GLOSSA psycholinguistics

Rivera-Vera, N., Andringa, S., Kronmüller, E., Monaghan, P., & Rispens, J. (2022). The effect of speaker reliability on adult cross-situational word learning. *Glossa Psycholinguistics*, 1(1): 7, pp. 1–33. DOI: https://doi.org/10.5070/G601113

eScholarship University of California

The effect of speaker reliability on adult cross-situational word learning

Natalia Rivera-Vera, Amsterdam Center for Language and Communication, University of Amsterdam, the Netherlands, n.a.riveravera@uva.nl

Sible Andringa, Amsterdam Center for Language and Communication, University of Amsterdam, the Netherlands, s.j.andringa@uva.nl

Edmundo Kronmüller, Escuela de Psicología Pontificia Universidad Católica de Chile, Chile, ekr@puc.cl

Padraic Monaghan, Department of Psychology, Lancaster University, United Kingdom, p.monaghan@lancaster.ac.uk

Judith Rispens, Amsterdam Center for Language and Communication, University of Amsterdam, the Netherlands, j.e.rispens@uva.nl

Word learning is guided by the statistical co-occurrence between spoken words and potential referents, through which learners gradually map labels to objects across situations. Given that word learning does not occur in a vacuum, rather in a communicative context, it is relevant to evaluate the role that speakers play. Because we do not evaluate the information provided by every person equally, it is reasonable to think that someone who makes lexical errors is not a reliable speaker from whom to learn new words. The current study focuses on speaker reliability in adult cross-situational word learning (CSWL). In two experiments we investigated the extent to which adults attend to the reliability of the speaker and how this affects word learning in a CSWL task. We varied the consistency with which a speaker mapped novel words to familiar objects. We hypothesized (1) that the speakers' reliability would be judged differently depending on their past object-labeling accuracy, and (2) that new words would be more difficult to learn when presented by an unreliable speaker. Experiment 1 shows that the unreliable speaker was assessed as less reliable, compared to the reliable speaker, but this effect disappeared in Experiment 2, when participants were taught new words by two speakers, a reliable and an unreliable one. Furthermore, we found no evidence to support the hypothesis that being exposed to an unreliable speaker impairs CSWL in adults. We discuss the relevance of these findings and the importance of further research on the role of speaker reliability in CSWL.

1 Introduction

How do learners learn to map a word to an object successfully? Social and pragmatic approaches to word learning argue that the word learning process is inherently social (e.g. Baldwin 1993; Tomasello 1992; Tomasello and Akhtar 1995; Tomasello 2000). The potential word meaning considered by a learner is constrained by the situation at hand, specifically by those aspects relevant to the communicative context. Pragmatic and social cues, fundamental to any communicative interaction, play a major role in guiding learners' attention to the multiple potential meanings of a word (Baldwin 1993; Tomasello and Akhtar 1995). For example, research on early word learning has shown that infants and children consider the intention of their interlocutors (Baldwin et al. 1996; Bloom 2000), or their reliability as a source of information when learning new words, which may skew the processes through which learners determine the meaning of a word (Lev-Ari 2015; Mills 2013). Moreover, socio-communicative cues like eye-gaze can influence the representations adult learners store during word learning across different situations (MacDonald et al. 2017); specifically, the number of potential referents for a given word.

At the same time, human learners are capable of detecting and exploiting the statistical properties of their environment, a phenomenon known as statistical learning (Aslin et al. 1998). Associative learning approaches argue that what guides learners to gradually learn new words are the co-occurrences between spoken words and potential referents, by means of which learners map new labels to referents across multiple communicative contexts (Gleitman 1990; Pinker 2013; Siskind 1996; Smith 2000; Yu and Smith 2007). In spite of the intrinsic referential uncertainty of the environment (Quine 1960), individuals are thus able to learn new words by using a cognitive system that is highly sensitive to the distributional co-occurrence between words and objects (Smith and Yu 2008; Suanda et al. 2014; Vouloumanos 2008).

Given the paucity of research on the effects of socio-communicative and pragmatic information on adult word learning, the two experiments presented here investigate the extent to which characteristics of a speaker influence adult cross-situational word learning. Specifically, we investigated to what extent the reliability of the speaker, i.e. the consistency with which they label objects across situations, affects cross-situational word learning in adults. In light of the evidence supporting the importance of speaker characteristics on early word learning, the primary objective of this study is to provide evidence for the role that speaker reliability may play in cross-situational word learning (e.g. Frank et al. 2009; Frank and Goodman 2014; MacDonald et al. 2017; Yu and Ballard 2007; Yurovsky 2018).

1.1 Cross-situational word learning

Cross-situational word learning (CSWL, also referred to as cross-situational statistical learning) has been proposed as the strategy learners use to track the co-occurrences between a word and

an object (Siskind 1996; Yu and Smith 2007). Learners determine the correct meaning of a word by aggregating its co-occurrences with an object across multiple exposures (e.g. Smith et al. 2011; Smith and Yu 2008; Yu and Smith 2007).

The hypothesis that word learning is based on tracking the cross-situational statistics of the environment, is usually tested in the following way (Yu and Smith 2007): Participants are exposed to a series of learning trials, each presenting an equal number of unknown objects and novel words (e.g. 2, 3 or 4, depending on the condition). Multiple words and potential referents are presented in a single trial. Every time participants hear a word, its corresponding referent is present, but on an individual trial it is not clear which label refers to which referent. This may lead to confusion about the association between a word and an object within a single trial. However, the co-occurrence between a particular word and a particular object across trials (situations), in addition to different sets of foils accompanying the target object across trials, enables inferring the label of an object. After this learning phase, participants are tested on an alternative forced-choice task – they are presented with one word and a minimum of two objects, and are asked to choose the object being referred to by the word. Yu and Smith (2007) found that adult learners are remarkably sensitive to the cross-situational regularities, in spite of the ambiguity of the learning environment. This finding has been replicated in similar studies on both adults and children, and in both noun and verb learning (e.g. Monaghan et al. 2012; 2015; Smith et al. 2011; Smith and Yu 2008; Smith et al. 2011; Suanda et al. 2014; Vouloumanos 2008).

Even though these studies show that cross-situational statistical information guides word learning, they do not consider factors that are inherent to the communicative situation outside a lab setting, like the characteristics of the speaker. In this regard, it is reasonable to ask to what extent learners are influenced by the reliability of their informants. If the cross-situational statistics presented are inconsistent, rendering the speaker an unreliable source, to what extent do learners take this information into account? Furthermore, does an unreliable speaker affect cross-situational word learning?

1.2 A social approach to cross-situational word learning

Current approaches to CSWL study how the learner's knowledge of the goals, intentions, and general characteristics of a speaker is integrated with the environment's perceptual information (i.e. statistical cues) the learner has to process during word learning (e.g. Frank et al. 2009; Frank and Goodman 2014; MacDonald et al. 2017; Najnin and Banerjee 2018; Yu and Ballard 2007; Yurovsky 2018; Yurovsky and Frank 2017). These so-called hybrid CSWL accounts (Roembke and McMurray 2016), which take into account social theories of early word learning in children (e.g Baldwin 1993; Bloom 2000; Tomasello 1992; 2000), can be described as unifying

approaches to word learning, since they investigate both the statistical regularities of word-object co-occurrences, and the social cues intrinsically encoded in the word learning process. Even though little research has been carried out to test hybrid CSWL approaches on adults, some studies have set the stage for further research on this topic.

With the aim to test whether contextual information modulates CSWL in adults, Poepsel and Weiss (2014) examined the role of the speaker's identity. The authors predicted that the speaker's identity (male vs. female speaker) would facilitate word learning because it would aid learners to make specific associations between a particular speaker (e.g. a female speaker) and the word-object pair they were supposed to learn from her. However, no statistically significant effect of speaker identity was found when participants were tested on a 4-AFC task. Based on these results, Poepsel and Weiss argue that the unambiguity of the target mapping during the familiarization phase may have prompted learners to focus on the word-object co-occurrence only, rather than on the social cue associated to it (i.e. the speaker's identity). Following this reasoning, Poepsel and Weiss' participants may have figured out that, even though the identities of the speakers were a clear cue, the perceptual information (i.e. the word-object co-occurrences) was sufficient to solve the task at hand. In light of this, the extent to which the speakers' identity – operationalized in this study as a gender difference – can instantiate a clear social or pragmatic cue in a CSWL context deserves further evaluation.

In another series of experiments that investigated the relation between social and statistical cues, MacDonald et al. (2017) tested how adults represent the word-object co-occurrences during CSWL depending on the gaze of the speaker (i.e. a social-communicative cue). Specifically, they asked whether the presence and reliability of the speaker's gaze in a CSWL trial modulate whether learners store one or many potential referents for a given word across situations. In Experiment 3, they manipulated the reliability of the speaker's gaze, operationalized as the consistency with which the speaker looked at the same object displayed on the computer screen across two trials (Exposure and Test). If the speaker's gaze was a reliable cue, then the object she was 'looking at' during Exposure trials was present in Test trials (Same condition). If gaze was unreliable, then the gazed object in the Exposure trials was *not* present in Test trials (Switch condition). The authors hypothesized that learners exposed to a higher number of Same trials would infer that the speaker's gaze was a reliable cue, leading them to store fewer referents for a given word. Conversely, learners exposed to a higher number of Switch trials would infer that the gaze cue was less reliable, prompting them to store more possible mappings in order to learn the correct word-object mapping. The results, though a small effect, confirmed the authors' hypothesis, suggesting that when the speaker's gaze is a more reliable cue, learners show a behavior consistent with using a fast-mapping mechanism (e.g. Trueswell et al. 2013). When

the gaze cue is more unreliable, instead, learners store more referents, a behavior consistent with an associative mechanism of word learning (e.g. Smith et al. 2011). MacDonald et al. (2017) conclude that social cues modulate the mechanisms involved in CSWL, suggesting that the representations generated by this process are rather flexible.

