# Lexical Complexity of Child-Directed and Overheard Speech: Implications for Learning

**Ruthe Foushee (foushee@berkeley.edu)**
Department of Psychology, University of California, Berkeley
Berkeley, CA 94704

**Thomas L. Griffiths (tom_griffiths@berkeley.edu)**
Department of Psychology, University of California, Berkeley
Berkeley, CA 94704

**Mahesh Srinivasan (srinivasan@berkeley.edu)**
Department of Psychology, University of California, Berkeley
Berkeley, CA 94704

## Abstract

Although previous studies have found a link between the quantity and quality of child-directed speech learners receive and their vocabulary development, no previous studies have found a parallel link between overheard speech measured at a very young age and vocabulary development (Shneidman & Goldin-Meadow, 2012; Shneidman, Arroyo, Levine, & Goldin-Meadow, 2013; Weisleder & Fernald, 2013). This is despite the fact that children are able to learn words from overheard speech in laboratory settings (Shneidman & Woodward, 2015). Drawing on the idea that children preferentially attend to stimuli that are at a manageable level of complexity (Kidd, Piantadosi, & Aslin, 2012, 2014), the present research explores the possibility that children do not initially tune into overheard speech because it is initially too complex for their stage of lexical development (i.e., contains too great a proportion of unfamiliar words). Using transcripts from CHILDES and the Santa Barbara Corpus, and estimates of vocabulary by age from the MB-CDI, we find that child-directed speech is significantly less complex than overheard speech through at least 30 months. If attention based on complexity at least partially accounts for the statistical independence of overheard speech and vocabulary development in early childhood, then children might only begin learning from more complex, overheard speech sometime after 30 months.

**Keywords:** lexical development; attention; corpus analysis

## Introduction

In every study designed to investigate what constitutes effective input for language-learning, researchers find that child-directed speech reliably predicts children's vocabularies months later, while overheard speech does not (Shneidman & Goldin-Meadow, 2012; Shneidman et al., 2013; Weisleder & Fernald, 2013). Studies like these record a period of children's everyday linguistic input at 1.5–2.5 years, and relate the quantity (number of words) and quality (number of word types, or lexical diversity) of speech coded as child-directed or overheard to vocabulary assessed at 2.5–3.5 years. Surprisingly, the independence of overheard input and later vocabulary development persists even in contexts where the majority of children's input comes from overheard speech (Shneidman & Goldin-Meadow, 2012), which characterizes many cultural communities (Correa-Chávez & Rogoff, 2009; Ochs, 1982; Schieffelin, 1990; Ward, 1971). Also adding to this puzzle are indications that children *can* learn from overheard speech: preschoolers pick up profanity, and can learn new words in laboratory tasks by as early as 18 months (Akhtar, Jipson, & Callanan, 2001; Akhtar, 2005; Floor & Akhtar, 2006; Gampe, Liebal, & Tomasello, 2012; Martínez-Sussman, Akhtar, Diesendruck, & Markson, 2011; Shneidman, Buresh, Shimpi, Knight-Schwarz, & Woodward, 2009).

Outside of child-directed speech, recent work suggests that infants may selectively attend to input that is in an optimal, intermediate zone of complexity: neither too simple nor too complex (Kidd et al., 2012, 2014). In looking time studies with 7–8-month-olds, children's probability of looking away from a visual display is lowest when the visual or auditory events are within an intermediate range of predictability based on the preceding sequence of events. Infants are highly likely to look away both when the event is too predictable (e.g., identical to what had happened on the previous several trials), and when it is too unpredictable (i.e., a completely unexpected event). Such studies indicate that children may be implicitly tracking the complexity of different inputs and budgeting their attentional resources, attending only to those stimuli which are in a learnable range. From this work, it seems one of the many ways to capture children's attention is by providing stimuli that are neither too simple nor too complex, according to their knowledge state.

In this paper, we explore the hypothesis that children's implicit attention management might provide an explanation for the limited impact of overheard speech on language acquisition. Might overheard speech initially fail to predict vocabulary development because it is excessively complex, causing children to ignore it until they can learn from it? If so, the lack of a relationship between overheard speech and vocabulary outcomes might be due to how early previous studies took measures of each input type.

