**Title: The impact of socio-economic variables on toddlers’ word comprehension: A touch-screen study in Argentina and France**

**Short title:** Word comprehension and SES across cultures

**Authors:** Celia, … , Alex 🡪 middle portion to be defined based on contribution as noted below

**Division of work:**

* Stimuli: Celia > French team (Charlotte > Laia > Camila) >= Argentina team (order tbd) > Alex
* Procedure: Celia + Alex
* Lit review: Alex + Celia
* Data gathering: French team (Charlotte > Laia > Alex > Camila) =< (all more or less equally up to daycares, if longitudinal included a lot more work done in Arg than in Fr!) Argentina team (order tbd)
* Data analysis: Alex *[ > Camila > Flor (planned) ]*
* First draft intro, exp 1-2: Alex
* First draft exp 3: tbd
* First draft discussion: tbd
* Comments: All

Last updated : May 25, 2016

TODO:

* redo all analyses using binomial models for pc
* reflect on whether to use standardized coefficients (to control for factors we have also collected) or just straight correlations
* reflect on “non-monolingual” category among French data
* collect more data in France
* add individual information for Arg data - N/A?
* decide whether to add longitudinal group or not…

# Abstract

A number of studies have documented differences in children’s language development in general, and vocabulary in particular, as a function of their primary caregiver’s level of education. Many of these studies are built on parental vocabulary checklists, language samples, or standardized tests, all of which instruments suffer from certain weaknesses. Moreover, it is rarely the case that non-American, non-English speaking populations are studied, and prior results on these other cultures do not support the generalized view that socio-economic status is a major determinant of lexical development. In this project, we present the results from three studies investigating young children’s word comprehension with a paired forced choice task built on a child-friendly touch screen, and administered in the children’s everyday environment. In Study 1, we tested 53 French 3-year-olds in three daycares in the south of Paris, and found that proportion of trials was not predicted by the educational level of their primary caregivers. In Study 2, 120 Argentinean 3-year-olds were tested in two daycares, one in a residential neighborhood and the other in an urban-marginal area. Results showed strong differences between the two groups. The discrepancy between Studies 1 and 2 may have been due to the latter children undergoing the test in very different settings, whereas the French daycares were not stratified by parental SES. Therefore, in Study 3 we tested young children from low- and mid-SES households in their own homes. Results show… ***to be continued!***

# Introduction

Nearly all children acquire language effortlessly in all known human cultures. One crucial aspect of language acquisition involves comprehending and learning words, a unique skill that appears to differ from other linguistic domains, such as syntax and phonology, in various ways. Firstly, many phenomena of syntax and phonology are productive: Once a child learns that their language follows a subject-verb-object word order, this can often be applied to the large majority of transitive sentences the child hears and produces; similarly, if the sequence “kn” is not encountered word-initially, then the child is safe to deduce that “tn” and “pn” will also be rare. In contrast, the very essence of words lies in the arbitrary pairing of form and meaning. While a vanishingly small proportion of any language’s lexicon may seem to be iconic, by and large, lexical acquisition relies on learning the random association between a sequence of sounds and a given concept. Secondly, whereas strong maturational effects have been uncovered for the latter two domains, suggesting that they rely on the emergence of relatively specialized neural pathways, lexical acquisition appears to continue to occur throughout our lifetime, it is dependent on a more widely distributed brain network, and it is thought to be a great deal more dependent on environmental factors than other domains.

It therefore comes as no surprise that wide variation in vocabulary size and speed of word comprehension can be found within populations. In their seminal work, Hart and Risley (1995) argued that one major source of variance in young children’s language development is in fact their caregivers’ socio-economic status (SES), with children whose parents are professionals having larger vocabularies than children born to parents on welfare. There are many reasons why SES could predict vocabulary, at variable levels of specificity. At a highly general level, food uncertainty, exposure to toxins, and poorer sleep quality are more widespread among less wealthy individuals (REF), and these organic factors could have generalized effects for instance on brain development (REF), and thus cause general attentional and memory problems that would interfere with learning in general, including vocabulary learning (REF).