The study of socio-communicative cues on CSWL is motivated (as in MacDonald et al. 2017) by a growing body of experiments on speaker characteristics' effects on word learning in preschoolers. These studies have shown that children are more inclined to learn new words from reliable speakers (i.e speakers presented as knowledgeable, certain and accurate with respect to their linguistic competence) than from unreliable speakers (i.e. speakers portrayed as ignorant, uncertain and inaccurate) (e.g. Brosseau-Liard et al. 2014; Buac et al. 2019; Jaswal and Neely 2006; Koenig et al. 2004; Koenig and Harris 2005; Scofield and Behrend 2008; Sobel et al. 2012). Crucially, in most of these studies the speakers' epistemic state is explicitly exposed, either by making them admit they are (un)certain about the correct word-object mapping of familiar words and objects (e.g. Sabbagh et al. 2003), or by having two speakers use the same label for two different objects (e.g Scofield and Behrend 2008).[1]

In light of these findings, speaker reliability presents itself as a socio-communicative cue that relates to a pragmatic principle of language conversation central to language learning, namely, Cooperation (Clark 2018; Grice 1989). In the context of word learning, the speaker's goal – e.g. to teach someone a new word – guides their choices on how and what to say, as well as on what they expect the learner to acquire. Assuming that the speaker is collaborative, the learner, in turn, develops certain expectations with regard to the speaker (e.g. that they should know the names for certain objects). In the context of word learning, learning from someone who labels objects inconsistenly (i.e. an unreliable speaker) should, in principle, disrupt those expectations and affect the learning process. Given the evidence suggesting that social-communicative cues have an effect on CSWL (e.g. MacDonald et al. 2017), it is reasonable to assume that cues pertaining the reliability of the speaker as an informant may also affect CSWL. Since both children and adults are faced with the indeterminacy of the referent when learning new words (Verga and Kotz 2013), it seems plausible to consider that adults, as children, also attend to speaker characteristics when learning new words. In the present study, we tested (1) the extent to which speaker reliability is detectable from the cross-situational information presented by the speaker, and (2) to what extent speaker reliability affects CSWL in adults.

[1] It is worth noting that, even though children prefer to learn novel words (e.g. fep or dax) from knowledgeable and reliable speakers, learning a new word initially presented by an ignorant speaker does not seem to affect subsequent word learning, provided another speaker introduces this word (Sabbagh et al. 2003).

2 The present study

The aim of the present study was to investigate the extent to which adult learners perceived the reliability of the speaker and how this affected their cross-situational word learning. In the following two experiments we manipulated the reliability of the speaker – which depended on how consistently (i.e. accurate) words were mapped to objects across situations – and measured participants' learning of new words in a CSWL task. While the first experiment required the learner to respond to the reliability of a speaker, the second required them to respond to the reliability of two speakers.

Two research questions were tested in this study. First, we asked to what extent participants judged the speaker's reliability differently depending on the consistency with which the speaker labeled objects across situations. This allowed us to investigate the degree to which participants had taken into account the reliability of the speaker, and developed different perceptions towards them as a source of information. We expected participants in the Unreliable Speaker condition to asses the speaker as less reliable compared to participants in the Reliable Speaker condition. We also asked the extent to which the reliability of the speaker impacts on adults' ability to learn words across-situations. We hypothesized that if participants are sensitive to the reliability of the speaker, then their word learning would differ depending on the reliability of the speaker they were exposed to. Thus, we expected a lower word learning accuracy across blocks in participants in the Unreliable Speaker condition compared to participants in the Reliable Speaker condition. Furthermore, we expected to see the same pattern for those participants who rated the speaker as less reliable, compared to those who rated them as more reliable.

3 Experiment 1

3.1 Methods

3.1.1 Participants

Sixty Dutch-native speakers, aged between 19 and 35 years old ($M = 26.6$, $SD = 4.14$; 39 female), participated in the experiment for payment of €5. All participants had normal or corrected-to-normal hearing and sight. Participants were randomly assigned to one of the two experimental conditions, with both conditions having the same number of participants (Reliable Speaker condition: $M = 27.16$, $SD = 4.37$ years old; 18 female; Unreliable Speaker condition: $M = 25.5$, $SD = 3.39$ years old; 21 female).

3.1.2 Materials and design

Sixteen word-object pairs were created for this study (see Supplementary material 1). All words were CVCV noncewords (e.g. pabe, modi, gade), extracted from the Novel Object and Unusual Name (NOUN) Database (Horst and Hout 2016), and recorded by a male native speaker of

Dutch in a soundproof room. The objects represented concrete, inanimate nouns (see **Figure 1**), and were taken from the Multilingual Picture (MultiPic) databank (Duñabeitia et al. 2018), a standardized set of drawings normed for name agreement and visual complexity in six European languages. The selection criterion was to choose objects with the same level of agreement in labeling a given object in Dutch (i.e. with an H index of 0). This means that these objects received the same name across Dutch participants in Duñabeitia et al. (2018)'s study.

We adapted the classical CSWL paradigm (Yu and Smith 2007) by merging and distributing the exposure and testing phases across four learning blocks (for a similar design, see Monaghan et al. (2012)). This design allowed us to see how participants learn *while* they perform the task. Additionally, by having two objects and only one label being presented, we introduced situations in which not all objects being perceived are referred to, which better resembles what occurs in naturalistic word learning than studies where all objects are labeled in each situation (as in Yu and Smith (2007)). Furthermore, we manipulated the consistency with which the word-object mappings were presented across situations (i.e. trials), resulting into two types of co-occurrences: reliable and unreliable. In the former, a word was consistently mapped to one unique object (**Figure 1**, upper row); whereas in the latter, a word was mapped to four different objects across situations, making the mapping inconsistent (**Figure 1**, bottom row). We provide a 16 × 16 matrix that illustrates this manipulation for each type of mapping in the Supplementary material 2.

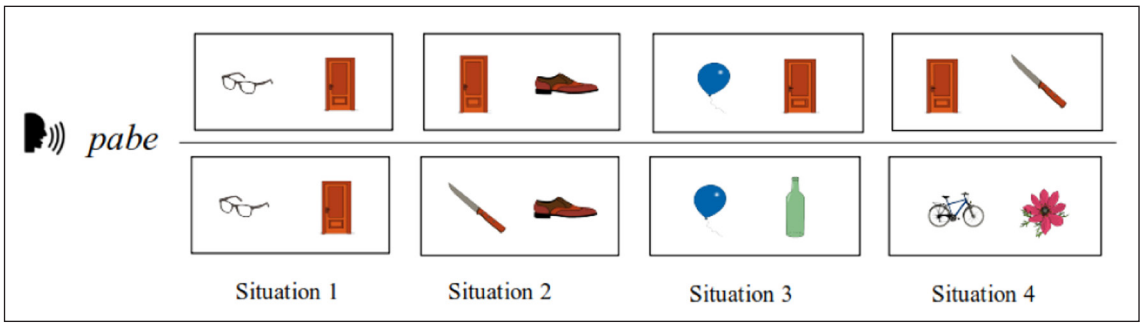

Figure 1: Illustration of an item for the pabe-door word-object pair. The upper row depicts the reliable mapping, whereas the bottom row corresponds to the unreliable mapping.

We coupled these two types of mappings with a specific speaker, introduced to participants as someone who would teach them new words. This coupling yielded our two experimental conditions, Reliable Speaker and Unreliable Speaker. In the Reliable Speaker condition, the speaker taught only reliable mappings to the learner. Hence, participants in this condition were exposed to 16 word-object pairs that were always consistently mapped – each of the 16 objects consistently co-occurred with one of the 16 words across trials. This means, for example, that

upon hearing a word, its corresponding target object was always present across trials. The conditional probabilities of word-object co-occurrences for this condition are shown in **Table 1**.[2]

SPEAKER RELIABILITY	STIMULI				CONDITIONAL PROBABILITY	
	Number of items	**Type of mapping**	**Auditory (A)**	**Visual (V)**	***p* (V|A)**	***p* (A|V)**
Reliable	16	reliable	word	target	1	0.5
			word	foil	0.14	0.07
Unreliable	8	reliable	word	target	1	0.5
			word	foil	0.14	0.07
	8	unreliable	word	target & foil	0.25	0.125

Table 1: Conditional probabilities per Speaker Reliability condition.

In the Unreliable Speaker condition, in turn, the speaker taught both reliably and unreliably mapped items. Here, eight out of the 16 word-object pairs were inconsistently mapped – i.e. a given word was mapped to eight different objects across trials, as is shown in **Table 1**, bottom row. This translates into a conditional probability of 0.25 for the word and both its target and foil object. This manipulation rendered these eight items unlearnable, and as such there was no correct answer for them.

The experiment comprised a set of four blocks of 64 trials each. Within each block, each word was heard four times. Each of the 16 objects was presented eight times within a block, four times as target and four times as foil (for the reliably mapped items). Objects were randomly paired with each of the 16 words once the experiment started, yielding unique word-object pairs per participant. The auditory words were recorded by a male speaker.

In addition to the CSWL task, and with the aim to estimate learners' perception of the speaker's reliability, we designed a 7-point Likert-type Subjective reliability questionnaire that required participants to assess the speaker. Thus, participants had to estimate how much they agreed or disagreed with the following statements about the speaker (1 = strongly disagree, and 7 = strongly agree):

- His voice sounded very calm
- His intelligibility was very good
- His knowledge of the words was very good
- He can reliably teach the new words to someone else

[2] $p(V|A)$: conditional probability of an object co-occurring with a given word; $p(A|V)$: conditional probability of a word given the object.

Whereas the first two statements were used as filler questions about the speaker's quality of voice and articulation, the last two assess the speaker as a source of information, specifically, his history of (in)accuracy in providing labels for objects, and his reliability as an informant (e.g. Koenig and Harris 2005; Scofield and Behrend 2008). The last two items were used to construct a reliability score, which we describe below (section 3.1.4).

3.1.3 Procedure

The experiment was carried out in a quiet area of a public space, or a laboratory, depending on the participant's preference. After filling in the questionnaire, reading the information brochure, and signing the consent form, the experiment started. Participants were instructed to sit in front of a laptop computer and wear headphones. They were randomly assigned to one of the two experimental conditions.

Before the first block started, participants were told they would learn words in a new language, and that the speaker they would hear, named Tijmen, was a second language learner of the language they were about to learn. Additionally, we included questions about the speaker in between blocks (see Supplementary material 3). Participants of both conditions were presented with the same information and questions about the speaker. These questions did not require an answer from the participants, rather they were used to prompt participants' perception of the speaker, which would be assessed after the CSWL task. Importantly, no explicit information about the speaker's reliability was provided to participants during the experiment. The objective of this was twofold. First, to determine if speaker reliability could be detected from the speaker's language use itself; and second, to make the experiment more ecologically valid. The latter stems from the observation that overt signaling of speaker unreliability is typically unavailable in real-life contexts (Gardner et al. 2021).

On each trial, two objects were displayed on the laptop screen and, after 200 ms, a word was played. Participants were instructed to choose the object they thought the word was referring to. They had to press either 'Z' or 'M' to choose the object on the left or right side of the screen, respectively. This instruction was repeated at the beginning of each block. No feedback was given after participants had responded. Once they had selected an object, the next trial started 500 ms after. Trial order presentation was pseudo-randomized within blocks, such that none of the words was heard in two consecutive trials. The target and foil objects occurred equally often on the right and left side of the screen within each block.

