (1, 2, and 6). We treated each of the 10 different versions of the training patterns as a separate subject during analysis, and treated gesture cue availability and number of objects as within-subject variables.

Testing

The model's ability to accurately detect the word-object mapping for each of the 100 pairings was tested under different conditions than its training: the model was tested instead where the target object appeared along with two other foil objects (simulating a three-alternative forced choice test). To assess the robustness of learning, we also determined whether the model could identify the target pairing without any gestural cue being present. The model's accuracy was determined in the same way as during training: if it produced a semantic representation closer to the target than to any of the other 99 semantic representations.

Data, code, and models run are available on the Open Science Framework (OSF) (http://osf.io/6frcw/?view_only=72344789a6294aa19d63a8bd93a628f3).

Results and discussion

Length of training

Figure 2A shows the time taken for the model to identify 95% or more of the word-object patterns in four consecutive tests. Additional simulations that were trained to a lower threshold of 90% correct criterion were also run, as some initial simulation runs failed to reach the 95% criterion by the end of training (Supporting Information, Figure S2).

We tested linear mixed effects (LME) models on length of training time (*lmer* and *lme4;* R [v3.6.3, 2020]), with number of objects during training (condition: 1, 2, or 6) as a categorical fixed effect (categorical so the difference between each of these contextual conditions on performance could be determined), gesture cue condition (0%, 33%, 67% and 100%) as a numeric fixed effect, and simulation run (1 to 10) as a random effect. We included number of objects during training and gesture condition as random slopes, but adding gesture cue condition, or the interaction between number of objects and gesture cue condition, resulted in the model not converging. The models were built including one fixed effect at a time, and using log-likelihood comparison to compare the contribution to model fit of each fixed effect (Barr et al., 2013).

Cues during training

Adding number of objects during training resulted in a significant improvement in fit, ($\chi^2(2) = 10.10$, $p = .006$). Quicker word learning was achieved with one object than two objects ($t(106.89) = 5.075$, $p < .001$), and two objects than six objects ($t(106.99) = 18.129$, $p < .001$). Gesture cue also significantly improved fit ($\chi^2(1) = 45.70$, $p < .001$), with greater cue availability resulting in quicker learning. The interaction also significantly improved fit ($\chi^2(2) = 14.23$, $p < .001$), with increasing availability of gesture cue having a stronger effect on learning speed in the two- and six-object conditions compared to the one-object condition ($t(114) = -3.572$, $p < .001$; $t(114) = -2.881$, $p = .005$, respectively). The effect of gesture cue on the two- and six-object conditions was not significantly different ($t(114) = 0.690$, $p = .491$). The resulting model is shown in Table 1 and the mean learning times for each object condition is shown in Figure 2A.

The model could learn word-referent mappings using cross-situational statistics and performed better with a cue: the addition of gesture (enhancing input activation from one location in the visual input layer) increased the associative learning signal from this region of the visual input. The model learned more quickly when there was no referential uncertainty about the target object – the one object condition learned faster than when two or six objects were present, but as we predicted, the gesture cue had a larger influence on learning under conditions of referential uncertainty. This of course makes perfect sense: when there is only one object, the model does not need support for disambiguating the referent. There was also a larger effect of gesture cue availability on the two-object than the six-object condition.

Accuracy at test

For testing performance, we constructed a series of generalised LME models in a similar way to the analyses of training length, with fixed effects of number of objects present during training and gesture cue condition, and random effects of simulation, but also an additional random effect of test item. Slopes for both fixed effects and their interaction were included for each random effect.

Number of objects present during training contributed significantly to fit ($\chi^2(2)$ = 18.43, $p < .001$), with one object resulting in lower accuracy than two and six objects (z = 18.77, z = 12.033, both $p < .001$, respectively), and six objects resulting in lower accuracy than two objects (z = -3.34, $p < .001$). Adding gesture cue did not significantly improve fit ($\chi^2(1)$ = 0.936, p = .333), but the interaction between gesture cue and number of objects during training was significant ($\chi^2(2)$ = 23.54, $p < .001$). As with the training time analysis, the effect of gesture cue availability had a stronger facilitative effect on accuracy for the two- and six-object conditions compared to the one-object condition (z = -8.64, z = -5.88; both $p <$ .001, respectively), and the effect of gesture cue availability on the two- and six-object conditions was not significantly different (z = 1.30, p = .194). The final model is shown in Table 1 and Figure 2B.

As there was a confound between training length and availability of gestural cues, additional simulations were run where the model was trained to the same amount of exposure for each of the different levels of availability of gestural cues, with similar accuracy results (Supporting Information, Table S1, Figure S2).

Unexpectedly, the model demonstrated more robust retention of the word-object mappings during testing when it had been trained under referential uncertainty; the two- and six-object conditions achieved higher accuracy than the one-object condition. In Monaghan (2017), the MIM performed best when there was some variability in the cues (when present 33% or 67% of the time) rather than with no variability (present 0% of the time) or a large degree of variability (present 100% of the time). However, in the current simulations, the effect of altering the number of potential referents in the environment for a given word also affected learning – some, but not a great deal, of referential uncertainty resulted in better learning, with the model demonstrating the highest accuracy in the two-object condition.

Thus, the computational model confirms our expectations about gesture being more important in the presence of referential uncertainty. We predict that if caregivers are sensitive to the potential value of a cue, then they ought to use more gestures in word learning situations when two unfamiliar referents are present rather than one. We might also predict that gestural cue use increases when six potential referents are present, though the model learned under these conditions to a similar degree irrespective of gesture cue availability.

However, the model also generated additional predictions that were unexpected: that word learning could actually be more successful when learning takes place under conditions of referential uncertainty. These results imply that variability in the environment can support learning. These hypotheses generated by the MIM were then tested in a behavioural word learning study with children aged 18–24-months-old and their caregivers.

Behavioural study

This experiment examined gesture use when caregivers taught their children novel word-object mappings under different degrees of referential uncertainty, and also explored whether gesture use under referential uncertainty predicts word learning. During training, caregivers taught their child three novel word-object pairs across the same conditions of referential uncertainty as simulated in the computational model – one, two, or six novel objects with a single target object per condition. Children were then tested on the novel word-object pairs taught by their caregiver during training.

Method

Participants

Forty-seven caregiver and child dyads, recruited through Lancaster Babylab, completed training (*M* = 20.5 months, *SD* = 1.7, male = 27; Table 2). All caregivers gave informed consent for the dyad. All dyads were from monolingual English homes, with no history of developmental or sensory disorders. The data from an additional six dyads were excluded due to child fussiness (Supporting Information, Table S9). Twenty-seven of the dyads that completed training also completed testing (*M* = 20.8 months, *SD* = 1.6, male = 13), with the remaining dyads excluded due to incomplete trials (16) or child fussiness (4). Dyads received a storybook for participation and reimbursement for travel expenses.

Stimuli

Three novel words were used: *darg*, *noop*, and *terb* (NOUN database; Horst & Hout, 2016). Nine similarly sized novel objects with different colours and shapes were used as stimuli (e.g. Figure 3). Three of these objects were randomly paired with the three novel words per participant. The remaining six objects then served as foils.

Training

Caregivers were familiarised with the three novel word-object pairs prior to the experiment without the child present. During training, the novel word and a three-word description of the target object were visible to the caregiver as a memory aid. Caregivers were told to imagine they were in an everyday setting, such as a shop with items on a shelf out of reach, and instructed to teach the novel words to their children as if they were real words for objects that the child had not seen before. Children then sat on their caregiver's lap and were presented with stimuli on a tray 70 cm away for 30 seconds, during which caregivers taught their child the novel word-object mapping (three training trials; 30 seconds each; one per novel word-object mapping). During training, dyads could not reach or handle the objects.