To explore this hypothesis, we examine whether we can account for findings about effective input for word-learning by defining an optimal window of complexity for an aspect of language acquisition. In particular, we focus on defining complexity for lexical development, since vocabulary was the outcome measure used in previous studies of overhearing.

While language input varies in complexity along many other dimensions (e.g., morphological, syntactic), for the present purposes, we will be focusing on the familiarity of words in the speech stream. Even so constrained, complexity is a moving target: as the abilities of the learner change, so does what she would consider unpredictable, or engaging. Our measure of lexical complexity therefore takes into account the expanding body of words in the child's vocabulary.

Our analysis uses tools from information theory and estimates of the current vocabularies of children at different ages to compute the complexity of the content of child-directed and overheard speech. Our results show that child-directed speech is consistently less complex than overheard speech through at least 30 months of age. These results provide support for the notion that children might be adaptively selecting their input by allocating attention to speech that falls into an intermediate zone of complexity given their current stage of lexical development.

## Goals in Relating Attention & Linguistic Input

Before presenting the methods and results of our analysis in detail, it is worth considering the goals of this investigation. There are at least two ways in which the speech adults direct to children might elicit more attention than speech they direct to other adults. First, across many cultures and contexts, speech to very young children takes a distinctive, exaggerated acoustic form (see Soderstrom, 2007, for a review). Current work is still elucidating the degree to which typical child-directed speech is tailored to children's learning at all levels of linguistic analysis (Eaves, Feldman, Griffiths, & Shafto, in press; Graf Estes & Hurley, 2013; Rafferty & Griffiths, 2012, *inter alia*). However, studies of listening preferences indicate that it at least captures children's attention (Werker, Pegg, & McLeod, 1994), a prerequisite for acquiring new word-to-meaning mappings (Graf Estes & Hurley, 2013; Ma, Golinkoff, Houston, & Hirsh-Pasek, 2011; Singh, Nestor, Parikh, & Yull, 2009). Second, adults may selectively use more words that are familiar to the child in child-directed speech than they would in adult-directed speech. This greater proportion of known words, and therefore decreased lexical complexity, may help capture the child's attention, and facilitate their learning of the words they don't yet know. The early disparity between the style and complexity of child-directed versus overheard input might then explain why previous studies have found an asymmetric relation between them and young children's vocabulary outcomes.

By examining the complexity of child-directed versus overheard speech as the child's word knowledge increases, we can begin to understand the ways in which the findings of studies relating input and vocabulary might be different (1) if the children were older, or (2) if they were exposed to overheard speech of lesser complexity. The hypothesis that young children's apparent lack of vocabulary learning from overheard speech is due at least in part to its greater complexity suggests that we might see a correlation between overheard input and vocabulary development in both these cases. To this end, we will investigate how input type (Child-Directed or Overheard), the child's own state of vocabulary knowledge, and learnability-driven attention might interact across lexical development. Taking into account the words familiar to the child, we approximate the relative lexical complexity of child-directed and overheard speech when the child is 12–30 months old.

In addition to validating the greater complexity of overheard speech, our results suggest an important role for the exaggerated style of child-directed speech in getting lexical development off the ground, at least in cultures where it is available. Children's earliest input is likely too full of unknown words to attract any attention based on manageable lexical complexity, but the acoustic profile of typical child-directed speech may maintain the attention necessary for their learning.

This investigation is preliminary, but the idea is that for a given child, with a given vocabulary, we might be able to estimate when she will begin to learn reliably from overhearing. That is, identify the point after which measures of *both* overheard and child-directed speech should predict vocabulary development. In reality, this will vary at least based on the vocabulary size of the individual child, and the typical complexity of her caretakers' adult-versus child-directed speech. Here, we first determine whether child-directed and overheard speech are generally differentiated in terms of lexical complexity across early development, using a complexity measure based on the proportion of words familiar to the average child. If overheard speech is more complex, and if adults calibrate the complexity of their child-directed speech as the target child matures, then when the complexity of the two input types are comparable, we would expect children to attend to overheard speech, and subsequently be able to learn from it. Estimating the beginning of this developmental window is our second goal.