After each block participants had the opportunity to take a short break. Once they finished the CSWL task, they filled out the subjective reliability questionnaire (in writing format) and signed the payslip. After that, the experimenter explained the purpose of the study. Each testing session lasted approximately 30 minutes.

The experiment was programmed using PsychoPy2 Experiment Builder (v1.85.3) (Peirce et al. 2019), and was run on a Windows 7 laptop. Auditory stimuli were processed using the computer software Praat (Boersma and Weenink 2019).

The experiment was approved by the Ethics Committee of the Faculty of Humanities of the University of Amsterdam.

3.1.4 Analysis

We report the data of all 60 participants. For our statistical analyses we used R (R Core Team 2013), the rstanarm package (Goodrich et al. 2018) for fitting our models in the Bayesian framework, and the bayestestR package (Makowski et al. 2019a) to describe them. This allowed us to directly test the weight of the evidence for (or against) our research hypotheses (Nicenboim and Vasishth 2016). Furthermore, we used weakly informative priors for all the regression parameters, which allowed us to make the most conservative inference possible given the structure of the data (Aslin et al. 1998; Barr et al. 2013).

To determine how participants' perception of speaker reliability differed per condition, we constructed a score from the analysis of the two seven-point Likert-type items of the Subjective reliability questionnaire. Recall that these two items tapped into the speaker's history of (in) accuracy in providing labels for objects, and their reliability as an informant. A correlational analysis between these two items showed that they were significantly correlated ($r = 0.83$, $p < 0.001$). Consequently, we averaged the scores of these items to build our Subjective reliability score. We then ran a linear model using the *lm* R function, with participants' subjective reliability scores as a dependent variable, and Speaker Reliability as a predictor. We used orthogonal sum-to-zero contrast coding for the dichotomous Speaker Reliability variable (Reliable Speaker = +0.5; Unreliable Speaker = –0.5).

To analyse how participants' performance was influenced by speaker reliability across blocks, we fitted two generalized mixed effects linear models (GLM) and used the logit link function to model participants' probability of a correct response (Baguley 2012; Baayen 2008). We took participants' responses in the CSWL task as our dependent variable (1 = correct; and 0 = incorrect), with Speaker Reliability (Reliable Speaker v. Unreliable Speaker; GLM 1) and Subjective Reliability (GLM 2) as between-subjects fixed effects, and Block as within-subjects fixed effect. Following Barr et al. (2013), we fitted maximal models, justified by our design. We included a random intercept for subject, as well as a by-subject random slope for Block. Given that the word-pairings were fully randomized per participant, i.e. unique for each of them, we did not include a random intercept for item. An orthogonal sum-to-zero contrast coding scheme was used for the Speaker Reliability variable (Reliable Speaker = +0.5; Unreliable Speaker = –0.5). Block was treated as numeric predictor, with each of its values (1, 2, 3 and 4)

corresponding to each of the four learning blocks of the CSWL task. We centered this variable for ease of interpretation. Recall that since there was no correct response for the unreliably mapped items in the Unreliable Speaker condition, we removed these items from the analysis. Hence, to account for participants' word learning we analyzed participants' responses to the reliably mapped items only.

Following Makowski et al. (2019b)'s reporting guidelines, for each of the fixed factors we report their posterior distribution by providing their 1) probability of direction (pd; an index of effect existence)[3]; 2) point estimate (estimated by MCMC sampling), in this case, the median; 3) Credible Interval (CI) set at 95%; and 4) Region of Practical Equivalence (ROPE).[4] The latter is an index for null hypothesis testing under the Bayesian framework (Kruschke 2014); in other words, a test of significance.

For plots, we used the ggplot2 package (Wickham 2009), as well as the sjPlot package (Lüdecke 2019). The R code for our Bayesian analyses can be found in our Open Science Framework (OSF) project page (DOI: 10.17605/OSF.IO/E5K4C).

3.2 Results

We begin by reporting the results of the Subjective reliability score, which provides a first glimpse of the effect that our speaker reliability manipulation had on participants. We then report the results of the CSWL task.

3.2.1 Subjective reliability

With the aim to answer the extent to which participants perceived the different speakers' reliabilities, we analyzed the scores of the Subjective reliability questionnaire. **Figure 2** suggests that participants in the Unreliable Speaker condition assessed the speaker as less reliable, compared to those in the Reliable Speaker condition.

---

[3] The Probability of Direction (pd) is an index of effect existence. It ranges from 50% to 100%, and it represents the certainty with which an effect is negative or positive (i.e., its direction). It strongly correlates with the frequentist p-value, such that a two-sided p-value of .1, .05, .01 and .001 would correspond to a pd of 95%, 97.5%, 99.5% and 99.95%, respectively.

[4] As an index of effect significance, the Region of Practical Equivalence allows us to know whether a parameter is related (or not) to a non-negligible change in the outcome. Following Kruschke and Liddell (2018), the ROPE could be set, by default, to a range from –0.1 to +0.1 of a standardized parameter. For linear models, the parameters can be generalized to –0.1 * SD_y, 0.1 * SD_y. For logistic models, the parameters (expressed in log odds) were converted to standardized difference through the formula $\pi/3$–√. This results in a range of –0.18 to +0.18. For ease of interpretation, the following reference values are used (for a detailed description, see Makowski et al. (2019b)): >99% in ROPE: negligible (we can accept the null hypothesis); >97.5% in ROPE: probably negligible; ≤97.5 & ≥2.5% in ROPE: undecided significance; <2.5% in ROPE: probably significant; and 1% in ROPE: significant (we can reject the null hypothesis).

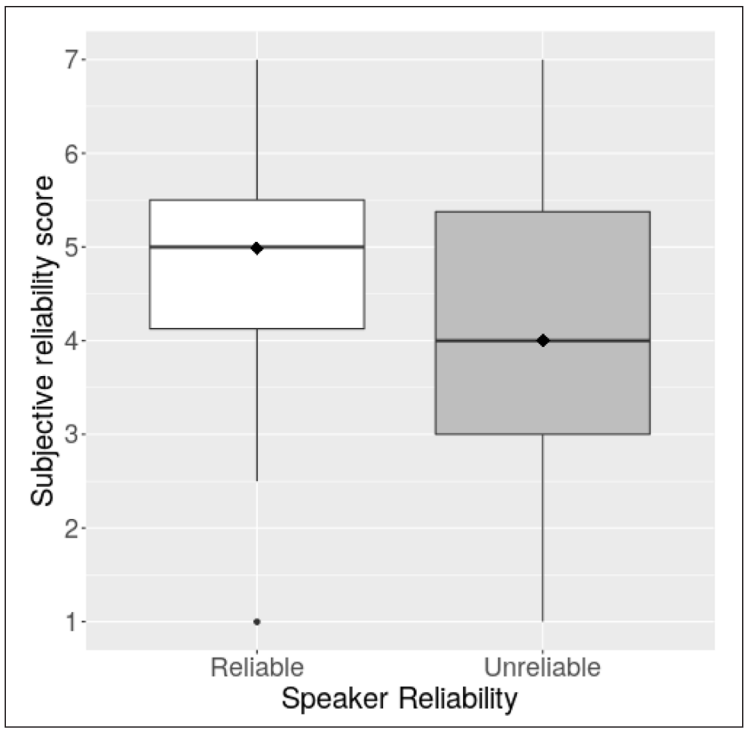

Figure 2: Boxplot of the Subjective reliability scores per condition. Y-axis represents the 7-point Likert-type scale measuring how reliable the speaker was perceived (1 = strongly unreliable; 7 = strongly reliable)

The results of the linear model showed that this difference is significant and it probably exists (Median = 0.96, 95% CI [0.2, 1.7], 0% in ROPE, pd = 99.21%). From these results we can conclude that, when presented with unreliably mapped word-object pairs, we have evidence this has an effect on how reliable the speaker is perceived by adult word learners. This is reflected in participants rating the speaker as a less reliable source of information when exposed to both reliable and unreliable word-object co-occurrences, compared to those participants exposed solely to reliable word-object mappings.

3.2.2 Cross-situational word learning task

The analysis of the mean accuracy scores showed that participants' performance was well above chance (0.5) for both conditions in each of the four blocks (**Figure 3**), in line with the results observed in both the classical CSWL task (e.g. Yu and Smith 2007) and adapted versions, where different word-object probability co-occurrences are tested (e.g. Vouloumanos 2008), and other socio-communicative cues are studied (e.g. MacDonald et al. 2017; Poepsel and Weiss 2014). As a group and across all four blocks, participants in the Reliable Speaker condition selected the target object with a mean accuracy of 86.2% (SD = 34), whereas participants in the Unreliable Speaker condition selected the target object with a group mean accuracy of 83% (SD = 38).

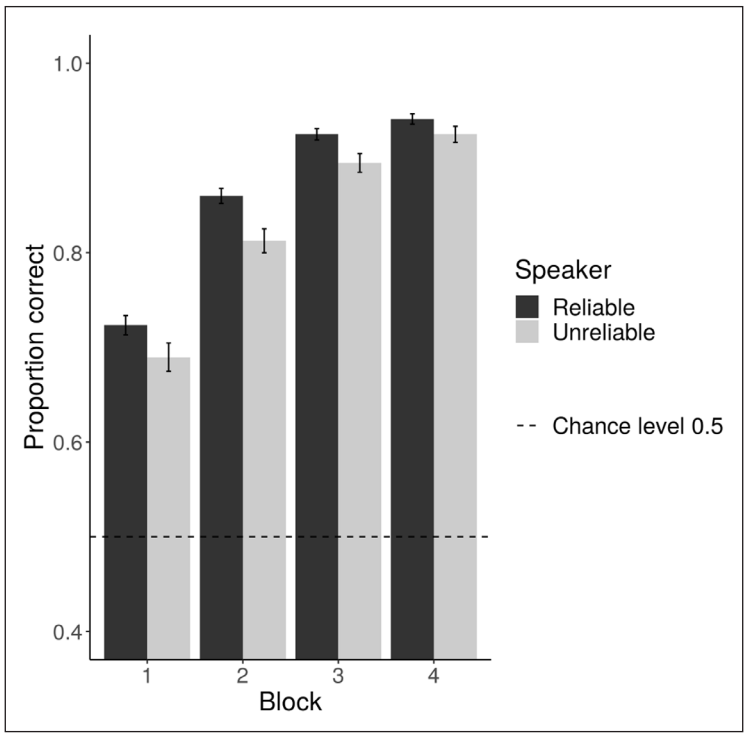

Figure 3: Average accuracy across blocks per Speaker Reliability condition. Error bars mark the standard error of the mean.