Dyads began with a warm-up trial where a red ball was presented on the tray and caregivers practised teaching their child the word 'ball'. All dyads were then administered all three conditions where target objects would appear alone (one-object condition), with another foil (two-object condition), or with five foils (six-object condition), reflecting the computational model's learning conditions (Figure 3A). A Latin Square was used to counterbalance the order in which training conditions were administered, and the position of targets per condition was also randomised in the same way as the computational model's training.

Testing

After training, children were tested by the experimenter on the three novel word-object mappings they had just learnt in a three alternative forced choice test, mirroring the computational model, with each word tested on separate trials (each word tested twice, six test trials in total; Figure 3B).

For each trial, the tray was arranged out of sight and then made visible. The then experimenter asked the child "Where is the [novel word]? Can you see the [novel word]? Point to the [novel word]." The tray was moved forward within the child's reach, and the child pointing towards, reaching for, or touching an object was recorded as a response. If the child did not respond, this was repeated; if the child still did not respond, the experimenter advanced to the next test trial. A Latin Square was used to counterbalance the order of conditions during testing across participants.

Coding

Training trials were video-recorded and coded per utterance for total gestures and speech co-occurring with gesture by a trained coder (see Supporting Information for details). An independent second rater coded 20% of the videos (randomly selected), with an inter-rater reliability of Cohen's $\kappa = 0.78$ for categorisation of gesture into subtypes (*deictic, representational, other*; $N = 284$; 85.21% agreement) and Cohen's $\kappa = 0.86$ for categorisation of speech with gesture into subtypes (*complementary* or *supplementary*; $N = 160$; 92.5% agreement).

An utterance was defined as a string of words or gestures preceded and followed by a pause or changes in conversation turn or intonation (Rowe et al., 2008). For gesture subtypes, we adapted Rowe et al.'s (2008) coding system: *deictic* gestures were intentional, clear movements that singled out the target, including pointing towards the target (e.g. finger points with the arm in extension) and reaches towards the target (e.g. extension of the arm with the palmar aspect of the hand exposed, or extension of the arm with the fingers in extension). *Representational* gestures included upper limb or body movements depicting object attributes such as shape or size (e.g. indicating a ball is round with two hands cupped and fingers flexed) and actions with the object (e.g. cupping the palmar aspect of one hand with fingers flexed, followed by arm movement forward from the shoulder joint, to indicate a ball rolling). *Other* gestures included all gestures not directed towards the referent; these included both deictic and representational gestures towards foils, to the experimenter, or caregiving-related gestures such as a parent hugging a child.

We adapted Iverson and Goldin-Meadow's (2005) coding system for speech with gesture in order to account for the effect of combined gesture and speech on learning as either *complementary*, where speech contained the target label, or as *supplementary*, where speech contained related information about the target referent such as size, colour, or function. Deictic gestures and occurrences of complementary speech with gesture correspond to the gestural cue conditions of the computational model. We also recorded the total number of times the *referent label* was used.

Vocabulary measures

Caregivers completed a demographics questionnaire that included socioeconomic status (SES; determined by parent education level). A parent-report measure of child vocabulary, the UK Communicative Development Inventories (CDI; Alcock, Meints, & Rowland, 2017) was also administered. The UK CDI measures expressive, receptive, and gesture vocabulary (communicative and symbolic). Communicative gestures include declarative and imperative gestures. Symbolic gestures are representational gestures that include actions, games, and pretend play.

Data, code, and models run are available on OSF (http://osf.io/6frcw/?view_only=72344789a6294aa19d63a8bd93a628f3).

Results and discussion

All dyads were from similar, mid-high SES backgrounds. Dyads that only completed training and those that completed both training and testing, did not yield any significant differences in demographics or CDI scores (Table 2).

To compare behavioural results to the computational model prediction that cue importance increased with referential ambiguity, we tested whether the number of objects during training affected caregiver behavioural cue use; in particular, deictic gesture use. LME models (*lmer* and *lme4*; R [v3.4.1, 2017]) were constructed to predict caregiver deictic gesture use, complementary speech with gesture, and referent label use separately. For each analysis, the number of objects during training (condition: 1, 2, or 6) was included as a categorical fixed effect, and child vocabulary was included as a numeric fixed effect. Due to high correlation between expressive and receptive vocabulary, separate linear mixed effects models were carried out – one with fixed effects of expressive, symbolic, and communicative gesture vocabulary, and one with receptive, symbolic, and communicative gesture vocabulary. Only the latter analysis is included here as the task required children to understand, rather than produce, novel words. Analyses with expressive vocabulary resulted in similar effects and are reported in the Supporting Information (Tables S3-S4). The models also contained random effects of participant, child age, target word, and target item. Slopes of condition per participant resulted in the model not converging. As for the computational model analysis, we included one fixed effect at a time, and used log-likelihood comparison to compare the contribution to model fit for each fixed effect (Barr et al., 2013). Separate LME models were also constructed in the same way to predict caregiver and child behaviour for each subtype described in our coding scheme to examine the range of caregiver communication with their children. We report here complementary speech with gesture and referent label use as these also highlight the referent in a similar manner to deictic gestures; all other subtypes can be found in Supporting Information (Tables S3-S4, Figure S4).

Cues during training

Caregiver data demonstrated a significant effect of condition on overall gesture use ($\chi^2(2) = 11.73$, $p = .003$). Consistent with the MIM results, this was largely due to deictic gesture cues ($\chi^2(2) = 9.48$, $p = .009$; Table 3, Figure 2C), with caregivers using more deictic gesture cues in the two-object ($t(90.24)= 2.32$, $p = .023$) and six-object ($t(91.79) = 3.08$, $p = .003$) conditions when compared to the one-object condition. Caregivers demonstrated no significant increase in deictic gesture use between two- and six-object conditions ($t(93.35) = 0.77$, $p = .445$). There were no significant fixed effects of child vocabulary or significant interactions found, and representative and other gestures did not yield any significant effects or interactions (Supporting Information, Figure S4A).

When examining caregiver complementary speech with gesture the addition of child symbolic gesture vocabulary improved model fit with a main effect of condition ($\chi^2(3) =$

0.43, $p < .001$; Table 4). Caregivers used more complementary speech with gesture in the two-object than the one-object condition ($t(80) = 2.58$, $p = .012$), but there was no significant difference between the two-object and six-object conditions ($t(80) = -0.89$, $p = .375$). A significant effect of condition on their overall use of the novel label was also found ($\chi^2(2) = 11.90$, $p = .003$, Table 4). The novel label was uttered significantly more by caregivers in the two-object compared to the one-object condition ($t(89.493) = 2.37$, $p = .020$), but significantly less in the six-object compared to the two-object condition ($t(89.658) = -3.52$, $p < .001$). No other significant effects of child vocabulary or interactions were found.

Overall, these results were consistent with the MIM model showing the largest effect of gesture availability in the two- and six-object conditions.

Accuracy at test

We used Generalised Estimated Equations (GEE; *geeglm* and *geepack*; R[v3.4.1, 2017]) to examine the effect of condition, caregiver behaviour, and child behaviour during training on test trial accuracy.[1] Separate GEEs were constructed to examine child vocabulary variables, condition, and each training behaviour gesture subtype as independent variables; here we report the effect of caregiver deictic gesture use with child receptive vocabulary. For all other subtypes and child vocabulary variables, please see Supporting Information (Tables S5-S8).