## Method

### Child's Lexicon

The Macarthur Bates Communicative Development Inventory (MB-CDI) is a family of parental report vocabulary instruments for children ages 0–30, many of whose administrations are publicly accessible online (Frank, Braginsky, Yurovsky, & Marchman, under revision). To obtain a lexicon for a given age, we pulled the list of words reportedly comprehended or produced on 50% or more of the administrations archived on wordbank.com for children of that age, using the `wordbankr` package. We did this for children of each month is age from 12–30 months, the period across which we know from previous studies that child-directed, but not overheard, speech predicts later vocabulary development.

### Child-Directed Speech

We obtained corpora of speech directed to children of each of our 19 ages from the Child Language Data Exchange System

(CHILDES) Database (MacWhinney, 2000). The corpora were comprised of transcripts representing English-language, naturalistic (not activity-oriented) parent-child interactions at which no other adult or child was present. All files in the database that met these specifications were included for analysis, so the number of children and files for each age varied, from 869–91701 tokens directed to 2–27 target children of different households.

### Adult-Directed Speech

While it would be ideal to obtain samples of overheard speech from the same CHILDES transcripts, they contained little verifiably inter-adult speech. Therefore, transcripts meant to reflect the speech that children might overhear came from the Santa Barbara Corpus (Du Bois et al., 2000–2005), a database of transcribed audio recordings of American English conversations from diverse contexts and regions. All files which documented informal, in-person and monolingual conversations between young to middle-aged adults were included. This resulted in a set of 19 transcripts, and a total of 87,496 words.

### Complexity Measures

All morphological roots and their frequencies were obtained for each set of transcripts using the *freq* command in CLAN. For each age and input type, 1,000 words were randomly sampled from the total set, using its frequency distribution. Following Kidd, Piantadosi, and Aslin (2012), a complexity measure for each sample of speech was calculated by taking the negative log probability of the words in the child's lexicon in that sample. For an example, say 600 of the 1,000 words in a sample of 18-month-old-directed speech were among those 310 words reported as known by at least half the 18-month-olds on the MB-CDI. The proportion of known words would be $(600/1000)$, or 0.60. The complexity measure for that sample for that age would be $(-\log(0.60))$, or 0.51. This measure is intended to reflect the density of novel words via its complement: the density of known words. Therefore, as the proportion of words children of a given age *already* know in a sample increases, complexity decreases.

Each child- and adult-directed corpus was sampled 100 times for each age, and the complexity of each sample relative to the average child's lexicon at that age was calculated. We expect the complexity of both child-directed and overheard speech to decrease with age as children's vocabularies increase. Beyond that, we can get a sense of whether parents modify their speech to be less complex for children, which may or may not occur across our age span.

To know this, we would want to compare the mean relative complexity of the child- and adult-directed speech samples at each age, with the prediction that child-directed speech should always be significantly lower. We are also interested in the rates at which the two input types decrease, and whether there is an age at which their complexity scores become comparable.

## Results & Discussion

### Complexity of Child-Directed vs. Overheard Speech

As predicted, mean complexity of adult-directed speech (ADS) was significantly greater than that of child-directed speech (CDS) at all ages except one, 21 months (paired t-tests, all significant at $p < 0.001$ Bonferroni-corrected for multiple comparisons). As can be seen in Figure 1, the complexity of child-directed speech is significantly *greater* than adult-directed speech at 21 months ($t(99) = 41.86$, $p < 0.001$), which given the similarity in complexity of 21-month "overheard" speech, we expect to be an artifact of the specific CHILDES samples used. The child-directed speech at this age came from 23 transcripts spanning 10 children in a variety of situations, which may have contributed to the greater number of unknown words, though this is the case at many of our ages. Another curious trend is the apparent short-term increase in complexity of both input types after 18 months. As the general relationship between overheard and child-directed speech is maintained during this period, we speculate that the cause may be related to our vocabulary measure, though this and the 21-month-old data will be points for future investigation.