In order to answer how speaker reliability affects CSWL, we first examined the main effect of the Speaker Reliability variable (GLM 1), followed by the main effect of Subjective Reliability (GLM 2). As shown in **Figure 4**, learners in the Reliable Speaker condition were 1.20 times more likely to perform better on the reliably mapped items than those in the Unreliable Speaker condition (Median (OR) = 1.20, 95% CI [0.77, 1.86]). However, the significance of this effect is undecided (45.39% in ROPE), and its existence uncertain (pd = 74.88%). This finding can be supplemented by the results of interaction between Speaker Reliability and Block,[5] which also

[5] As suggested by the editor, we run a model with Block treated as a 4-level categorical predictor. We used an orthogonal sum-to-zero contrast coding scheme with 1) block 1 coded as –0.5, block 2 as +0.5, and both blocks 3 and 4 set to 0; 2) blocks 1 and 2 coded as –1/4 and blocks 3 and 4 coded as +1/4; and finally 3) block 3 coded as –0.5, block 4 as +0.5, and both blocks 1 and 2 set to 0. The results show that the significance of the main effect of Speaker Reliability remained undecided and its existence uncertain (Median (OR) = 1.23, 95% CI [0.76, 1.97]; 41.53% in ROPE, pd = 80.23%). The interaction between Speaker Reliability and Block shows that, compared to participants in the Unreliable speaker condition, participants in the Reliable Speaker condition were: 1) 1.15 times more likely to perform better in Block 1 vs. Block 2 (Median (OR) = 1.15, 95% CI [0.70, 1.89]; 48.49% in ROPE, pd = 71.26%); 2) 0.98 times more likely to choose the target object in both Blocks 3 and 4 vs. Blocks 1 and 2 (Median (OR) = 0.98, 95% CI [0.51, 1.94]; 42.62% in ROPE, pd = 51.25%); and 3) 0.92 more likely to perform better in Block 4 vs Block 3 (Median (OR) = 0.92, 95% CI [0.56,1.51]; 53.69% in ROPE, pd = 62.88%). Each of these interactions were of undecided significance and uncertain existence.

shows that the significance of this effect is undecided and its probability of existence uncertain (Median (OR) = 1.00, 95% CI [0.79, 1.25]; 89.71% in ROPE, pd = 50.98%).

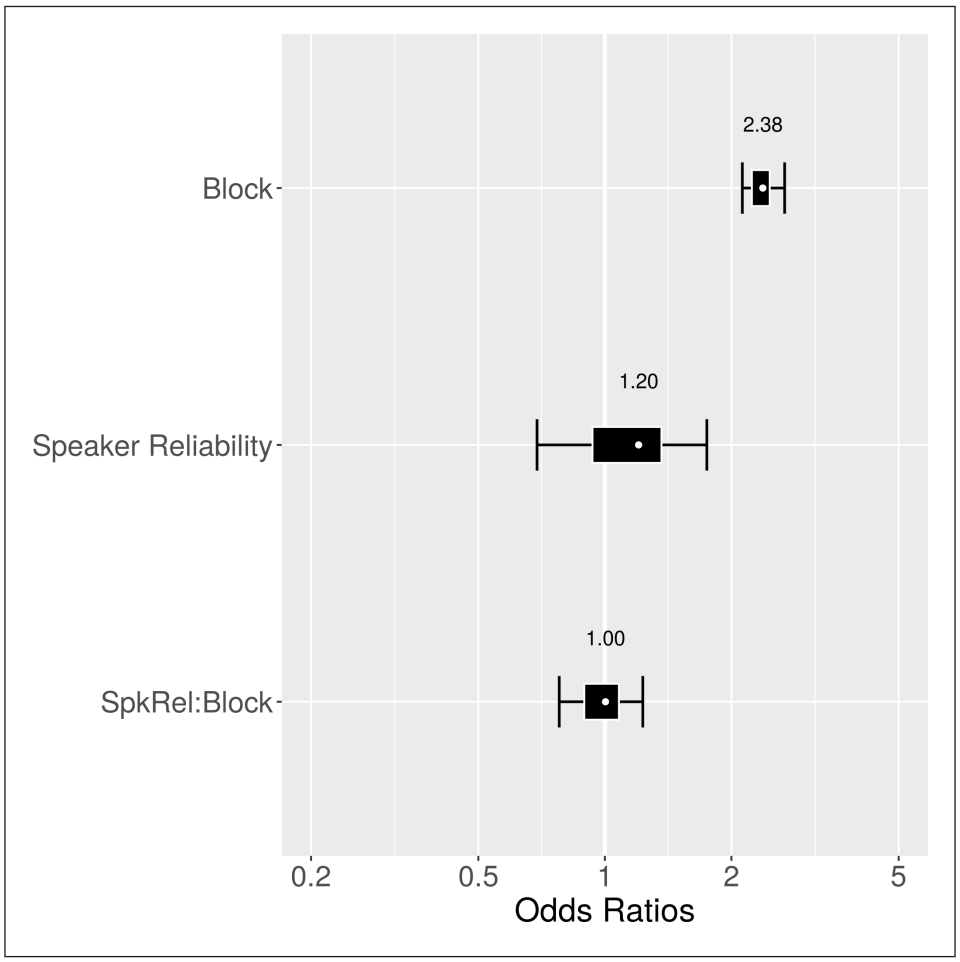

Figure 4: Odds ratio for each of the fixed effects of the first generalized mixed effects model. Solid bars indicate the posterior distribution of coefficient values for each parameter; white dots the point estimates (median); and black lines the 95% Credible Interval. SpkRel:Block = interaction between Speaker Reliability and Block

The analysis of the main effect of Subjective Reliability on CSWL shows a similar pattern of results (**Figure 5**). Participants who perceived the speaker as more reliable were 1.03 times more likely to perform better than those who perceived the speaker as less reliable (Median (OR) = 1.03; 95% CI [0.89, 1.1], 100% in ROPE, pd = 63.91%). However, the existence of this effect is uncertain, and its significance negligible. Furthermore, when we examine its interaction with Block, a similar pattern of results is observed (Median (OR) = 1.02; 95% CI [0.95, 1.09], 100% in ROPE, pd = 73.18%).

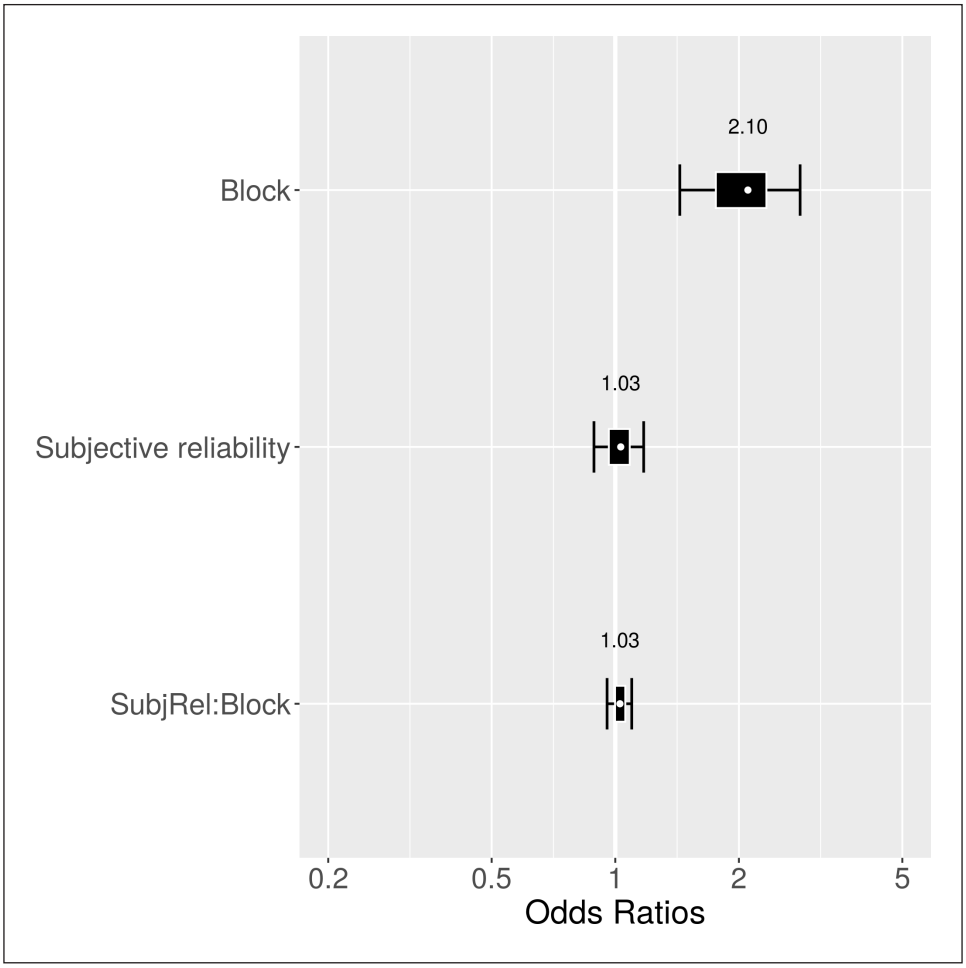

Figure 5: Odds ratio for each of the fixed effects of the second generalized mixed effects model. Solid bars indicate the posterior distribution of coefficient values for each parameter; white dots the point estimates (median); and black lines the 95% Credible Interval. PerRel:Block = interaction between Subjective Reliability and Block.

Taken together, these findings do not support the prediction that participants in the Reliable Speaker condition are significantly more likely to perform better on the reliably mapped items than those in the Unreliable Speaker condition. In response to our first research question, we found no evidence that being exposed to either a reliable or unreliable speaker has an effect on adults' CSWL.

3.3 Discussion

The aim of Experiment 1 was to answer, first, to what extent learners would judge the speaker's reliability differently depending on the consistency with which the objects were labeled across situations; and second, the degree to which speaker reliability affects CSWL. The negligible effect

of Subjetive Reliability, together with the undecided significance and uncertain existence of a Speaker Reliability effect, revealed that our speaker reliability manipulation – i.e. the speaker labeling objects unreliably in the Unreliable Speaker condition – did not have an effect on participants' word learning rate.

Interestingly, the unreliable word-object mappings very likely affected the perception that learners had about the speaker's reliability, reported in the perceived speaker reliability measurement. When exposed to a speaker that maps the same label to different objects across several situations, i.e. unreliably, learners' perception of that speaker led them to judge them as a less reliable source of information. The opposite holds when the speaker mapped the same word to a unique object only, i.e. that participants were more likely to judged the speaker as a more reliable source of information. In spite of these results, inspection of participants' individual answers in the perceived reliability questionnaire shows that several of them in the Unreliable Speaker condition still assessed the speaker as a reliable source of information. A potential explanation for this is that our speaker reliability manipulation was not strong enough: participants had to notice that some of the word-object pairs were unreliably mapped and attribute this to a speaker who could be making mistakes in labeling objects. Even though some participants reported making this association, this may have been weakened by the salience of the statistical information they had to process. The large number of trials per item may have turned out to be more relevant to process than the speaker's characteristics. Furthermore, participants had to learn the correct word-object mapping before they could notice the unreliability at all.