In the other extreme of specificity, perhaps SES is correlated with certain parental behaviors that promote or hinder lexical acquisition. This is indeed, according to Hart and Risley (1995), *the* main reason why children from lower SES households never catch up with their more fortunate peers. According to their longitudinal work and that of others (e.g., Hoff, REF), input quantity, and to a certain extent quality, fully explain away SES effects on children’s productive vocabulary. These studies, however, often rely on children’s spontaneous speech samples gathered in the same conversations from which the parental input measures are extracted. In such cases, parental education may not explain any additional variance above and beyond parental speech directed to the child simply because the predictor and outcome are factually collinear. Indeed, one previous study (Rowe, 2012) used independent evidence for the children’s vocabulary outcomes (the Peabody Picture Vocabulary Test, to which we will return below), and found parental education to be a stable predictor of vocabulary above and beyond both quantity and qualitative variation in the input, at 30 and 42 months of age, with smaller effects at 54 months of age.

In short, there seem to be many indirect and general, and potentially also direct and specific, pathways through which parental socio-economic status should predict vocabulary, even in early childhood. Indeed, a rapid review of the literature could leave readers with the impression that this effect is profound and widespread. However, we believe the matter is far from closed, as previous evidence suffers from a number of limitations that will be detailed below. Before reviewing this literature, we would like to reiterate that properly describing the statistical relationship between SES and lexical development is not only interesting from a societal viewpoint, given that early vocabulary is an important predictor of school achievement (REF), but also on theoretical grounds, given the many factors noted above that predict large inter-individual differences should ensue from SES variation. To the extent that this predicted statistical relationship is not found, this would trigger further reflection on whether our theories of lexical acquisition hold up, or whether instead there is remarkable resilience in that somewhat peripheral, non-productive aspect of the language system.

*Previous work on SES effects on young children’s vocabulary: Three mainstream outcome measures*

As a function of the outcome measure used, there are three distinct bodies of literature reporting on the relationship between caregiver socio-economic status and young children’s vocabulary, a selection of which is shown on Table 1. The first body investigates how children’s spontaneous speech relates to SES, often with the goal of understanding the role that parental input plays in the process. The (near?) totality of this work has been carried out in the United States, where SES is often confounded with linguistic differences, namely increased degrees of bilingualism among poorer compared to richer samples. Even when bilingualism is controlled for, caregivers varying in SES may also vary crucially on cultural factors that relate to child verbosity. For instance, some have argued that African-American caregivers do not value or promote children’s volubility as much as European-American caregivers (REF). If children’s word processing is viewed only via a behavior that is less culturally appropriate, then such measures of knowledge are intrinsically biased.

A second body of literature uses parental vocabulary checklists to measure children’s lexical comprehension and production, and mostly the widely used MacArthur Communicative Development Inventory (CDI). The English CDI has two forms that have been normed for two distinct age ranges, and each form has been produced in two formats, a long one (about 300 words for 8-16 months, about 600 words for 16-32 months) and a short one (100 words in both cases). Adaptations have been carried out in several other languages, and a recent initiative has built a centralized, cross-linguistic repository containing literally thousands of individual children’s records. Reports on SES-vocabulary relations within this body of work are often incidental, and do not always show the expected positive correlation. For instance, in Fenson et al. (1995), the opposite was found for early vocabulary comprehension scores, leading the authors to hypothesize that caregivers with lower education levels were less realistic in their estimations of the child’s word knowledge and tended to overestimate this (CITE DAN’S CONFIRMATION?? Also, check wordbank to see if this holds up in all CDIs and countries etc.). This example illustrates one important issue with this instrument as a measure of children’s word processing, namely that it is subject to diverse biases as a function of parental SES and thus the relationship between childhood vocabulary and SES cannot be studied cleanly. It could be argued that these issues are less problematic when word production is assessed rather than comprehension, but in this case presumably the same cultural issues just identified, namely parental perception of the desirability of child volubility could emerge. At the very least, if parents do not promote children’s production by talking with them, they should have fewer chances of observing the child producing words.