These abnormalities also serve to highlight how remarkable it is that child- and adult-directed speech were consistently different in this analysis, given the variability in the transcripts from which the speech samples were pulled, and the general sparsity of the data at each age. The persistent difference between the two input types suggests adults do in fact adjust their vocabulary based on whether they are speaking to a child or fellow adult. Whether they do so appropriately, that is, whether they continuously calibrate their speech as the child ages, we will explore in future analyses.

Of course the words on the MB-CDI do not represent all the words children know, and the complexity range, particularly for child-directed speech, is therefore likely systematically overestimated. The undoubtedly frequent occurrence of the child's own, known name, along with other household-specific lexical items in speech likely directed more often to her than to another adult, were completely ignored.

### Modeling Complexity Trajectories

The ages for which the above analysis was possible were constrained by the vocabulary measure we used, so the dataset will necessarily be expanded in future work. For each input type, we fit an exponential function with the complexity score as our dependent variable, and age as our independent variable. The function was selected on a primarily theoretical basis, based on the shape we expected the relationship between complexity and age to take. Choosing this form necessarily means the two models will converge to zero, however, so we speculate on their relationship past the age of 30 months with caution. We were interested in predicting the age at which the two input complexities would cease to be significantly different, with the hypothesis that this might be when we would expect children to regularly learn from overheard