In order to make the reliability of the speaker more salient, we designed a second experiment where the speaker reliability manipulation was implemented within participants. Thus, participants were exposed to two speakers, a reliable and an unreliable one. Given that participants of Experiment 1 were able to gain a different perception of reliability from a single speaker, this within-subject manipulation allowed us to test the extent to which participants are able to track the reliability of multiple speakers. This design also allowed us to avoid the shortcomings of between-subject experimental designs, such as sample size and individual variability.

4 Experiment 2

In this experiment we presented participants with both a reliable and an unreliable speaker, differentiated by their linguistic competence. This was reflected in the consistency with which each speaker mapped a word to an object across trials. This allowed us to test the extent to which learners were able to track the reliability of the speakers and, as in Experiment 1, to what extent their word learning was affected by the speaker's reliability as a source of information.

4.1 Methods

4.1.1 Participants

Sixty Dutch-native speakers aged between 18 and 31 years old (M = 23.30, SD = 3.24; 34 female; 23 male; and three who identified as 'other'), participated in the experiment for payment of €5. All participants had normal or corrected-to-normal hearing and sight.

4.1.2 Materials and design

We employed the same stimuli and software programs used in Experiment 1. Half of the words were recorded by a female Dutch native speaker, and the other half consisted of the same recordings used in Experiment 1.

Even though the CSWL task was the same as in Experiment 1, the within-subject design changed both the distributional co-occurrence of the reliably and unreliably mapped word-object pairs, and their conditional probabilities. The Reliable Speaker condition consisted of two sets of four (instead of eight) reliably mapped word-object items, i.e. objects that consistently co-occurred with one particular word across trials (see Supplementary material 2). The co-occurrence of an object given a word (p (V|A)), and that of a word given a foil object (p (A|V) were higher than in Experiment 1 because foils were chosen from and within each set of four word-object pairs (**Table 2**). However, the conditional probability of the target object co-occurring with a given word, and that of the same word given the target object remained the same, i.e. 1 and 0.5, respectively.

Speaker reliability	**Stimuli**				**Conditional probability**	
	Number of items	**Type of mapping**	**Auditory (A)**	**Visual (V)**	***p* (V|A)**	***p* (A|V)**
Reliable	4	reliable	word	target	1	0.5
	4		word	foil	0.33	0.16
Unreliable	4	reliable	word	target	1	0.5
			word	foil	0.33	0.16
	4	unreliable	word	target & foil	0.5	0.25

Table 2: Conditional probabilities per Speaker Reliability condition.

For the Unreliable Speaker condition, the set of inconsistently mapped word-object pairs was reduced to four items. These consisted of four different objects paired to four different words across trials. This translates into the same conditional probability for both the word and its target object, and the same word and its foil, but with a higher value than in Experiment 1, that is, a p (A|V) of 0.25. Similarly, the same conditional probability is shared by an object given a word,

i.e. a p (V|A) of 0.5, regardless of that word being that object's label or of another object (see **Table 2**). Again, this conditional probability is higher than in Experiment 1 due to the increased co-occurrence between certain words and certain objects. Finally, the unreliable mapping items remained unlearnable and without a correct answer, just as in Experiment 1.

As in the first experiment, the coupling between a type of mapping and one of the two speakers (female and male) yielded our experimental conditions. Thus, in the Reliable Speaker condition the speaker presents only reliably mapped items. In the Unreliable Speaker condition, instead, the speaker presents both reliably and unreliably mapped items. The speaker-mapping combination was counterbalanced across participants, and randomly assigned to each of them.

Finally, we administered the same 7-point Likert-type Subjective Reliability questionnaire employed in Experiment 1, after the CSWL task. This time participants had to assess both speakers.

4.1.3 Procedure

We followed the same procedure as in the previous experiment, with a slight change in the instructions we provided to participants before the third block. Specifically, when asking participants whether they were able to learn new words from the speakers, we removed the information about the latter being non-native speakers of the new language because we had already provided that information at the beginning of the experiment (see Supplementary material 3). The experiment was approved by the Ethics Committee of the Faculty of Humanities of the University of Amsterdam.

4.1.4 Analysis

We report the data of all 60 participants. For our statistical analyses we used the same program and packages as in Experiment 1. For a detailed description of our models, see our OSF project page (DOI: 10.17605/OSF.IO/E5K4C).

For the analysis of the Subjective Reliability, we followed the same procedure as that of the previous experiment regarding the creation of a Subjective Reliability score, although this time we had two scores, one per speaker. Each score was built from the two items (per speaker) assessing the speaker's reliability, which were significantly correlated (reliable speaker: $r = 0.79$, $p < 0.001$; unreliable speaker: $r = 0.68$, $p < 0.001$). From these scores we created a relative Subjective Reliability score by subtracting the score for the unreliable speaker from the score for the reliable speaker for each participant. We used this score as a dependent variable in our linear model, with Speaker Reliability as a fixed factor.

We performed a similar analysis as that of Experiment 1 for the accuracy data of the CSWL task, with the only difference that the models included a by-subject slope for Speaker Reliability (GLM 3) and Subjective Reliability (GLM 4) in the random structure.

4.2 Results

We first report the results of the Subjective Reliability score, followed by the results of the CSWL task.

4.2.1 Subjective reliability

With the purpose to see how participants would assess the speakers differently, we analyzed the relative Subjective Reliability scores we built from participants' assessments of each speaker. **Figure 6** suggests no difference between speakers in terms of how (un)reliable they were perceived by the participants.

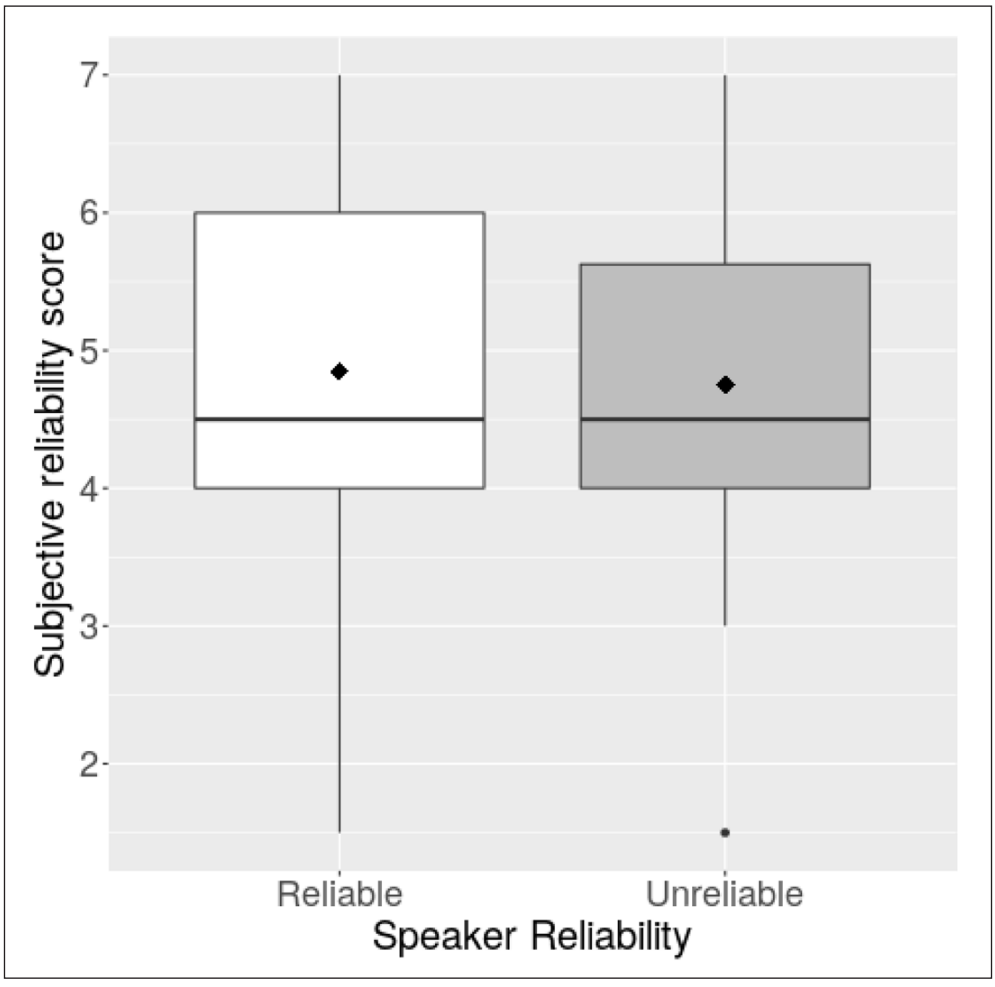

Figure 6: Boxplot of the Subjective Reliability score per condition. Y-axis represents the 7-point Likert-type scale measuring how reliable the speaker was perceived (1 = strongly unreliable; 7 = strongly reliable).

The results of the linear model confirm this observation by showing that the existence of a distinctive perception of each speakers' reliability was uncertain, with an undecided significance associated to it (Median = 0.00,[6] 95% CI [–0.250, 0.255], 45% in ROPE, pd = 50.22%). This means that participants did not perceive the reliable and unreliable speaker differently with respect to their reliabilities, given that they assessed both speakers as equally reliable as sources of information. We can conclude that we do not have evidence that adults tracked the different speakers' reliability.

4.2.2 Cross-situational word learning task

Descriptive statistics of the mean accuracy scores per condition and block are shown in **Figure 7**. As a group, and similar to the results in Experiment 1, participants' performance was well above chance (0.5) for both conditions in each block. Participants in the Reliable Speaker condition selected the target object with a mean accuracy of 77% (SD = 42), whereas participants in the Unreliable Speaker condition selected the target object with a group mean accuracy of 79% (SD = 41).