The third body of work uses direct tests of children’s word comprehension, typically a standardized task administered by a trained experimenter and where the child needs to produce an appropriate behavioral response for this individual. By far, the most common instrument in this case is the Peabody Picture Vocabulary Test (or PPVT), which has been standardized in the United States, and adapted (and sometimes also normed) in other countries. The PPVT is administered by showing the child a set of 4 pictures at a time, and asking him/her to “point to the X”. Undoubtedly, children who are more used to such a task (e.g., who are engaged in joint book reading and whose parents prompt pointing) will have an easier time with this task. Additionally, it has been argued that the very activity of asking “obvious” questions is culturally biased, in that this behavior is seldom observed in certain cultures or subcultures (REF). The forced choice task is repeated for a maximum of XX sets of pictures, which are sorted in terms of difficulty, and the task is stopped when the child produces XX errors in a row. The vocabulary level is derived from the maximum trial number the child could solve correctly. As a consequence, differences in attention and ability to concentrate on a repetitive task will lead to lower scores, even if the children are equally good at word comprehension (REF? or unnecessary?).

*Two novel laboratory-based measures*

A very recent strand within this third body tries to circumvent this problem by either proposing children a passive task that requires no overt response, or by developing an active, but ecological task. We review both, but will ultimately focus on the latter approach as the most promising one.

Fernald, Marchman, and Weisleder (2008) followed the former route, using an implementation of the looking-while-listening paradigm to investigate word comprehension in 48 English-learning toddlers tested in one of two sites within the state of California. One of the sites was close to Palo Alto and the other was in Northern California, and thus they drew from populations whose salaries, education levels, and racial/cultural/linguistic composition was widely diverse. All children were selected after interviewing the parent on the phone to exclude children with a history of health issues or premature birth, or who were exposed to languages other than English. Additionally, children were classified into a “low” or “high” SES using the family’s Hollingshead Index. Maternal education thus could overlap across the groups, and both lower and higher groups had a range encompassing 12 to 18 years of education, although the averages were significantly different, at 14 (some college) versus 17 (post-baccalaurate) years. Children were tested twice, at 18 and 24 months, and both times the infant sat on a caregiver’s lap in front of a large screen, and viewed up to 32 (at 18 months) or 16 (at 24 months) experimental trials (additional filler trials were included – see paper for further details). In each trial, a pair of pictures was shown, and shortly after the audio asked “Where’s the X? Look at the X”. There were 8 lexical items used, all judged to be well-known to most toddlers: *baby, doggy, birdie, kitty, ball, shoe, book,* and *car*; moreover, if a given child did not know a word, according to parental report, then all trials including that item were excluded from consideration. These items were repeated counterbalancing side and/or which item served as target or distractor four times for the 18-month test, and twice for the 24-month test. Accuracy was defined as the proportion of time the toddler looked at the picture matching the verbal prompt within a certain window. Additionally, if the child was looking at the distractor at the onset of the keyword and shifted to the right picture within the analysis window, then a measure of response time was defined as the latency between the word onset and the beginning of the visual shift. Additionally, caregivers filled in a CDI. Results showed that SES predicted all three lexical measures: Children whose parents had a higher SES had larger vocabularies, higher accuracies and shorter response times, at both 18 months and 24 months (with no age interaction). Interestingly, the correlation coefficients for the two direct measures (e.g., at 18 months r=.52 and .5, respectively) were higher than that found for the parental checklist (r=.34). In subsequent work, Fernald and colleagues have replicated the above-cited finding that quantity of speech directed to the child predicts vocabulary size, but have further shown that the child’s individual response time measure drawn from this task mediates that input-outcome relationship, which the authors interpreting as the potential mechanistic pathway (input influences speed of processing, and therefore informational uptake, and therefore lexical development).

One potential limitation of this otherwise promising and ecologically valid technique is that it requires specialized equipment, as families need to come into the lab for the child to perform the task in a darkened room. As a result, children from different SES could react to such a novel situation quite differently. Additionally, the task carefully isolates processing speed by using highly known words, that are furthermore repeated. It is, in this sense, not surprising that in other work Fernald and Marchman (2010?) find that this response time measure is a good predictor of childhood IQ. Indeed, it is unclear to what extent the predictive power of this task or the individual differences it measures relate exclusively or primarily to lexical processing rather than information processing more generally. It would be interesting, in the future, to distinguish the two by using a non-speech information processing task, or perhaps by using diverse lexical items and showing that response times are sensitive to lexically relevant variants.