In line with the computational model results, children performed most accurately in the two-object condition (Table 3, Figure 2D), although there was no significant difference in accuracy between the two-object and six-object condition (*Wald* = 0.01, $p = .921$). However, children responded significantly more accurately in the two-object than the one-object condition, even when child receptive vocabulary and caregiver deictic gesture use were accounted for (*Wald* = 4.36, $p = .037$).

Although the lack of referential ambiguity would suggest that word-object mapping should be easier in the one-object condition, a higher success of word learning in the two- and six-object conditions was consistent with the MIM computational results. Additionally, although children were offered the least amount of gesture information by caregivers in the one-object condition, adding caregiver behaviour subtypes during training to the analysis did not contribute any significant value to predicting accuracy during testing (Table 3; Supporting Information, Tables S5-S8).

General Discussion

Natural language learning environments are noisy and variable, and yet children still manage to accurately map words to objects. In this study, we predicted that a computational model of word learning (MIM) trained under conditions of varying referential uncertainty would learn faster with fewer potential referents. We also predicted that a gestural cue would be most helpful to word-referent mapping when there was an increase in potential referents.

Contrary to our first prediction, but consistent with literature highlighting the value of variability during word learning (e.g. Apfelbaum & McMurray, 2011; Monaghan, 2017), the computational model predicted the most robust learning when there were several potential referents, rather than just one. Although the MIM learnt quickest in the one-object condition, there was higher accuracy at test when it had been trained under referential uncertainty during the two- and six- object conditions. The addition of a gestural cue during training significantly improved learning when there were more potential referents as predicted, but the model also benefited from the presence of variability via the availability of gestural cues, learning most robustly when cues were presented 33% and 67% of the time.

This generated two hypotheses for testing in behavioural settings. Firstly, if caregivers are sensitive to the role of gestural cues in supporting word learning, they ought to use more gestures when there is referential uncertainty, and secondly, children might actually learn best when trained under referentially uncertain conditions. The experimental study did identify that caregivers adapt their gestural cues to support learning in the face of referential uncertainty, but with significant increases only from the one-object to the two-object or six-object condition, and no significant increase from the two-object to the six-object condition. Finally, the experimental study also found that children learnt best under referential uncertainty, performing most accurately in the two- and six-object conditions, in line with the model's surprising predictions.

These results were somewhat counterintuitive; one might expect the highest test accuracy in the behavioural study for words learnt in the one-object condition. This would be consistent with the fast-mapping literature, where children are able to identify a new word after a single exposure (Carey & Bartlett, 1978), and with cross-situational word learning in adults that indicates increasing the number of potential referents results in less accurate and slower learning (K. Smith et al., 2011; Trueswell et al., 2013; Yu & L.B. Smith, 2007). Despite this, our task differed in several ways that could have affected performance at test. Firstly, children were not tested on each word after the corresponding training trial as in referent selection trials during fast-mapping tasks (Horst & Samuelson, 2008) – they were tested after all training trials. Secondly, the co-occurring foils were novel, whereas fast-mapping tasks involve familiar objects alongside novel objects. Cross-situational word learning paradigms also usually offer the opportunity to learn from within- and across-trial competition as all objects are named (Yurovsky et al., 2013). In our study, there was no such opportunity, as different foils were used within-subject for each condition, and testing trials consisted of forced-choice between the three target objects.

Rather, it is possible that the presence of referential uncertainty in the two- and six-object conditions might have supported learning through enabling comparison. The role of two or more competing alternatives is well established in internal constraint accounts of language learning, including mutual exclusivity (Halberda, 2006; Markman & Wachtel, 1988) and the novel name-nameless category principle (Golinkoff et al., 1992). Similarly, children's learning of categories is aided by having an alternative, either by using comparison, where one object appears with others in the same category, or by contrast, where an object appears with a non-category object (Ankowski et al., 2013). Such a beneficial effect may also apply in our study where the referent is identified among a range of other unknown objects.

Few studies have examined cross-situational referential ambiguity in infants and children, with most limiting referential ambiguity to two potential referents per training trial (e.g. L.B. Smith & Yu, 2008; Yu & L.B. Smith, 2011). Those that have examined older children (5–7-years-old) suggest that they may struggle most when a specific foil, termed a high probability competitor, co-occurs with a target more often than other foils (Suanda et al., 2014). Bunce and Scott (2017) examined 2.5-year-old children with four potential referents per trial. Children could identify the correct target using cross-situational statistics with four potential referents without exhaustive labelling when all distractors were different (no across-trial competition), and even with a high probability competitor – but only if a different foil appeared by the last trial, allowing disambiguation at the end of training. This suggests that children are able to learn under certain circumstances with increased referential ambiguity, subject to limitations in cognitive and memory capacity.

Another potential explanation for performance in the one-object condition is that children were less interested compared to when there were several objects present. Future research could use an eye-tracker to measure attention more precisely and determine how foils are fixated on alongside targets. Testing immediately after training trials using both target and foil objects may also help illuminate whether children process all objects present.

The present computational model and experimental study also highlighted that some variability in both the environment and in the use of cues in communication may facilitate learning. We have demonstrated that the former influences the latter, establishing that caregiver gesture cue use when teaching their children novel words was contingent on the presence of referential uncertainty. This is consistent with the theory that gestures singling out target referents are particularly valuable during word-object mapping (Cartmill et al., 2013; Rader & Zukow-Goldring, 2012). However, although we expected gesture use during training to increase from the two- to the six-object condition, this was not the case. Hence, caregivers gestured and offered cues according to the presence, rather than the degree, of referential uncertainty, and did not offer significantly more cues when referential uncertainty was high.

Taken together, these results indicate that referential uncertainty is perhaps subject to some degree of cognitive management by both the caregiver and child, where high uncertainty can be reduced to a more tractable sense of 'this, not that'. The use of gestural cues may reduce cognitive load for the infant (Goldin-Meadow, 2000; McGregor et al., 2009; McNeil et al., 2000); the key difference in our study seemed to be between having either one choice of word-object mapping, or more than one – beyond this, the benefits of gestural cues may begin to decline. Caregivers appeared to be sensitive to this lack of discrimination between the two- and six-object conditions, as there was no significant difference in their behaviour. A switch to laborious counting during the six-object condition, rather than being able to immediately perceive the number of items in the one- and two-object conditions (Cowan, 2001; Xu & Chun, 2009), may have affected how the caregiver then packaged information for their child. This could potentially lead to the treatment of the two- and six-objects as analogous by the caregiver, and thus the child. Similarly, gestural cues did not have a large effect on speed of learning in the six-object condition in the computational model compared to the one- and two-object conditions.

Studies of how children acquire representations of number additionally indicate that children around 20-months-old are not able to comprehend more than three or four objects (Feigenson et al., 2004; Le Corre & Carey, 2007; Wynn, 1990), which could also render performance across the two- and six-object conditions in our study somewhat analogous. Despite this, being able to distinguish only a limited number of stimuli may also help constrain word-referent mappings. Head-mounted cameras during toy exploration laboratory studies show that, despite having multiple objects in front of them, 20-month-olds tend to hold single objects in view at a time (L.B. Smith et al., 2011) and learn the names of objects that dominate their view simultaneously with label utterance (Pereira et al., 2014).

However, we did not test incremental increases in referential uncertainty, opting instead for no ambiguity, some ambiguity, and high ambiguity. An interesting avenue for future research would be to investigate whether there is a precise ‘tipping point’ in the number of potential referents at which caregivers cease to offer more gestural cues to their children and whether this then affects children’s learning – although the similar performance between the two-object and six-object conditions may suggest that anomalies in behaviour and learning are unlikely to occur with intermediate ambiguity between two and six objects.