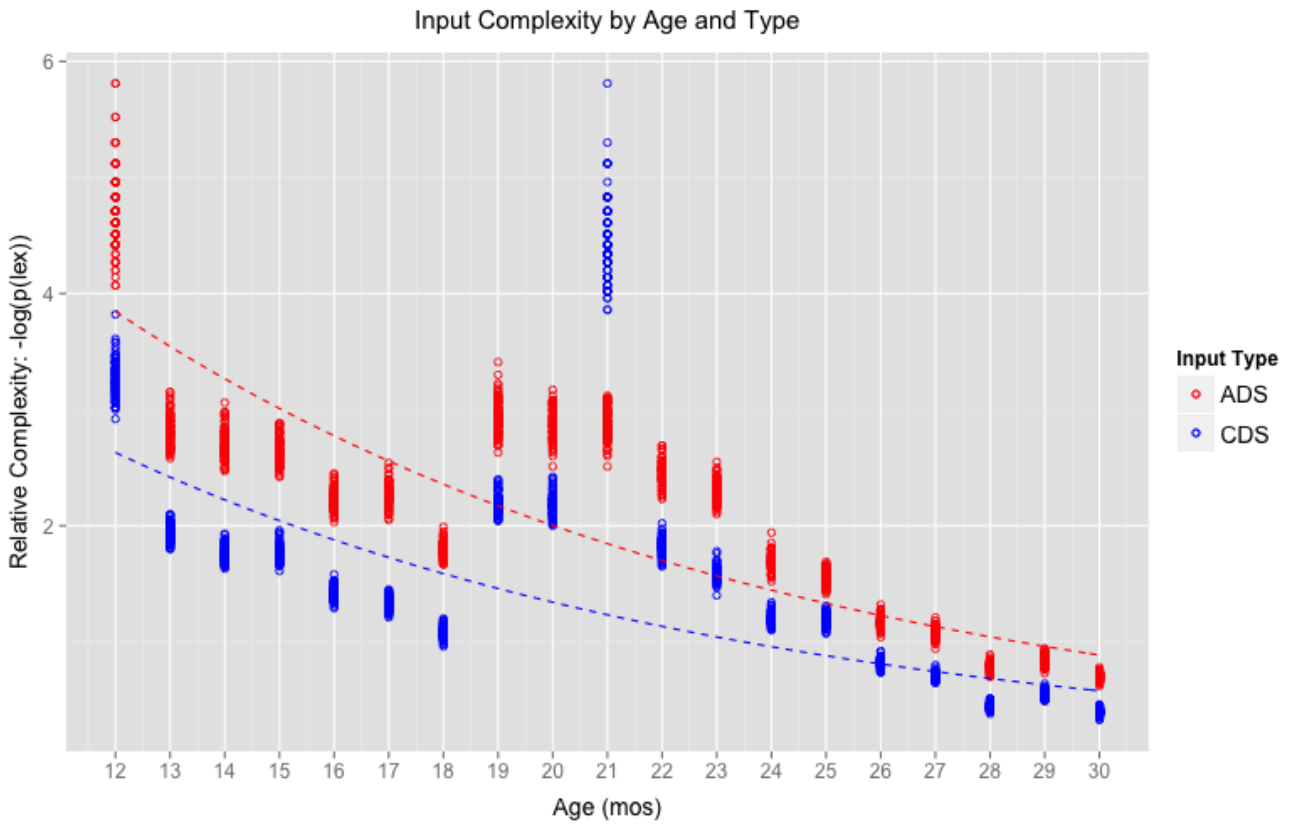


Figure 1: Measure of complexity relative to median child's lexicon at four time points in Child-Directed and Overheard, or Adult-Directed, Speech. Exponential curves for each input type are shown with dotted lines.

Table 1: Coefficients and statistics for exponential regressions ($Y \sim A * \exp(-B * X)$) predicting speech complexity from age for each input type.

| Input | $A$ | $B$ | $R^2$ | $SE$ |
|---|---|---|---|---|
| CDS | 1.98 | −0.08 | 0.69 | 0.32 |
| ADS | 2.33 | −0.08 | 0.79 | 0.24 |

speech. Looking at Figure 1, we might expect this milestone sometime after the child's third birthday, and certainly not before 30 months.

Our longitudinal analysis of relative complexity sheds light on the puzzle we started with, as well as the trajectory of lexical development. As demonstrated in empirical studies, children are capable of learning from overheard input, but may not recruit this skill typically because the overheard speech they hear is too complex to elicit their attention. While there has been some effort to alter the syntax in which new words are presented in typical overhearing experiments, the scripts for the overheard conversation are designed so that the novel noun the child is supposed to learn is the only unfamiliar word in the dialogue (Akhtar et al., 2001; Akhtar, 2005; Floor & Akhtar, 2006; Gampe et al., 2012; Martínez-Sussman et al., 2011; Shneidman et al., 2009). Thus, the speech is extremely simplified, and likely at or below the typical complexity of speech directed to children of the participants' ages.

The analysis also provides a new way of looking at child-directed speech. Is attention and an exaggerated acoustic style necessary for the link between child-directed speech and vocabulary development, or is the high proportion of familiar words largely responsible?

During the ages we studied, adult-directed speech was always more complex than child-directed speech, indicating that adults do selectively use words that are accessible to their young interlocutors. However, even the reduced vocabulary apparently used for very young audiences appears highly complex given the lexicon of the typical 12-month-old. This may be where the acoustic profile of child-directed speech critically compensates for unavoidably high lexical complexity in maintaining the child's attention.

By the second year of life, the relative complexity of both types of input fall rapidly as the child's vocabulary increases, and appear to approach each other. That the two are still not equivalent by the age when studies have failed to link overheard input and vocabulary development suggests complexity differences may indeed help account for that finding.

## Future Directions

Our complexity measure both gives us leverage on the learning asymmetry between overheard and child-directed speech, and introduces further questions to address. For example, because complexity ratings represent the interaction between adults' choice of vocabulary and children's own lexicons, we can't know how much their rates of decline across time are due to adults' appropriate calibration to the child's knowledge state versus the child's own expanding vocabulary.

In addition, we don't know how much inattention to overheard speech might be influenced by *preference* for child-directed-speech. Do the levels of lexical complexity we found for child-directed speech reflect something universal about children's information processing capacities, or would the 'optimal' level of complexity to attend to and therefore learn from be higher if child-directed speech were unavailable? This question points to the need for a similar analysis in cultures where child-directed speech is not a primary source of children's linguistic input.