This precise feature is found in another laboratory task, this time using a large touch screen. The Computerized Comprehension Task (CCT) has been developed and utilized for over more than 10 years by Friend and colleagues, who have used it to look at word comprehension in toddlers from diverse linguistic and social backgrounds. In the CCT, the child sits by an experimenter in front of a large touch screen, on which two images appear. Shortly after, the experimenter produces a verbal prompt “Where is the X”, which is repeated after 3.5 seconds if the child does not produce a response. If after 7 seconds there is still no response, the images disappear and the next trial starts. When a touching response is produced, a touch to the prompted target will entail a positive reinforcement (from the screen and the experimenter), whereas a touch to the competitor will have no effect. After 4 training trials (which can be repeated if the child misses), the experimenter presents, in a pace appropriate to the child and until he/she seems no longer interested in the task, up to 41 trials, which contain nouns, adjectives, and verbs as targets. Although all of these words can be found on the CDI checklist, a third of the words within each lexical category are words that are typically produced by most toddlers; another third are produced by about half of the toddlers, and the final third is more rare for this age. Given the structure of the task, the authors can extract a number of dependent measures: The number of trials the child has completed, the number for which the child actually attempted a response, the proportion of responses that were correct, and the response times. Over several studies, Friend and her collaborators have documented the results from all these dependent measures in English-, Spanish-, and French-learners, typically aged between 15 and 21 months. Moreover, in some papers, when children remained engaged after the 41 experimental trials, a third of the trials was presented again, allowing an estimation of immediate test-retest reliability.

Overall, Friend and colleagues find that older children complete and attempt more trials, have higher accuracy scores, and lower response times. In our view, one reason why this technique is preferable for the study of early word comprehension over the looking-while-listening implementation just discussed is the fact that, quite consistently, Friend and colleagues find better performance for the items selected for being high frequency over those that are rarely attested in children’s CDIs; and for nouns over verbs and adjectives, which also mirrors patterns of lexical development that have been attested for comprehension and production in the languages Friend and colleagues study.

Although the CCT, like the Peabody, relies on an overt behavioral response triggered by the experimenter, we believe two design features make it preferable over the other standardized task. First, since “easy” and “difficult” items are mixed together, the task allows a separation of accuracy from number of trials completed/attempted, thus leveling the ground for children varying in executive functions. Second, whereas the PPVT is administered on paper (thus resembling a task that may be more or less common to the different children), the touch screen actually produces a contingent response on the screen that could become reinforcement in itself, and thus increase the relevance of the task for the child. A recent study by Frank and colleagues (in press) directly compared word comprehension in 1- to 5-year-olds measured with these three instruments. After testing over 200 children in one of the three, the authors conclude that a task like the CCT preferable over both the looking-while-listening and a paper-based PPVT-like task, based on the combination of robust psychometric properties and attractiveness to the child. Finally, in a direct comparison of the predictive power of CDIs and CCT data from the same children with respect to later expressive language from a spontaneous sample, Friend and colleagues find that the CCT data is more predictive than CDI, further bolstering the interest of using such a measure.

A thorough reading of the studies using CCT reveals only one instance in which SES was discussed. Deanda and colleagues gathered CCT and CDI data from 72 American 15- to 18-month-olds exposed mostly to English, varying in SES as well as degrees of exposure to other languages (completely monolingual versus hearing 5 to 15% other languages). Controlling for the latter, SES explained a significant proportion of variance in CCT accuracy, whereas this relationship was only marginally significant for the CDI scores. In a second experiment, the authors compared a group of children learning only Mexican Spanish against another group exposed to English marginally, with SES also varying. In this second sample, there was no effect of SES on CCT performance. The authors interpret this as a true null result, and cite two sources of convergent evidence, a study by Hurtado et al. (2008) that replicated Fernald et al. (2008)’s looking-while-listening paradigm, but not the SES results, on a Mexican Spanish-speaking sample, and a study with a Mexican Spanish adaptation of the CDI (Jackson-Maldonado et al., 1993). The authors argue that this differential degree of statistical association between SES and toddler lexical development in American English-speaking and Mexican Spanish-speaking population is due to divergent rearing practices, citing work suggesting that the latter “encourage obedience and collaboration more so than verbal communication and independence” (p. XX).