Although cues are useful for supporting learning, they are also individually highly variable within naturalistic environments. Caregivers may not gesture towards intended referents on 85% of occasions (Iverson et al., 1999), articles may precede adjectives rather than nouns (Monaghan et al., 2007), and prosodic cues also are not always consistent (Fernald, 1991). The computational MIM simulations also found that the most robust learning occurred when gestural cues were present some of the time, rather than when they were exclusively present or absent. Why is this? Firstly, it has been established that a system that relies on perfectly reliable cues learns quickly, but learning is brittle when those cues are no longer reliable (Monaghan, 2017). Secondly, when identifying a target from amongst different competitors, the occasional lack of a cue may make the presence of one more salient, avoiding potential habituation effects (Veale et al., 2011) and preventing inhibition of other useful information (Kamin blocking effect; Shanks, 1985). A system where gestural cues vary may then have a higher degree of sensitivity to those cues than one where gestural cues are either always there, or always absent. Thus, variability of cues is not only more similar to real-world settings, but also benefits learning. This raises the intriguing possibility that the variability of cues when children are acquiring vocabulary may not be an accident of a noisy environment, but rather the stochasticity of adults' use of cues may be by design.

In our experimental study, we did not find any effect of training response variables on testing data – inclusion of caregiver gesture and speech use did not predict child accuracy after controlling for condition. If referential uncertainty and the cues in response to it are so vital to learning, why did this not manifest in our data? This may be partly due to our sample of mid-to-high SES families who had actively expressed interest in developmental research. Families from higher SES backgrounds have been found to use gesture more than those from lower SES backgrounds, with an increase in parental gesture correlating with increased child gesture and later vocabulary skill (Rowe & Goldin-Meadow, 2009). Our participants may well have been at a ceiling level of caregiver input, resulting in gesture adding very little. Gesture may be particularly beneficial to language development in environments with limited resources and a diminished quality of parental input (Kirk et al., 2013), and may be useful as part of language interventions in low income families (Vallotton, 2012). Consequently, we recommend that caution should be exercised when generalising our conclusions across different SES backgrounds.

Additionally, as our sample inclusion criteria precluded developmental delay, our findings may not extend to these populations (Hartley et al., 2019, 2020). Our results confirmed that caregivers appeared to be sensitive to task demands, and models predicting speech with gesture during training were improved with the addition of CDI subscales. Although these estimates were very small, the impact of child vocabulary could be more prominent in a language delayed sample (Wray and Norbury, 2018).

An alternative explanation concerning why caregiver behaviour did not predict children's behaviour relates to our sample's age (20-months-old on average). Previous literature links caregiver gesture and early child gesture use at 10–14-months-old (Liszkowski et al., 2012; Liszkowski & Tomasello, 2011) and caregiver gesture use in Rowe et al. (2008) predicted early child gesture use at 14-months-old, but not expressive vocabulary at 42-months-old. Caregiver gesture use appears relatively stable over time, whereas child gesture use may take a supportive role to speech once verbal ability is established (Goldin-Meadow, 2007; Iverson et al., 1999; Rowe et al., 2008). Subsequently, children in our study may have been at a stage where verbal input is weighted more heavily than gesture input. Although we examined some of these factors, our primary focus was deictic gestures. Future research could consider speech input in greater depth, including Mean Length of Utterance and temporal relations of naming events with gesture.

We also found a higher level of child dropout when testing trials commenced, reducing power for GEE analysis (which could also reduce the effect of caregiver gesture on child behaviour; Liszkowski et al., 2012; Liszkowski & Tomasello, 2011). Although we observed no significant differences between children that completed testing and those that did not, child fussiness may have been caused by objects being out of reach during training, resulting in frustration by the time testing commenced. This may mean that differences in temperament and attention could be present that were not accounted for. Additionally, whereas previous studies enabled children to freely explore an environment, we constrained the objects in our study to be out of reach to control for exposure times and interaction with the objects. This could have resulted in less gesture, particularly by children, who had no immediate receipt of the objects to which they gestured. Studies that compare objects within reach across varying environmental referential uncertainty, and that measure broader child traits, will usefully address these points. To isolate any effect of referential uncertainty itself from caregiver behaviour, future studies could also test children's word learning across referential uncertainty without caregiver interaction.

In conclusion, we found variability in gesture cue availability combined with referential ambiguity produced optimal learning in a computational model of word learning. This was supported by an experimental study that demonstrated that: (a) caregivers gestured according to the presence, rather than degree, of referential uncertainty, and (b) children learnt best in the presence, rather than absence, of referential uncertainty. These results advance understanding of communicative exchange during word learning, indicating that caregivers contingently adapt their gesture use according to the presence of referential uncertainty.

References

Alcock, K. J., Meints, K., & Rowland, C. F. (2017). *UK-CDI Words and Gestures: Preliminary Norms and Manual*. http://lucid.ac.uk/ukcdi

Ankowski, A. A., Vlach, H. A., & Sandhofer, C. M. (2013). Comparison versus contrast: Task specifics affect category acquisition. *Infant and Child Development*, *22*(1), 1–23. https://doi.org/10.1002/icd.1764

Apfelbaum, K. S., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science*, *35*(6), 1105–1138. https://doi.org/10.1111/j.1551-6709.2011.01181.x

Bahtiyar, S., & Küntay, A. C. (2009). Integration of communicative partner's visual perspective in patterns of referential requests. *Journal of Child Language*, *36*(3), 529–555. https://doi.org/10.1017/S0305000908009094

Bannard, C., Rosner, M., & Matthews, D. (2017). What's worth talking about? Information Theory reveals how children balance informativeness and ease of production. *Psychological Science*, *28*(7), 954–966. https://doi.org/10.1177/0956797617699848

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3). https://doi.org/10.1016/j.jml.2012.11.001

Brooks, R., & Meltzoff, A. N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: A longitudinal, growth curve modeling study. *Journal of Child Language*, *35*(1), 207–220. https://doi.org/10.1017/s030500090700829x

Brown-Schmidt, S., & Duff, M. C. (2016). Memory and common ground processes in language use. *Topics in Cognitive Science*, *8*(4), 722–736. https://doi.org/10.1111/tops.12224

Bunce, J. P., & Scott, R. M. (2017). Finding meaning in a noisy world: Exploring the effects of referential ambiguity and competition on 2·5-year-olds' cross-situational word learning. *Journal of Child Language*, *44*(3), 650–676. https://doi.org/10.1017/S0305000916000180

Callanan, M. A., & Sabbagh, M. A. (2004). Multiple labels for objects in conversations with young children: Parents' language and children's developing expectations about word meanings. *Developmental Psychology*, *40*(5), 746–763. https://doi.org/10.1037/0012-1649.40.5.746

Capone, N. C., & Mcgregor, K. K. (2004). Gesture development: A review for clinical and research practices. *Journal of Speech, Language, and Hearing Research*, 173186.

Carey, S. (1988). Conceptual differences between children and adults. *Mind & Language*, *3*(3), 167–181. https://doi.org/10.1111/j.1468-0017.1988.tb00141.x

Carey, S., & Bartlett, E. (1978). *Acquiring a Single New Word.*

Carpenter, M., Nagell, K., & Tomasello, M. (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–vi, 1–143.