### Alternative Measures & Sources

There are several methodological alternatives we might consider to complement the current study. To address the limitations of our vocabulary measure, we might conduct the same analysis using another vocabulary measure with more longevity, like the Peabody Picture Vocabulary Test (Dunn & Dunn, 2007). Older children's lexicons might also be approximated using age of acquisition ratings (Kuperman, Stadthagen-Gonzales, & Brysbaert, 2012), scaled with reference to established norms for a given age range. Extending the analysis to later ages is vital to determining if and when child- and adult-directed speech converge in lexical complexity.

Some of the uncertainties remaining in the current analysis might be resolved by limiting our complexity calculation to longitudinal datasets, which would also enable us to more accurately approximate the child's vocabulary development via her production. Others can only be addressed with longitudinal data that capture both the speech directed to a child, and the speech that same child has the opportunity to overhear.

Alternative complexity measures capturing the same idea are also valuable to explore. In our measure, for example, samples with the same proportion of known words, but different levels of diversity of unknown words, are equivalent. This does not reflect the intuition that a sample of speech which contains 600 words you know, but 400 *different* words you don't should be more complex than a sample with 400 of the same, unknown word. Thus, a measure which incorporates the entropy of the unknown words in a sample may even better reflect lexical complexity. Studies interested in the same question may choose to model complexity instead using word-by-word predictability in the input, sampling continuous segments of speech rather than random tokens, and judging complexity based on transitional probability.

Information about optimal levels of complexity might also be informed by children's book standards. Children's books typically have a recommended audience age. It might be informative to calculate our complexity measure for a corpus of children's books either recommended for a given age, or reported by teachers and parents to enthrall children of a given age, and compare it to the age-typical complexity of informal speech input. Previous studies have ascertained that children's books have greater lexical diversity than child-directed speech (Jones & Smith, 2015), but our measure would enable us to assess how calibrated storybooks are for vocabulary learning at different ages.

### Experimental Predictions

The current analysis gives us a complexity range for lexical development that can be tested in the lab. If a period of overheard speech were crafted to receive a complexity rating of approximately 0.33 (the mean complexity of child-directed speech at 30 months), could we get a 30-month-old attending to, and consequently learning from, it? What about a younger child with a large vocabulary? The thus-far neglected aspect of the findings linking attention and learnability—that children should also not attend to overly simple input—can now be investigated as well. Will children preferentially attend to slightly more complex, but learnable, overheard speech, compared to completely predictable child-directed speech? Studies with infants terminate the experiment when they look away, so we don't know how children manage their attention when a stimulus continues. We also can't know without empirical support whether our sample size of 1,000 words is a reasonable window in which to assess complexity, and how that window changes as children's attentional resources increase. If children stop tuning in when speech becomes excessively complex, can their attention be regained later on?

Trade-offs between complexity and speech style should be systematically investigated at different ages and vocabularies, along with ways of encouraging learning from indirect speech beyond manipulating lexical complexity. If learning from overhearing is driven at least in part by attention, then we should see learning of a new word by a pig-loving preschooler from a complex overheard dialogue about pigs before learning from a dialogue about architecture of the same complexity. Syntactic complexity could be manipulated as well: unknown words occurring in high proportion in familiar syntactic frames might be easier to learn than words in unfamiliar frames at the same density.

As we move forward in this investigation, it is critical to acknowledge that there are other ways in which child-directed and overheard speech diverge. While children have been shown to be able to infer the meaning of a new word via its discourse context (Sullivan & Barner, 2015), learning new words from overhearing might remain more difficult than from child-directed speech—even at the same complexity level—merely because referents in parent-child interaction are qualitatively different and easier to identify. We can't know how much explanatory power the idea of lexical development based on input complexity might give us, however,

until we expand our analysis to further vocabulary measures, transcripts, and methods.

## References


Akhtar, N. (2005). The robustness of learning through overhearing. *Developmental Science*, *8*(2), 199–209.

Akhtar, N., Jipson, J., & Callanan, M. (2001). Learning words through overhearing. *Child Development*, *72*(2), 416–430.