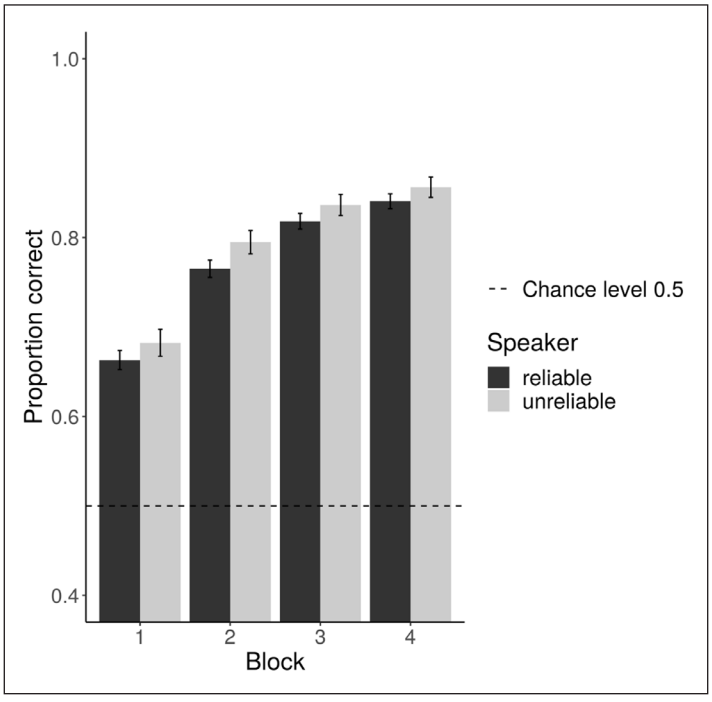

Figure 7: Average accuracy across blocks per Speaker Reliability condition. Error bars mark the standard error of the mean.

[6] An estimate value of 0 indicates that both speakers were rated as equally reliable. A value < 0 indicates that the reliable speaker was perceived as less reliable than the unreliable speaker. Finally, a value > 0 indicates that the reliable speaker was perceived as more reliable than the unreliable speaker.

With the aim to answer to what extent participants' CSWL is affected by the reliability of the speaker, we fit two generalized linear mixed-effects models and assessed, first, the main effect of Speaker Reliability (GLM 3), and second, the main effect of Subjective Reliability (GLM 4). As illustrated in **Figure 8**, the results show that participants had a lower probability of performing better on items presented by the reliable speaker than on those presented by the unreliable speaker (Median (OR) = 0.71; 95% CI [0.58, 0.86], 3.60% in ROPE, pd = 99.5%). Even though this effect probably exists, its significance remains undecided. To complement these results, we also report the interaction between Speaker Reliability and Block, which shows that this effect possibly exists, but its significance is undecided (Median (OR) = 0.88; 95% CI [0.78, 0.98], 80.81% in ROPE, pd = 96%).

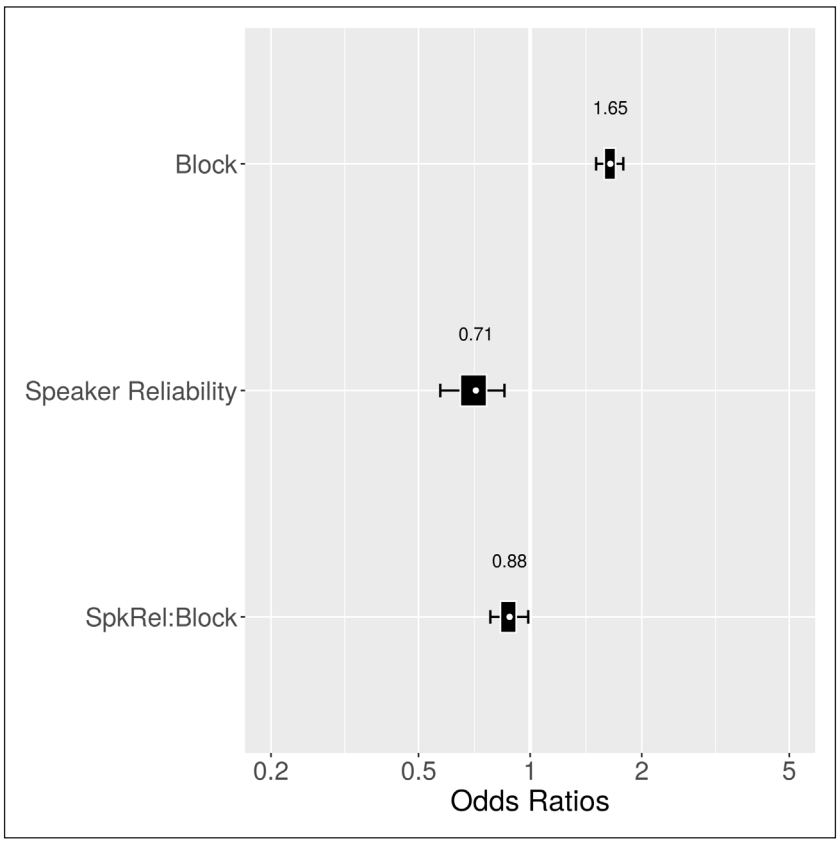

Figure 8: Odds ratio for each of the fixed effects of the first generalized mixed effects model. Solid bars indicate the posterior distribution of coefficient values for each parameter; white dots the point estimates (median); and black lines the 95% Credible Interval. SpkRel:Block = interaction between Speaker Reliability and Block.

The analysis of the main effect of Subjective Reliability on CSWL shows that participants who perceived the speaker as more reliable had a lower probability of performing better than those

who perceived the speaker as less reliable (Median (OR) = 0.02, 95% CI [–0.09, 0.12], 100% in ROPE, pd = 62.21%; see **Figure 9**). However, the significance of this main effect is negligible, and its existence probability is uncertain. Furthermore, and as a complement to this finding, the interaction between Subjective Reliability and Block is also negligible in terms of significance, as well as uncertain with respect to its existence (Median (OR) = –0.004, 95% CI [–0.07, 0.06], 100% in ROPE, pd = 56.53%).

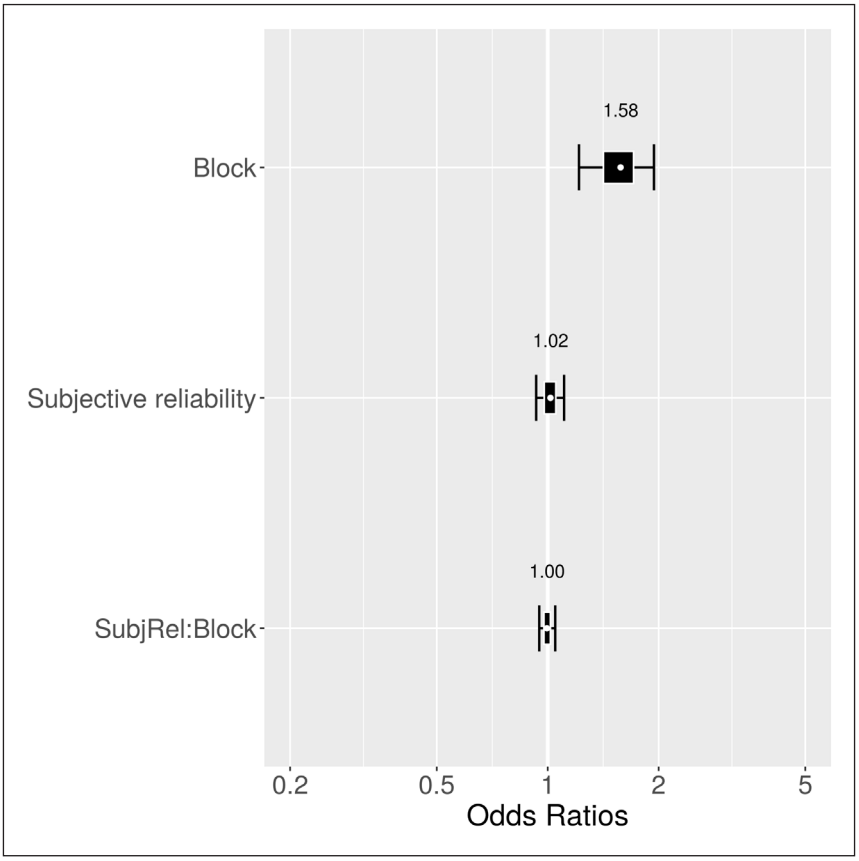

Figure 9: Odds ratio for each of the fixed effects of the second generalized mixed effects model. Solid bars indicate the posterior distribution of coefficient values for each parameter; white dots the point estimates (median); and black lines the 95% Credible Interval. PerRel:Block = interaction between Subjective Reliability and Block.

Altogether, these results provide weak evidence against the null hypothesis. Hence, answering our second research question, we conclude that our data do not reliably support the prediction that adults presented with a reliable speaker are more likely to perform better on the reliably mapped items across situations, compared to their performance on similar items presented by an unreliable speaker. Instead, the data show that participants in both speaker conditions are

likely to perform similarly over blocks with regard to their learning trajectory, regardless of the (subjective) reliability of the speaker.

4.3 Discussion

The aim of Experiment 2 was to implement a stronger speaker reliability manipulation that would allow us to investigate the extent to which the reliability of the speaker has an effect on adults' CSWL. As in Experiment 1, the results showed that participants in both conditions learned well above chance level, even though their accuracy was slightly lower. Nevertheless, these results are consistent with previous findings in adult CSWL (e.g. MacDonald et al. 2017; Poepsel and Weiss 2014; Roembke and McMurray 2016; Vouloumanos 2008; Yu and Smith 2007).

With respect to participants' perception of the speakers' reliabilities, participants did not assess the speakers differently. Even though we expected that, by having two distinct speakers, participinats would differentiate between their linguistic competence, no significant difference was observed. Similar to the results we observed for the effect of Speaker Reliability on participants' accuracy, the number of items in the Unreliable Speaker condition may account for this. Participants in this condition were presented with four unreliably mapped word-object pairs (vs. eight in Experiment 1). Even though the proportion was the same as that of Experiment 1, this number of unreliable items might not have been sufficient for triggering a perception that the speaker was unreliable. Furthermore, upon noticing that some words did not pair with an object consistently, participants may have attributed this to their own performance on the task, and not to a specific speaker labeling objects 'incorrectly'.

Regarding the Speaker Reliability effect on CSWL, the results show the opposite pattern of those observed in Experiment 1, and of what we predicted. Participants in the Unreliable Speaker condition tended to perform better than those in the Reliable Speaker condition across blocks, an effect that the analysis showed that it possibly exists. In spite of its undecided significance, we argue that a possible explanation for this behavior may come from the experimental design itself. Recall that we performed our analyses on the reliably mapped word-object pairs, and that due to the within-subjects design we had fewer of these items in the Unreliable Speaker condition (four, versus eight in the Reliable Speaker condition). This could have guided participants' attention towards these items only because half of the words they had to learn were virtually unlearnable. Thus, participants could have figured out that since some of the items were more difficult learn, it was better to focus on those for which there was a clear mapping between the new word and its target object across situations.