Cartmill, E. A., Armstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences*, *110*(28), 11278–11283. https://doi.org/10.1073/pnas.1309518110

Cleave, P. L., & Bird, E. K.-R. (2006). Effects of familiarity on mothers' talk about nouns and verbs. *Journal of Child Language*, *33*(3), 661–676. https://doi.org/10.1017/S0305000906007549

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. https://doi.org/10.1017/S0140525X01003922

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314. https://doi.org/10.1016/j.tics.2004.05.002

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S. J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, *59*(5), 1–173; discussion 174-185.

Fernald, A., & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Developmental Psychology*, *27*(2), 209–221. https://doi.org/10.1037/0012-1649.27.2.209

Fries, C. C. (1952). *The Structure of English*. Longmans.

Goldin-Meadow, S. (2000). Beyond words: The importance of gesture to researchers and learners. *Child Development*, *71*(1), 231–239. https://doi.org/10.1111/1467-8624.00138

Goldin-Meadow, Susan. (2007). Pointing Sets the Stage for Learning Language—And Creating Language. *Child Development*, *78*(3), 741–745. https://doi.org/10.1111/j.1467-8624.2007.01029.x

Goldin-Meadow, S., & Wagner, S. M. (2005). How our hands help us learn. *Trends in Cognitive Sciences*, *9*(5).

Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L. M., & Wenger, N. R. (1992). Young children and adults use lexical principles to learn new nouns. *Developmental Psychology*, 99–108. https://doi.org/10.1037/0012-1649.28.1.99

Gorman, K. S., Gegg-Harrison, W., Marsh, C. R., & Tanenhaus, M. K. (2013). What's learned together stays together: Speakers' choice of referring expression reflects shared experience. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(3). https://doi.org/10.1037/a0029467

Halberda, J. (2006). Is this a dax which I see before me? Use of the logical argument disjunctive syllogism supports word-learning in children and adults. *Cognitive Psychology*, *53*(4), 310–344. https://doi.org/10.1016/j.cogpsych.2006.04.003

Hartley, C., Bird, L.-A., & Monaghan, P. (2019). Investigating the relationship between fast mapping, retention, and generalisation of words in children with autism spectrum disorder and typical development. *Cognition*, *187*, 126–138. https://doi.org/10.1016/j.cognition.2019.03.001

Hartley, C., Bird, L.-A., & Monaghan, P. (2020). Comparing cross-situational word learning, retention, and generalisation in children with autism and typical development. *Cognition*, *200*, 104265. https://doi.org/10.1016/j.cognition.2020.104265

Horst, J. S., & Hout, M. C. (2016). The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. *Behavior Research Methods*, *48*(4), 1393–1409. https://doi.org/10.3758/s13428-015-0647-3

Horst, J. S., & Samuelson, L. K. (2008). Fast mapping but poor retention by 24-month-old infants. *Infancy*, *13*(2), 128–157. https://doi.org/10.1080/15250000701795598

Iverson, J. M., Capirci, O., Longobardi, E., & Caselli, M. C. (1999). Gesturing in mother-child interactions. *Cognitive Development*, *14*(1), 57–75. https://doi.org/10.1016/S0885-2014(99)80018-5

Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science*, *16*(5), 367–371. https://doi.org/10.1111/j.0956-7976.2005.01542.x

Kirk, E., Howlett, N., Pine, K. J., & Fletcher, B. C. (2013). To Sign or Not to Sign? : The Impact of Encouraging Infants to Gesture on Infant Language and Maternal Mind-

Mindedness. *Child Development*, 574–590. https://doi.10.1111/j.1467-8624.2012.01874.x

Kuhn, L. J., Willoughby, M. T., Wilbourn, M. P., Vernon-Feagans, L., & Blair, C. B. (2014). Early communicative gestures prospectively predict language development and executive function in early childhood. *Child Development*, *85*(5), 1898–1914. https://doi.org/10.1111/cdev.12249

Le Corre, M., & Carey, S. (2007). One, Two, Three, Four, Nothing More: An Investigation of the Conceptual Sources of the Verbal Counting Principles. *Cognition*, *105*(2). https://doi.org/10.1016/j.cognition.2006.10.005

Liszkowski, U., Brown, P., Callaghan, T., Takada, A., & Vos, C. de. (2012). A prelinguistic gestural universal of human communication. *Cognitive Science*, *36*(4), 698–713. https://doi.org/10.1111/j.1551-6709.2011.01228.x

Liszkowski, U., & Tomasello, M. (2011). Individual differences in social, cognitive, and morphological aspects of infant pointing. *Cognitive Development*, *26*(1), 16–29. https://doi.org/10.1016/j.cogdev.2010.10.001

Luce, M. R., & Callanan, M. A. (2010). Parents' object labeling: Possible links to conventionality of word meaning? *First Language*, *30*(3–4), 270–286. https://doi.org/10.1177/0142723710370543

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meaning of words. *Cognitive Psychology*, *20*(2), 121–157. https://doi.org/10.1016/0010-0285(88)90017-5

Masur, E. F. (1997). Maternal labelling of novel and familiar objects: Implications for children's development of lexical constraints. *Journal of Child Language*, *24*(2), 427–439. https://doi.org/10.1017/S0305000997003115

Mcgregor, K. K., Rohlfing, K. J., Bean, A., & Marschner, E. (2009). Gesture as a support for word learning: The case of under. *Journal of Child Language*, *36*(4), 807–828. https://doi.org/10.1017/S0305000908009173

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*(4), 831–877. https://doi.org/10.1037/a0029872

McNeil, N. M., Alibali, M. W., & Evans, J. L. (2000). The role of gesture in children's comprehension of spoken language: Now they need it, now they don't. *Journal of Nonverbal Behavior*, *24*(2), 131–150. https://doi.org/10.1023/A:1006657929803

Mervis, C. B. (1987). Child-basic object categories and early lexical development. In *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 201–233). Cambridge University Press.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91–117. https://doi.org/0.1016/s0010-0277(03)00140-9

Monaghan, P. (2017). Canalization of language structure from environmental constraints: A computational model of word learning from multiple cues. *Topics in Cognitive Science*, *9*(1), 21–34. https://doi.org/10.1111/tops.12239

Monaghan, P., Brand, J., Frost, R. L. A., & Taylor, G. (2017). Multiple variable cues in the environment promote accurate and robust word learning. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, 817–822.

Monaghan, P., Christiansen, M. H., & Chater, N. (2007). The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology*, *55*(4), 259–305. https://doi.org/10.1016/j.cogpsych.2006.12.001

Monaghan, P., & Mattock, K. (2012). Integrating constraints for learning word-referent mappings. *Cognition*, *123*(1), 133–143. https://doi.org/10.1016/j.cognition.2011.12.010

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science*, *13*(4), 329–336. https://doi.org/10.1111/j.0956-7976.2002.00460.x

Nilsen, E. S., & Graham, S. A. (2009). The relations between children's communicative perspective-taking and executive functioning. *Cognitive Psychology*, *58*(2), 220–249. https://doi.org/10.1016/j.cogpsych.2008.07.002

O'Neill, D. K. (1996). Two-Year-Old Children's Sensitivity to a Parent's Knowledge State When Making Requests. *Child Development*, *67*(2), 659–677. https://doi.org/10.1111/j.1467-8624.1996.tb01758.x

Özçalişkan, Ş., & Goldin-Meadow, S. (2005). Do parents lead their children by the hand? *Journal of Child Language*, *32*(3), 481–505. https://doi.org/10.1017/S0305000905007002

Özçalışkan, S., & Dimitrova, N. (2013). How gesture input provides a helping hand to language development. *Seminars in Speech and Language*, *34*(4), 227–236. https://doi.org/10.1055/s-0033-1353447

Pearlmutter. (1989). Learning state space trajectories in recurrent neural networks. *International 1989 Joint Conference on Neural Networks*, 365–372 vol.2. https://doi.org/10.1109/IJCNN.1989.118724

Pereira, A. F., Smith, L. B., & Yu, C. (2014). A bottom-up view of toddler word learning. *Psychonomic Bulletin & Review*, *21*(1), 178–185. https://doi.org/10.3758/s13423-013-0466-4

Quine, W. V. O. (1960). *Word & Object*. MIT Press.