Correa-Chávez, M., & Rogoff, B. (2009, May). Children's attention to interactions directed to others: Guatemalan mayan and European American patterns. *Developmental Psychology*, *45*(3), 630–641.

Du Bois, J. W., Chafe, W. L., Meyer, C., Thompson, S. A., Englebretson, R., & Martey, N. (2000–2005). *Santa Barbara corpus of spoken American English, Parts 1–4.* Philadelphia: Linguistic Data Consortium.

Dunn, L. M., & Dunn, D. M. (2007). *PPVT–4: Peabody picture vocabulary test*. Minneapolis, MN: Pearson Assessments.

Eaves, B. S., Feldman, N. H., Griffiths, T. L., & Shafto, P. (in press). Infant-directed speech is consistent with teaching. *Psychological Review*.

Floor, P., & Akhtar, N. (2006). Can eighteen-month-olds learn words by listening in on conversations? *Infancy*, *9*, 327–339.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (under revision). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*.

Gampe, A., Liebal, K., & Tomasello, M. (2012). Eighteen-month-olds learn novel words through overhearing. *First Language*, *32*(3), 385–397.

Graf Estes, K., & Hurley, K. (2013). Infant-directed prosody helps infants map sounds to meanings. *Infancy*, *18*(5).

Jones, J. L. M. M. N., & Smith, L. B. (2015). The words children hear: picture books and the statistics for language learning. *Psychological Science*, 1–8.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS One*, *7*(5).

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2014). The Goldilocks effect in infant auditory attention. *Child Development*, *85*(5), 1795–1804.

Kuperman, V., Stadthagen-Gonzales, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, *44*(4), 978–990.

Ma, W., Golinkoff, R. M., Houston, D. M., & Hirsh-Pasek, K. (2011). Word learning in infant- and adult-directed speech. *Language Learning and Development*, *7*(3), 185–201.

MacWhinney, B. (2000). *The Database* (3rd ed., Vol. 2). Mahwah, NJ: Lawrence Erlbaum Associates.

Martínez-Sussman, C., Akhtar, N., Diesendruck, G., & Markson, L. (2011). Orienting to third-party conversations. *Journal of Child Language*, *38*(2), 273–296.

Ochs, E. (1982). Talking to children in Western Samoa. *Language in Society*, *11*, 77–104.

Rafferty, A. N., & Griffiths, T. L. (2012). Optimal language learning: The importance of starting representative. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Schieffelin, B. (1990). *The give and take of everyday life: Language socialization of Kaluli children*. Cambridge: Cambridge University Press.

Shneidman, L. A., Arroyo, M. E., Levine, S., & Goldin-Meadow, S. (2013). What counts as effective input for word learning? *Journal of Child Language*, *40*(3), 672–686.

Shneidman, L. A., Buresh, J. S., Shimpi, P., Knight-Schwarz, J., & Woodward, A. L. (2009). Social experience, social attention and word learning in an overhearing paradigm. *Language Learning and Development*, *5*, 266–281.

Shneidman, L. A., & Goldin-Meadow, S. (2012). Language input and acquisition in a Mayan village: How important is directed speech? *Developmental Science*, *15*(5), 659–673.

Shneidman, L. A., & Woodward, A. L. (2015). Are child-directed interactions the cradle of social learning? *Psychological Bulletin*.

Singh, L., Nestor, S., Parikh, C., & Yull, A. (2009). Influences of infant-directed speech on early word recognition. *Infancy*, *14*, 654–666.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, *27*(4), 501–532.

Sullivan, J., & Barner, D. (2015). Discourse bootstrapping: Preschoolers use linguistic discourse to learn new words. *Developmental Science*, *19*(1).

Ward, M. (1971). *Them children: A study in language*. New York: Holt, Rinehart, and Winston.

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143–52.

Werker, J. F., Pegg, J. E., & McLeod, P. J. (1994). A cross-language investigation of infant preference for infant-directed communication. *Infant Behavior and Development*, *17*(3), 323–333.