*The state of the question*

There are certainly numerous studies that mention SES differences, but as a whole the literature remains unconvincing. As we argued above, the literature suffers from potential confounds with cultural and linguistic differences, and limitations in terms of the independence of the instrument and the outcome. We would like to add two criticisms to the body of work as a whole, which are not meant to be critical of any one study. First, as we noted above, the great majority of this work has been carried out in English, with American samples. As a result, this work builds on relatively small levels of SES variation, since caregiver education in such work often ranges between high school and college, and sometimes only between some college and post-graduate education. Yet in many countries in the world, only primary education is mandatory and widespread; in some, it is even common to only complete a few years of formal education. As a result, these samples can certainly be described, to adapt a common phrase (REF), as WEIRE: Western, Educated, Industrialized, Rich, and English-speaking. It seems risky to assume that the workings of the human cognitive system serving lexical acquisition can be fully understood by the study of such a specialized sample alone, thus inviting further research, with multiple methodologies, in other countries.

The second worry we have is that there is a problem of selective reporting. We base our conclusions on work mostly confirming an expected positive relationship between caregiver education and children’s vocabulary, yet it is quite possible that many other researchers or studies have tested for an association between e.g., maternal education and vocabulary, but not reported it due to the fact that the correlation was not in the expected direction or was not significant.

*The present work*

To address all of these considerations, and contribute some much needed evidence on this important question, we carried out a series of experiments in France and Argentina. All of them used an adaptation of the CCT to a portable touch screen. We made this design choice so as to be able to use the exact same equipment in both countries, and to facilitate others’ reproducing it elsewhere. This design choice has the further advantage that we can go to the child’s everyday environment, and thus potentially reduce the effects of facing a laboratory task.

# Experiment 1: Individual variation within French daycares

In our first cross-linguistic extension, we investigated word comprehension among young children growing up in Paris, France. Not only would this cohort provide data on SES effects in French-learning children, but also this extension was particularly interesting because France and the United States are roughly similar in terms of gross internal revenue and level of development (cite OECD data), but very different along other dimensions. Specifically, families with lower incomes benefit from much greater degrees of support than their American analogues, which is evident in terms of the living aid, health care, and child care that are accessible to the population (cite OECD data).

As for our word comprehension measure, as was explained above we followed work from the CCT, and share with this test the general structure of the test, the number of lexical items and their distribution into word categories and difficulty levels, and the use of a touch screen to collect responses. We diverged from that work in several dimensions, which will be detailed in the Methods section. Most importantly, we used a personal touch screen (iPad ®) rather than a large touch screen, to be able to bring the equipment to the child, rather than having families visit the lab.

Indeed, we collected data in three daycares located close to the south of Paris *intra muros*, a relatively popular area characterized by high buildings and a large proportion of immigrants. We hoped that by targeting this neighborhood, we would be able to sample from a population that was varied in income and education. Nonetheless, the range of SES that can be found in daycares established within Paris will probably be smaller and biased towards the higher end than that found in France at large, and thus our results should not be interpreted as representing national variation. A potential issue with our data collection site was the fact that children could be growing up in multilingual households, and we therefore collected detailed information on children’s language exposure via a questionnaire.

Add EDEN data, CDI data…

## Methods

### Participants

A total of XX families agreed to allow their child to participate. The parents of XX of these failed to turn in the questionnaire by the end of the study phase, and we therefore have no SES information on them. Two children refused participation, and an additional one could not be tested because they were absent throughout all visits. Thus, there were 51 children with at least some touch screen data.