Rader, N. de V., & Zukow-Goldring, P. (2012). Caregivers' gestures direct infant attention during early word learning: The importance of dynamic synchrony. *Language Sciences*, *34*(5), 559–568. https://doi.org/10.1016/j.langsci.2012.03.011

Rowe, M. L., & Goldin-Meadow, S. (2009). Differences in early gesture explain SES disparities in child vocabulary size at school entry. *Science*, *323*(5916), 951–953. https://doi.org/10.1126/science.1167025

Rowe, M. L., Özçalişkan, Ş., & Goldin-Meadow, S. (2008). Learning words by hand: Gesture's role in predicting vocabulary development. *First Language*, *28*(2), 182–199. https://doi.org/10.1177/0142723707088310

Samuelson, L. K., Kucker, S. C., & Spencer, J. P. (2017). Moving word learning to a novel space: A dynamic systems view of referent selection and retention. *Cognitive Science*, *41*, 52–72. https://doi.org/10.1111/cogs.12369

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgement. *The Quarterly Journal of Experimental Psychology*, *37B*, 1–21. https://doi.org/10.1080/14640748508402082

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1–2), 39–91.

Smith, A. C., Monaghan, P., & Huettig, F. (2017). The multimodal nature of spoken word processing in the visual world: Testing the predictions of alternative models of multimodal integration. *Journal of Memory and Language*, *93*, 276–303. https://doi.org/10.1016/j.jml.2016.08.005

Smith, K., Smith, A. D. M., & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, *35*(3), 480–498. https://doi.org/10.1111/j.1551-6709.2010.01158.x

Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568. https://doi.org/10.1016/j.cognition.2007.06.010

Smith, L. B., Yu, C., & Pereira, A. F. (2011). Not your mother's view: The dynamics of toddler visual experience: Dynamics of toddler visual experience. *Developmental Science*, *14*(1), 9–17. https://doi.org/10.1111/j.1467-7687.2009.00947.x

Southgate, V., Maanen, C. V., & Csibra, G. (2007). Infant pointing: Communication to cooperate or communication to learn? *Child Development*, *78*(3), 735–740. https://doi.org/10.1111/j.1467-8624.2007.01028.x

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.

Suanda, S. H., Mugwanya, N., & Namy, L. L. (2014). Cross-situational statistical word learning in young children. *Journal of Experimental Child Psychology*, *126*, 395–411. https://doi.org/10.1016/j.jecp.2014.06.003

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*(1), 126–156. https://doi.org/10.1016/j.cogpsych.2012.10.001

Vallotton, C. D. (2012). Infant signs as intervention? Promoting symbolic gestures for preverbal children in low-income families supports responsive parent–child relationships. *Early Childhood Research Quarterly*, *27*(3), 401–415. https://doi.org/10.1016/j.ecresq.2012.01.003

Veale, R., Schermerhorn, P., & Scheutz, M. (2011). Temporal, environmental, and social constraints of word-referent learning in young infants: A neurorobotic model of multimodal habituation. *IEEE Transactions on Autonomous Mental Development*, *3*(2), 129–145. https://doi.org/10.1109/TAMD.2010.2100043

Vigliocco, G., Motamedi, Y., Murgiano, M., Wonnacott, E., Marshall, C., Maillo, I. M., & Perniss, P. (2019). Onomatopoeia, gestures, actions and words: How do caregivers use multimodal cues in their communication to children? *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, 1171–1177.

Wray, C., & Norbury, C. F. (2018). Parents modify gesture according to task demands and child language needs. *First Language*, *38*(4), 419–439. https://doi.org/10.1177/0142723718761729

Wynn, K. (1990). Children's understanding of counting. *Cognition*, *36*(2), 155–193. https://doi.org/10.1016/0010-0277(90)90003-3

Xu, Y., & Chun, M. M. (2009). Selecting and perceiving multiple visual objects. *Trends in Cognitive Sciences*, *13*(4), 167–174. https://doi.org/10.1016/j.tics.2009.01.008

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13), 2149–2165. https://doi.org/10.1016/j.neucom.2006.01.034

Yu, C., & Smith, L. (2011). What you learn is what you see: Using eye movements to study infant cross-situational word learning. *Developmental Science*, *14*(2), 165–180.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420. https://doi.org/10.1111/j.1467-9280.2007.01915.x

Yu, C., & Smith, L. B. (2012). Modeling cross-situational word-referent learning: Prior questions. *Psychological Review*, *119*(1), 21–39. https://doi.org/10.1037/a0026182

Yurovsky, D., Yu, C., & Smith, L. B. (2013). Competitive processes in cross-situational word learning. *Cognitive Science*, *37*(5), 891–921. https://doi.org/10.1111/cogs.12035

Figures

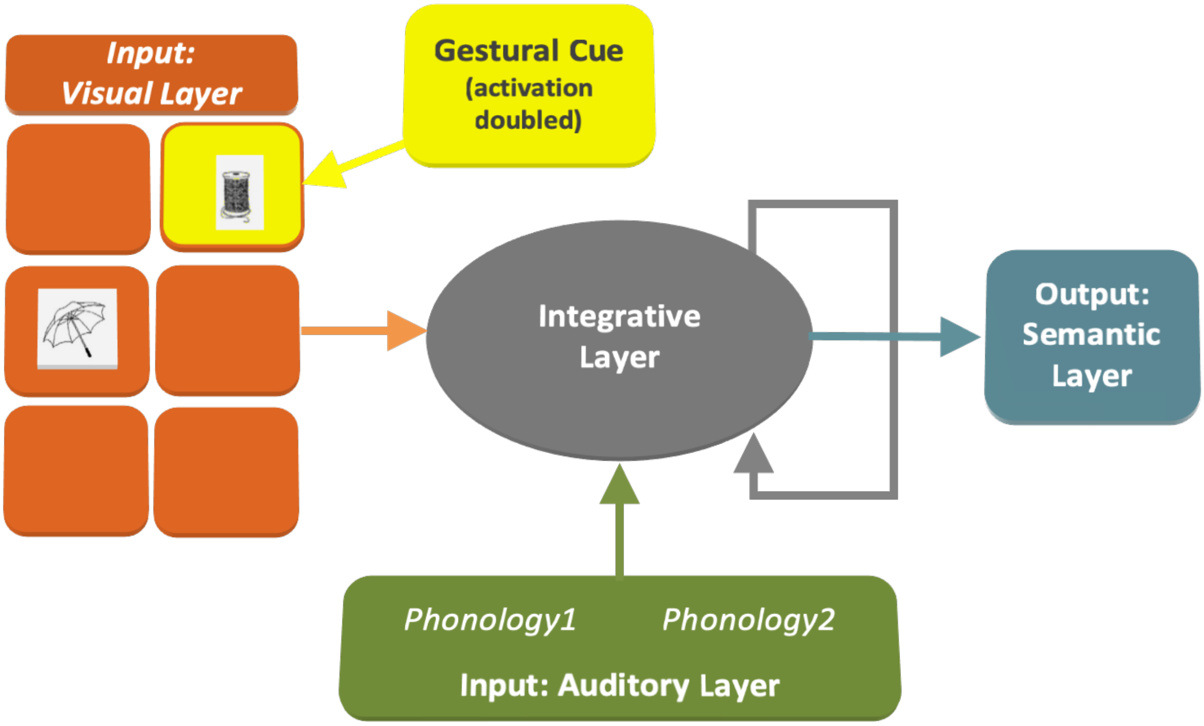

Figure 1. Architecture of the multimodal integration model (MIM) for word-object mapping (example of two-object training condition with gesture cue present).