5 General discussion

Adults (as well as children) are adept at picking up the statistical regularities in the environment (Siskind 1996). Studies testing CSWL have consistently shown that learners are able to track the

co-occurrences between words and objects, even in noisy and ambiguous situations (Smith 2000; Smith et al. 2011; Suanda et al. 2014; Vouloumanos 2008; Yu and Smith 2007). In a more natural learning environment, however, learners are usually exposed to information that goes beyond the statistical regularities present in the language. Socio-communicative and social-pragmatic approaches to word learning emphasize the important role that the speaker's characteristics and intentions play in language acquisition (Krogh-Jespersen and Echols 2012; Scofield and Behrend 2008; Sobel et al. 2012). These factors are central to word learning, and even more when the speaker inconsistently labels a referent across situations. In such cases, learners are exposed to a speaker who presents them with unreliable information, a non-cooperative behavior that might ultimately affect the way they learn new words. In light of this, the present study raised the questions about the extent to which it is possible for learners to detect the reliability of the speaker across situations, and how they can use this information to modulate the extent to which they learn from interlocutors.

The two experiments reported here show that adult learners are able to successfully learn new words in a CSWL task, well above chance level, and in line with previous CSWL studies (e.g. MacDonald et al. 2017; Monaghan et al. 2012; 2015; Roembke and McMurray 2016; Smith et al. 2011; Smith and Yu 2008; Suanda et al. 2014; Vouloumanos 2008; Yu and Smith 2007). Contrary to our predictions, though, they did so even when they had to learn from a speaker who offered unreliable input. Specifically, they effectively learned from someone who labeled half of the objects inconsistently across situations. This result is not unusual nonetheless. Recent research on how disfluencies affect children's word learning has shown that young learners are able to learn new words equally well from both fluent and disfluent speakers (White et al. 2020). Hence, any reduced confidence a word learner may develop when exposed to less credible speakers does not seem to affect their overall learning. Further research on this topic could explore the extent to which speaker characteristic effects take place during word learning, that is, before learners make a conscious choice. This could be a fruitful research direction in view of the extensive evidence on speaker characteristic effects on referential resolution in language comprehension (e.g. Barr et al. 2014; Hanna and Tanenhaus 2004; Heller et al. 2008; Keysar et al. 1998; 2000; Kronmüller and Barr 2015; Metzing and Brennan 2003; Nadig and Sedivy 2002).

Interestingly, when presented with a speaker who provided unreliably word-object co-occurrences, participants in Experiment 1 perceived them as a less reliable source of information, compared to participants who were not exposed to such type of co-occurrences. As pointed out by one of the reviewer, this could arise from participants' own learning profile, reflected in their overall accuracy in the task. Nevertheless, we found no significant differences in overall accuracy between conditions, which suggests that the assessment of the speaker's reliability was not a direct consequence of participants' performance. The results of Experiment 1 point to an apparent distinction between learning from speakers with different reliabilities and,

at the same time, developing a perception of the speaker's reliability. However, we do not have evidence as yet that learners can attribute different levels of reliability to more than one speaker at a time.

As the results of Experiment 2 showed us, participants did not seem to have tracked and recognized the reliability of each speaker. Recall that participants were exposed to fewer unreliable items in Experiment 2 compared to Experiment 1. The question that arises then is how much evidence is necessary for participants to detect the speaker's reliability, and for the latter to have an effect on their word learning. If we assume that the mechanism underlying participants' learning was error-based, i.e. that in proportion to the magnitude of the errors they encountered, they adjusted their expectations for future input (Fine and Jaeger 2013), then it is possible that the number of unreliable items in Experiment 2 was not enough data from which participants could detect reliability.

Despite the fact that the statistical significance of the effect of Speaker Reliability on the CSWL task was undecided in both experiments, the Bayesian parameter of Probability of Direction (pd) shows that its status as main effect differed from Experiment 1 to Experiment 2. Specifically, from an uncertain (pd = 74.88%) to a probably existing effect of speaker reliability on CSWL (99.55%). Nevertheless, and contrary to our predictions, input offered by an unreliable speaker led to slightly better learning in Experiment 2. Recall that our hypothesis was that if a speaker offers unreliable input, even the reliably mapped items presented by them would be difficult to learn. By tracking the statistics of the unreliable speaker's referential expression, learners experienced 'errors', i.e. words that could not be mapped to one, unique object. This would lead learners not to trust a speaker who makes labeling mistakes, and ultimately affect their word learning. To account for the results we observe in Experiment 2, we argue that since learning the words from an unreliable speaker requires more effort than learning from a reliable one, participants' attention was driven to specific items early on in the experiment. When items were presented by the unreliable speaker, those which mapping was inconsistent across situations may have prompted learners to focus their attention and effort on the items that were learnable, discarding those that were not. When learning from the reliable speaker, instead, learners' attention was not directed to any particular item. Since the statistical information presented by the reliable speaker was rather consistent and equal for each item, learners had no need for directing their attention and focus on learning a specific word. This might have led them to perform more inattentively across items, which could account for their lower performance in Experiment 2. These results, nonetheless, should be taken with caution, as the significance of this effect is undecided.

Finally, compared to most word learning studies, where the speakers' reliability is made explicit to learners (as in, for example, Sobel et al. (2012: p. 96): "She doesn't know all the right words for things"), we opted for a rather indirect way of describing the speakers' reliability in both experiments. We characterized the speakers as second language learners of the language

taught to participants, and matched each speaker to distinct word-object mappings. This entailed that participants had to 1) infer that the degree of the speaker's competence in the new language was a reflection of their reliability as a speaker, and 2) track the statistics of the speaker's referential expression. When presented with one speaker, learners do this efortlessly. However, when presented with two speakers, learners do not seem to make a distinction between speakers. We think that explicitly describing the speakers as either reliable or unreliable, along with an explicit question about speakers' reliability (as in MacDonald et al. (2017)), could have helped participants in Experiment 2 to track the differences between speakers, and assess their reliability distinctively. A design like this could help us determine the extent to which learners incorporate the speakers' information into their word learning process.

6 Conclusion

The two experiments presented here tested the extent to which adults' CSWL is affected by the reliability of the speaker, and how adult learners are able to perceive the speakers' reliabilities differently. We found no evidence to support the hypothesis that being presented with an unreliable speaker impairs CSWL in adults. We found nonetheless that participants being exposed to an unreliable speaker assessed them as less reliable, compared to being exposed to a more reliable speaker (Experiment 1). However, this effect disappears when participants are presented with both a reliable and an unreliable speaker (Experiment 2). Furthermore, the word learning trajectories in both experiments did not differ significantly.

Recently, hybrid approaches to CSWL have investigated whether learners process statistical information differently when socio-pragmatic cues are present in the learning environment. The evidence so far, however, is scarce and mixed. Whereas no significant effects of speaker's identity on CSWL has been found (e.g. Poepsel and Weiss 2014; Chan and Monaghan 2019), a small effect of speaker's gaze has been shown on a similar task (MacDonald et al. 2017). The present study tested a different cue, namely the reliability of the speaker, considering that social-pragmatic accounts of language acquisition have consistently highlighted and shown the importance of the speaker's reliability in early word learning (Brosseau-Liard et al. 2014; Buac et al. 2019; Jaswal and Neely 2006; Koenig and Harris 2005; Koenig et al. 2004; Scofield and Behrend 2008; Sobel et al. 2012). Even though we found no effect of speaker reliability on CSWL, two findings draw our attention and are relevant for further investigation. First, a different perception of the speakers' reliability in Experiment 1; and second, the possibility of existence of a speaker reliability effect in Experiment 2. We propose that such an effect could be observed in a design where 1) the ambiguity of the situation is higher than the ones we designed here, and 2) not only the characteristics, but also the intentions of the speaker are clearly expressed.

We think the starting point for further research on speaker reliability effects on CSWL should focus on how learners view others and, in particular, those who are teaching them new words. It is when people are viewed as knowledgeable and helpful teachers that learning becomes a more tractable problem (Shafto et al. 2012). In other words, when the speaker is reliable, solving the ambiguity of the referent and learning new words becomes easier.

Supplementary material

Supplementary material can be found in our OSF project page (DOI: 10.17605/OSF.IO/E5K4C).

Supplementary material 1: Stimuli. Stimuli used in the CSWL task.

Supplementary material 2: Word-object co-occurrences. Matrix of word-object co-occurrences for each type of item.

Supplementary material 3: Instructions. Instructions and information about the speakers given to participants in both experiments.

Funding information

This work was funded by the National Agency of Research and Development (in Spanish, ANID) of the Ministry of Science and Technology of Chile through a BECAS CHILE grant obtained by Natalia Rivera-Vera (project number 72170240).

Acknowledgements

We wish to thank Dirk-Jan Vet for his help with programming the experiment in Python and developing PRAAT scripts for pre-processing the data. We also thank the UvA (r)MA students Jeremy Bos, Anne Beltman, Marc van Waarden and Jenny Tsiara for their help with recruiting and testing participants.

Competing interests

The authors have no competing interests to declare.

Author contributions

The experiments described here were designed by Natalia Rivera-Vera, Sible Andringa, and Padraic Monaghan. Natalia Rivera-Vera implemented the experiments, and conducted them with the help of four research assistants: Jeremy Bos, Anne Beltman, Marc van Waarden and Jenny Tsiara. Natalia Rivera-Vera analyzed the data of both experiments, and consulted Sible Andringa, Padraic Monaghan and Edmundo Kronmüller for statistical advice. Natalia Rivera-Vera wrote the article with feedback and contributions from each of the co-authors.

References

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological science, 9*(4), 321–324. DOI: https://doi.org/10.1111/1467-9280.00063

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511801686

Baguley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioral sciences*. Macmillan International Higher Education.

Baldwin, D. A. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of child language*, *20*(2), 395–418. DOI: https://doi.org/10.1017/S0305000900008345

Baldwin, D. A., Markman, E. M., Bill, B., Desjardins, R. N., Irwin, J. M., & Tidball, G. (1996). Infants' reliance on a social criterion for establishing word-object relations. *Child development*, 67(6), 3135–3153. DOI: https://doi.org/10.2307/1131771

Barr, D. J., Jackson, L., & Phillips, I. (2014). Using a voice to put a name to a face: the psycholinguistics of proper name comprehension. *Journal of Experimental Psychology: General*, *143*(1), 404. DOI: https://doi.org/10.1037/a0031813

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), 255–278. DOI: https://doi.org/10.1016/j.jml.2012.11.001

Bloom, L. (2000). The intentionality model of word learning: How to learn a word, any word. In R. M. Golinkoff & K. Hirsh-Pasek (Eds.), *Becoming a Word Learner: A Debate on Lexical Acquisition* (pp. 19–50). Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780195130324.003.002

Boersma, P., & Weenink, D. (2019). Praat: doing phonetics by computer [computer program].