### Stimuli design

The stimuli for this French experiment was designed together with the Spanish stimuli described in Experiment 2. In both tests, we followed the CCT in having a few pairs of “easy” words used during an initial training phase (3 in this French version), and 41 pairs of words varying in difficulty, for a total of 44 pairs or 88 lexical items. As in the CCT, 53.7% of the test pairs were nouns, 26.8% were verbs, and 19.5 % were adjectives. These percentages are based on the results of Fenson’s et al (1994) regarding toddlers’ vocabulary knowledge assessed through CDI. See Table 1, in which the distribution of lexical pairs by word class and level of difficulty is presented.

Table 1: Distribution of pairs as a function of word category and difficulty level.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | nouns | verbs | adjectives | total |
| easy | 7 | 4 | 3 | 14 |
| moderate | 7 | 4 | 3 | 14 |
| difficult | 8 | 3 | 2 | 13 |
| total | 22 | 9 | 8 | 41 |

In the CCT, difficulty was defined in terms of the proportion of children who knew the words according to the CDI-based parental report gathered at 16 months in a separate sample: an item was classified as easy if it was comprehended by more than 1/3 of children, difficult if by less than 1/3 and moderate otherwise. This method thus relies crucially on ample CDI data for the relevant language, something that we did not have access to for French (and which is non-existent for Argentinean Spanish).

We therefore used frequency in child-directed speech in our difficulty classification. The frequency estimates were drawn from publicly available French-language corpora in CHILDES (Lyon REF; Paris REF), which contain XX transcripts from XX children aged XX and XX, for a total corpus size of XX. The transcripts were lemmatized using the CLAN (REF) command mor, and frequencies were added across all forms of a word type, a procedure that seemed appropriate given the fact that inflections in French often do not result in form difference (e.g., *chien* ‘dog’ and *chiens* ‘dog’ have exactly the same pronunciation). We made sure that the frequency distributions within a given word category did not overlap for easy, moderate, and difficult pairs, although it was impossible to match frequency across word categories within a difficulty level. Table XX shows the distribution of frequency by word category.

## The selection of the 88 words (44 pairs) was made trying as much as possible to use items that were present in the CCT and PPVT, as explained in more detail in Experiment 2. We first created the lists for the Argentinean Spanish test because there were more constraints to follow, and we adapted that list to French by re-classifying or replacing lexical items if needed (e.g., if an item selected as Easy for the Argentinean test actually had a frequency in the French corpus within the Moderate range, then it was treated as a Moderate item, rather than an Easy one; if the item was not present at all, then it was removed from the list). As in the CCT, there were two forms of the test, which differed in terms of which of the two members of the pair (chosen at random) would serve as target (the other serving as the distractor). Table XX shows the final selected pairs.

INSERT TABLE HERE – make sure to be clear RE form 1 and form 2

### Stimuli implementation

Images were mostly photographs rather than drawings, and they created by us or found using the images.google.com search set to find images with right to reuse. Images were edited using PowerPoint and Adobe when needed by removing or adding a background, cropping, etc. Audio stimuli were recorded from a native French speaker, using EQUIPMENT. We edited the stimuli by removing trailing silences and normalizing amplitude using Praat, and converted them into the mp3 required by our presentation program using Audacity. All nouns were embedded in the phrase *Touche le(s)/la …* ‘Touch the …’ (using the appropriate article for each noun). All verbs were embedded in the phrase *Il est où, celui qui…* ‘Where is the one that …’. Finally, adjectives appeared pre- or post-nominally as it was natural in French, making sure that the noun used did not give cues to the picture; for example in the pair fatigué-triste, the phrases were *Touche l’homme fatigué* ‘touch the tired man’ versus *Touche l’homme triste* ‘touch the sad man’.

The training pairs were always shown at the beginning in a fixed order. The remaining 41 pairs were shown in a fixed pseudo-randomized order where no more than 3 trials in a row were moderate or hard, and no more than two times in a row the correct response was on a given side.