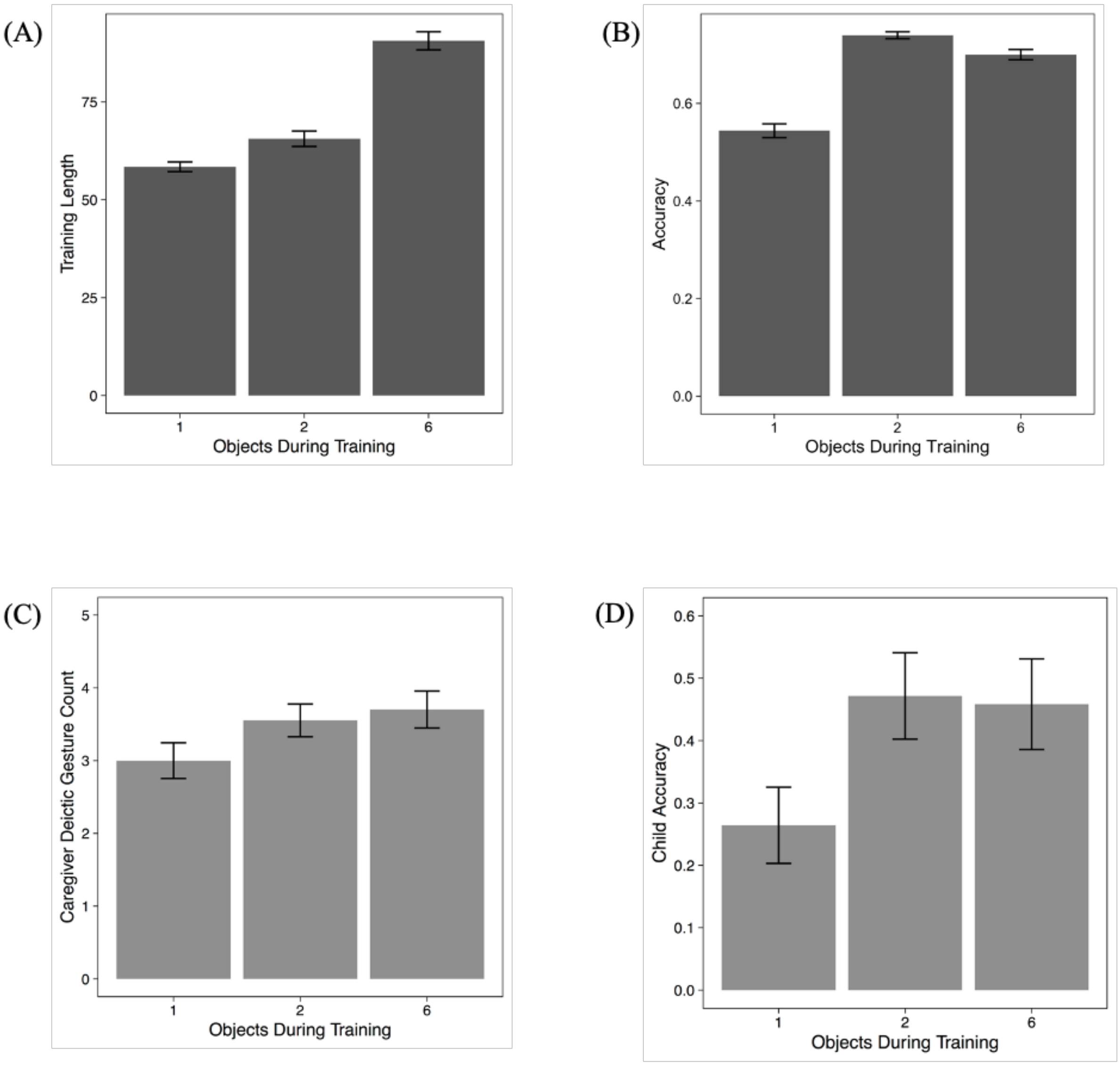

Figure 2. Mean and standard error bars for results of the MIM and behavioural study. Note that for testing accuracy, there were three objects present and no gesture cue. (A) MIM: Training length time by number of objects present during training (calculated across gesture cue condition);† (B) MIM: Testing accuracy proportion correct by number of objects present during training (calculated across gesture cue condition);† (C) Behavioural study: Count of caregiver deictic gesture use by number of objects present during training; (D) Behavioural study: Child testing accuracy proportion correct by number of objects present during training. †For MIM results by number of objects present during training and by individual gesture cue condition, please see Supporting Information, Figure S1.

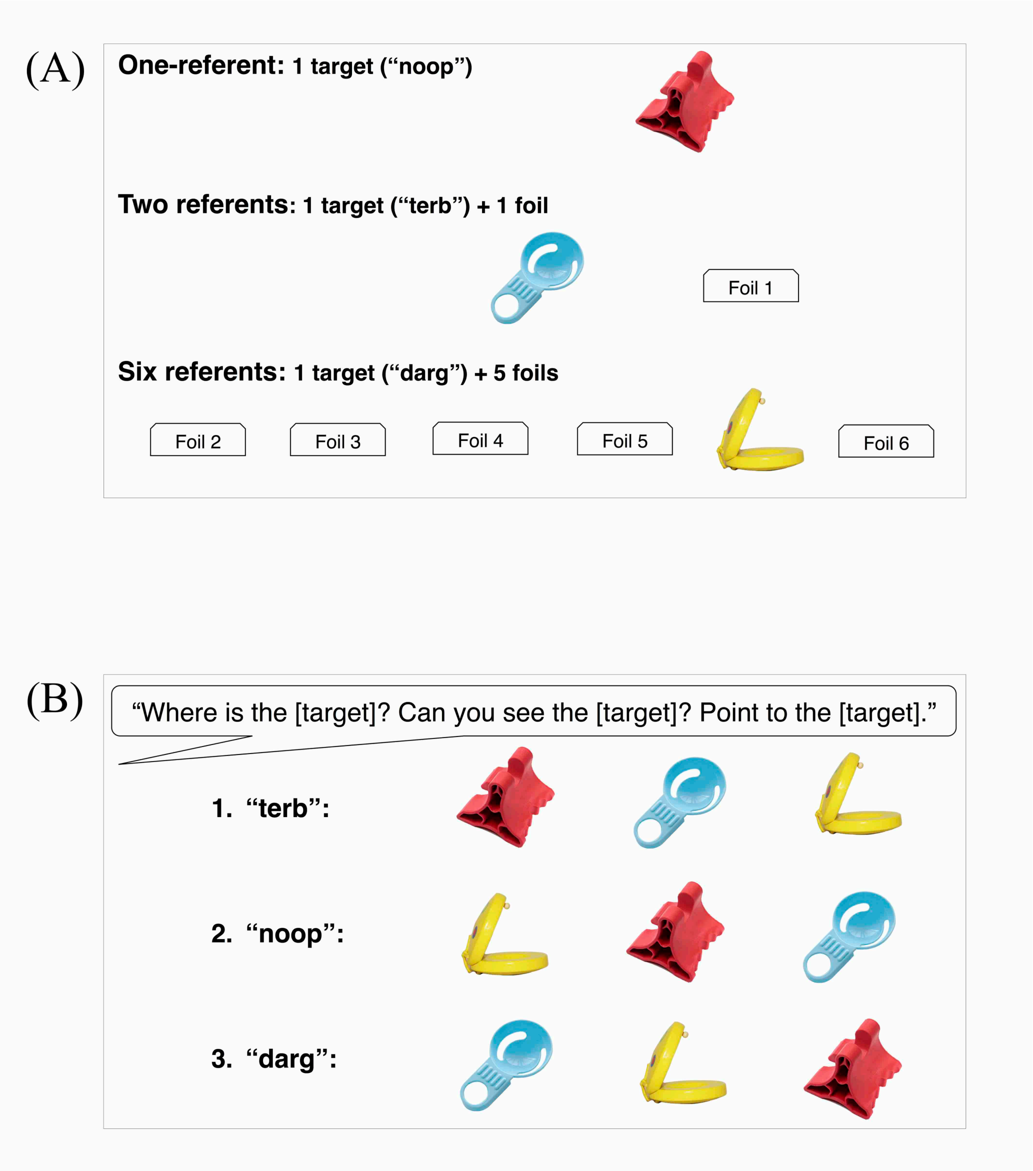
(A)
One-referent: 1 target ("noop")
Two referents: 1 target ("terb") + 1 foil
Foil 1
Six referents: 1 target ("darg") + 5 foils
Foil 2
Foil 3
Foil 4
Foil 5
Foil 6
(B)
"Where is the [target]? Can you see the [target]? Point to the [target]."
1. "terb":
2. "noop":
3. "darg":