Brosseau-Liard, P., Cassels, T., & Birch, S. (2014). You seem certain but you were wrong before: Developmental change in preschoolers' relative trust in accurate versus confident speakers. *PLoS ONE*, *9*(9), e108308. DOI: https://doi.org/10.1371/journal.pone.0108308

Buac, M., Tauzin-Larché, A., Weisberg, E., & Kaushanskaya, M. (2019). Effect of speaker certainty on novel word learning in monolingual and bilingual children. *Bilingualism*, *22*(4), 883–895. DOI: https://doi.org/10.1017/S1366728918000536

Chan, K. C. J., & Monaghan, P. (2019). Simulating bilingual word learning: Monolingual and bilingual adults' use of cross-situational statistics. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society* (pp. 1472–1478).

Clark, E. V. (2018). Conversation and Language Acquisition: A Pragmatic Approach. *Language Learning and Development*, *14*(3), 170–185. DOI: https://doi.org/10.1080/15475441.2017.1340843

Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). Multipic: A standardized set of 750 drawings with norms for six european languages. *Quarterly Journal of Experimental Psychology*, *71*(4), 808–816. PMID: 28326995. DOI: https://doi.org/10.1080/17470218.2017.1310261

Fine, A. B., & Jaeger, F. T. (2013). Evidence for implicit learning in syntactic comprehension. *Cognitive Science*, *37*(3), 578–591. DOI: https://doi.org/10.1111/cogs.12022

Frank, M. C., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, *75*, 80–96. DOI: https://doi.org/10.1016/j.cogpsych.2014.08.002

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–585. DOI: https://doi.org/10.1111/j.1467-9280.2009.02335.x

Gardner, B., Dix, S., Lawrence, R., Morgan, C., Sullivan, A., & Kurumada, C. (2021). Online pragmatic interpretations of scalar adjectives are affected by perceived speaker reliability. *PLoS ONE*, *16*(2 February), 1–22. DOI: https://doi.org/10.1371/journal.pone.0245130

Gleitman, L. (1990). The structural sources of verb meanings. *Language acquisition*, 1(1), 3–55. DOI: https://doi.org/10.1207/s15327817la0101_2

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2018). rstanarm: Bayesian applied regression modeling via stan. *R package version*, *2*(4), 1758.

Grice, P. (1989). *Studies in the Way of Words*. Harvard University Press.

Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: evidence from eye movements. *Cognitive Science*, *28*(1), 105–115. DOI: https://doi.org/10.1207/s15516709cog2801_5

Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition*, *108*(3), 831–6. DOI: https://doi.org/10.1016/j.cognition.2008.04.008

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. DOI: https://doi.org/10.3758/s13428-015-0647-3

Jaswal, V. K., & Neely, L. A. (2006). Adults don't always know best: preschoolers use past reliability over age when learning new words. *Psychological Science*. DOI: https://doi.org/10.1111/j.1467-9280.2006.01778.x

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The Role of Mutual Knowledge in Comprehension. *Psychological Science*, *11*(1), 32–38. DOI: https://doi.org/10.1111/1467-9280.00211

Keysar, B., Barr, D. J., Balin, J. A., & Paek, T. S. (1998). Definite Reference and Mutual Knowledge: Process Models of Common Ground in Comprehension. *Journal of Memory and Language*, *20*(39), 1–20. DOI: https://doi.org/10.1006/jmla.1998.2563

Koenig, M. A., Clément, F., & Harris, P. L. (2004). Trust in testimony: Children's use of true and false statements. *Psychological Science*, 15(10), 694–698. DOI: https://doi.org/10.1111/j.0956-7976.2004.00742.x

Koenig, M. A. and Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child development*, *76*(6), 1261–1277. DOI: https://doi.org/10.1111/j.1467-8624.2005.00849.x

Krogh-Jespersen, S., & Echols, C. H. (2012). The influence of speaker reliability on first versus second label learning. *Child Development*, *83*(2), 581–590. DOI: https://doi.org/10.1111/j.1467-8624.2011.01713.x

Kronmüller, E., & Barr, D. J. (2015). Referential precedents in spoken language comprehension: A review and meta-analysis. *Journal of Memory and Language*, *83*, 1–19. DOI: https://doi.org/10.1016/j.jml.2015.03.008

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, second edition. DOI: https://doi.org/10.1016/B978-0-12-405888-0.00008-8

Kruschke, J. K., & Liddell, T. M. (2018). The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic bulletin & review*, *25*(1), 178–206. DOI: https://doi.org/10.3758/s13423-016-1221-4

Lev-Ari, S. (2015). Comprehending non-native speakers: theory and evidence for adjustment in manner of processing. *Frontiers in Psychology*, *5*, 1546. DOI: https://doi.org/10.3389/fpsyg.2014.01546

Lüdecke, D. (2019). *sjPlot: Data Visualization for Statistics in Social Science*. R package version 2.8.7.

MacDonald, K., Yurovsky, D., & Frank, M. C. (2017). Social cues modulate the representations underlying cross-situational learning. *Cognitive Psychology*, *94*, 67–84. DOI: https://doi.org/10.1016/j.cogpsych.2017.02.003

Makowski, D., Ben-Shachar, M., & Lüdecke, D. (2019a). bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. *Journal of Open Source Software*, *4*(40), 1541. DOI: https://doi.org/10.21105/joss.01541

Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., & Lüdecke, D. (2019b). Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology*, *10*(December), 1–14. DOI: https://doi.org/10.3389/fpsyg.2019.02767

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*, 201–213. DOI: https://doi.org/10.1016/S0749-596X(03)00028-7

Mills, C. M. (2013). Knowing when to doubt: Developing a critical stance when learning from others. *Developmental psychology*, *49*(3), 404–418. DOI: https://doi.org/10.1037/a0029500

Monaghan, P., Mattock, K., Davies, R. A., & Smith, A. C. (2015). Gavagai Is as Gavagai Does: Learning Nouns and Verbs From Cross-Situational Statistics. *Cognitive Science*, *39*(5), 1099–1112. DOI: https://doi.org/10.1111/cogs.12186

Monaghan, P., Mattock, K., & Walker, P. (2012). The role of sound symbolism in language learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *38*(5), 1152–1164. DOI: https://doi.org/10.1037/a0027747

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. DOI: https://doi.org/10.1111/j.0956-7976.2002.00460.x

Najnin, S., & Banerjee, B. (2018). Pragmatically framed cross-situational noun learning using computational reinforcement models. *Frontiers in Psychology*, *9*(JAN), 1–18. DOI: https://doi.org/10.3389/fpsyg.2018.00005

Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational Ideas—Part II. *Language and Linguistics Compass*, *10*(11), 591–613. DOI: https://doi.org/10.1111/lnc3.12207

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). Psychopy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. DOI: https://doi.org/10.3758/s13428-018-01193-y

Pinker, S. (2013). *Learnability and Cognition: The Acquisition of Argument Structure*. MIT Press. DOI: https://doi.org/10.7551/mitpress/9700.001.0001

Poepsel, T. J., & Weiss, D. J. (2014). Context influences conscious appraisal of cross situational statistical learning. *Frontiers in Psychology*, *5*(JUL), 1–9. DOI: https://doi.org/10.3389/fpsyg.2014.00691

Quine, W. V. O. (1960). Translation and meaning. In *Word and Object* (chapter 2, pp. 23–71). MIT Press, Cambridge, Massachusetts, new edition edition.

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roembke, T. C., & McMurray, B. (2016). Observational word learning: Beyond propose-but-verify and associative bean counting. *Journal of Memory and Language*, *87*, 105– 127. DOI: https://doi.org/10.1016/j.jml.2015.09.005

Sabbagh, M. A., Wdowiak, S. D., & Ottaway, J. M. (2003). Do word learners ignore ignorant speakers? *Journal of Child Language*, *30*(4), 905–924. DOI: https://doi.org/10.1017/S0305000903005828

Scofield, J., & Behrend, D. A. (2008). Learning words from reliable and unreliable speakers. *Cognitive Development*, *23*(2), 278–290. DOI: https://doi.org/10.1016/j.cogdev.2008.01.003

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning From Others: The Consequences of Psychological Reasoning for Human Learning. *Perspectives on Psychological Science*, *7*(4), 341–351. DOI: https://doi.org/10.1177/1745691612448481

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1–2), 39–91. DOI: https://doi.org/10.1016/S0010-0277(96)00728-7

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

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

Smith, L. B. (2000). Learning how to learn words. In Golinkoff, R. M. and Hirsh-Pasek, K., editors, *Becoming a word learner: A debate on lexical acquisition*, chapter 3. Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780195130324.003.001

Sobel, D. M., Sedivy, J., Buchanan, D. W., & Hennessy, R. (2012). Speaker reliability in preschoolers' inferences about the meanings of novel words. *Journal of Child Language*, *39*(1), 90–104. DOI: https://doi.org/10.1017/S0305000911000018

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. DOI: https://doi.org/10.1016/j.jecp.2014.06.003

Tomasello, M. (1992). The social bases of language acquisition. *Social Development*, *1*(1), 67–87. DOI: https://doi.org/10.1111/j.1467-9507.1992.tb00135.x

Tomasello, M. (2000). The social-pragmatic theory of word learning. *Pragmatics*, *10*(4), 401–413. DOI: https://doi.org/10.1075/prag.10.4.01tom

Tomasello, M., & Akhtar, N. (1995). Two-year-olds use pragmatic cues to differentiate reference to objects and actions. *Cognitive Development*, *10*(2), 201–224. DOI: https://doi.org/10.1016/0885-2014(95)90009-8

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. DOI: https://doi.org/10.1016/j.cogpsych.2012.10.001

Verga, L., & Kotz, S. A. (2013). How relevant is social interaction in second language learning? *Frontiers in Human Neuroscience*, *7*, 550. DOI: https://doi.org/10.3389/fnhum.2013.00550

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, *107*(2), 729–742. DOI: https://doi.org/10.1016/j.cognition.2007.08.007

White, K. S., Nilsen, E. S., Deglint, T., & Silva, J. (2020). That's thee, uuh blicket! How does disfluency affect children's word learning? *First Language*, *40*(1), 3–20. DOI: https://doi.org/10.1177/0142723719873499

Wickham, H. (2009). Elegant graphics for data analysis.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13–15), 2149–2165. DOI: https://doi.org/10.1016/j.neucom.2006.01.034

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

Yurovsky, D. (2018). A communicative approach to early word learning. *New Ideas in Psychology*, *50*, 73–79. DOI: https://doi.org/10.1016/j.newideapsych.2017.09.001

Yurovsky, D., & Frank, M. C. (2017). Beyond naïve cue combination: salience and social cues in early word learning. *Developmental Science*, *20*(2), 1–17. DOI: https://doi.org/10.1111/desc.12349