Figure 3. Behavioural study: (A) Example of training trials; (B) Example of testing trials.

Tables

Table 1. Computational model: linear mixed effects model results of the MIM computational model's performance, testing the effects of number of objects during training and gesture cue condition on length of training time and accuracy.

Dependent variable	Independent variables	*Estimate*	*SE*	*df*	*t*	*p-value*
Length of training time	(intercept – one object)	68.62	1.49	114	46.05	< .001
	One v. two objects	13.21	2.11	114	6.267	< .001
	One v. six objects	37.04	2.11	114	17.58	< .001
	Two vs. six objects	23.84	2.11	114	11.31	< .001
	Gesture cue	-20.39	2.39	114	-8.54	< .001
	One v. Two object x Gesture cue	-12.06	3.38	114	-3.57	< .001
	One v. Six object x Gesture cue	-9.73	3.38	114	-2.88	.005
	Two v. Six object x Gesture cue	2.33	3.38	114	.690	.491
		Estimate	*SE*		*z*	*p-value*
Testing accuracy after training to criterion	(intercept – one object)	-0.532	0.16		-3.36	< .001
	One v. two objects	2.67	0.19		13.91	< .001
	One v. six objects	1.94	0.21		9.30	< .001
	Two vs. six objects	-0.66	0.20		-3.35	< .001
	Gesture cue	0.40	0.07		5.58	< .001
	One v. two objects x Gesture cue	-0.54	0.08		-6.80	< .001
	One v. six objects x Gesture cue	-0.45	0.09		-4.98	< .001
	Two v. six objects x Gesture cue	0.07	0.07		0.91	.365
		Estimate	*SE*		*z*	*p-value*
Testing accuracy after extended training	(intercept – one object)	-0.696	0.15		-4.51	< .001
	One v. two objects	2.945	0.18		16.20	< .001
	One v. six objects	2.107	0.21		9.93	< .001
	Two vs. six objects	-0.448	0.13		-3.50	< .001
	Gesture cue	0.484	0.07		6.68	< .001
	One v. two objects x Gesture cue	-0.657	0.08		-8.64	< .001
	One v. six objects x Gesture cue	-0.547	0.09		-5.88	< .001
	Two v. six objects x Gesture cue	0.259	0.20		1.30	.194

Table 2. Behavioural study: demographics and child vocabulary scores as measured by the UK-Communicative Development Inventories with Welch Two Sample T-Tests comparing those that completed training only, and those that completed training and testing.

	Completed training (total sample; *N* = 47)	Completed training + testing trials (*n* = 27)	Completed training only (*n* = 20)	Welch Two Sample T-Tests (completed training + testing, v. completed training only)		
Sex (m:f ratio)	27:20	14:13	13:7			
	mean (sd)	*mean (sd)*	*mean (sd)*	*t (df)*	*95% CI*	*p-value*
Age (months)	20.5 (1.7)	20.8 (1.6)	20 (1.8)	-2 (38)	[-1.85, 0.22]	.1
Receptive	276 (91.5)	294 (87.9)	251 (92.9)	-2 (40)	[-96.5, 11.7]	.1
Expressive	146 (114)	159 (119)	129 (108)	-0.9 (43)	[-97.2, 36.8]	.4
Comm. gesture	19.9 (3.79)	20.5 (3.9)	19.1 (3.6)	-1 (43)	[-3.60, 0.83]	.2
Symb. gesture	41.1 (6.9)	41.4 (7.4)	40.5 (6.4)	-0.4 (33)	[-5.40, 3.58]	.7

Table 3. Behavioural study: linear mixed effect model (LME) results testing the effects of number of objects during training and child vocabulary scores on caregiver deictic gesture use during training trials, and generalised estimated equation (GEE) results on the effects of number of objects during training and child vocabulary scores on child accuracy at test.

Dependent variables	Independent variables	*Estimate*	*SE*	*df*	*t*	*p-value*
Caregiver deictic gestures during training (LME)	(intercept – one object)	2.99	0.31	12.58	9.76	<.001
	One v. two objects	0.57	0.25	90.24	2.32	.023
	One v. six objects	0.76	0.25	91.79	3.08	.003
	Two v. six objects	0.19	0.25	93.35	0.77	.445
		Estimate	*SE*		*Wald*	*p-value*
Child testing accuracy (GEE)	(intercept – one object)	-1.76	0.66		7.05	.008
	One v. two objects	0.90	0.43		4.36	.037
	One v. six objects	0.85	0.46		3.32	.068
	Two v. six objects	-0.05	0.50		0.01	.921
	Receptive vocabulary	0.002	0.002		1.24	.265
	Caregiver deictic gesture	0.03	0.10		0.10	.749

Table 4. Behavioural study: linear mixed effects model results testing the effects of number of objects during training and child vocabulary scores on caregiver gesture and speech with gesture subtypes during training trials.

Dependent variable	Independent variables	*Estimate*	*SE*	*df*	*t*	*p*-value
Referent label use	(intercept – one object)	6.49	0.57	8.11	11.44	<.001
	One v. two objects	0.77	0.33	89.49	2.37	.020
	One v. six objects	-0.39	0.33	91.12	-1.18	.242
	Two v. six objects	-1.16	0.33	89.66	-3.52	<.001
Comp. speech with gesture	(intercept – one object)	0.43	1.04	41.57	0.41	.069
	One v. two objects	0.65	0.25	80.00	2.58	.012
	One v. six objects	0.43	0.25	80.00	1.68	.096
	Two v. six objects	-0.23	0.25	80.00	-0.89	.375
	Symb. gesture vocab	0.03	0.02	36.27	1.33	.193

comp. = complementary; symb. = symbolic; vocab = vocabulary

Endnotes

1. General linear mixed effects models (*glmer* package; *lme4* in R [v3.4.1, 2017]) were originally used but failed to converge, so GEEs were employed.