### Procedure and equipment

The computer program used to deliver the stimuli and gather the response had been developed in our lab and used in previous work (e.g., Dei et al., 2016). Data collection occurred in 2-3 visits to each daycare, and each time 2-4 members of our team were present. Typically, two of us (CM and LF) would administer the task and the rest would walk the children in and out, take notes, and liaise with the daycare staff. Children would therefore be tested 1-2 at a time in a quiet room, usually the library, while sitting on age-appropriate chairs.

After welcoming the child, the experimenter interacting with the child would show him/her a game (REF) on the ipad that consisted on popping bubbles. This game was repeated until the experimenter felt the child was comfortable with her and the equipment, usually about 1-2 minutes. The experimenter would then propose another game and explain that for this one there was sound, and therefore the child would need to wear headphones. We chose to deliver stimuli this way in order to be able to administer the test to two children at once without their respective sounds interfering, and also in the hope that this would establish a sort of joint game with the experimenter, minimizing other distractions, notably the other child playing. The experimenter also wore headphones, as the sound was being delivered via a splitter (see Figure XX). Most children agreed to this; in two exceptions, the experimenter would hold the headphones by the child’s head so that she/he could hear the sounds. At this point, the vocabulary test began.

The structure of each trial is exactly the same throughout training and test phases: The left image appears, then it becomes less neat and the right image appears; next, the images are covered with a transparent gray layer while a character on the screen appears. To pace the trials taking the child’s attention into account, the child or experimenter needs to tap on the character to launch the prompt phase. Furthermore, it is only when the prompt phrase finished that the gray layer disappears and the child can respond; all touches prior to this point are ignored by the application, thus discouraging random touching that does not take into account the command. A tap on one of the objects, regardless of whether it is the target or the competitor, causes a “plop” sound, that image to be neater, and the other to disappear. Additionally, if the tap was on the correct response, the character jumped up and down happily providing positive feedback, e.g. “yes, that’s it”. If the tap was on the wrong response, there was no change on the screen. Additionally, the CCT specifies that the experimenter repeats the prompt phrase after 3.5 seconds without response. Following this concept, in our test, if there was no response after 4 seconds, the character acted impatient and said something like “are you still there?” This was actually the cue for the experimenter to repeat the prompt phrase. In CCT, after 7 seconds the screen goes blue and the trial is considered “not attempted”. Similarly, if there was still no response after another 4 seconds, the experimenter said “the \_\_ was here” and touched the right image. This, however, happened very rarely. If at any point the child seemed disinterested (looking away, failing to respond), the experimenter reminded him/her he/she can stop anytime he/she wants.

As with the CCT, there were a few training trials, which used highly familiar words, at the beginning – in our case, we used 3 training trials. There are two main differences between the training and test trials. First, experimenters provided a great deal of feedback during training, but merely provided positive feedback seldom at random points during the test phase proper. Second, during the training, but not the test, phase if the competitor rather than the target was touched, the trial was repeated immediately. To proceed to the testing phase, the participant needed to have responded correctly to the very last training trial. The test trials were presented until the last trial was completed, or if the child said they wanted to stop. Prior to testing we had also decided to follow the CCT procedure of declaring the test over if the child failed to produce a response 4 trials in a row, but this criterion was never met.

It should be mentioned that we weaved together 2 training and 12 test trials from a separate experiment looking at learning of minimally different words. Those data will not be discussed in detail here; see Fibla (2016) further information.

### Questionnaire

Parents were asked to fill in a French adaptation of the MacArthur CDI (the CDFI, Kern, XXX), in its short form adapted to 24-month-olds, as well as a short questionnaire on language exposure, background health, and SES-relevant information. The first group of questions covered exposure to French versus other languages at 0-12, 12-24, and after 24 months of age. Children with at least 75% exposure to French at all three ages were classified as monolingual; children with at least 25% exposure to exactly two languages (including French) over this time were classified as bilinguals; the rest were termed “other”. Caregivers were also asked if the child had passed a hearing test, received a medical diagnosis that could entail language delays, and had colds recently. Finally, two questions were relevant to the SES analyses, one asking about the education level about the three primary caregivers, and another about their occupation. Prior to data inspection, it was decided that, for comparability with previous literature, only maternal education would be employed in analyses.

### Analyses

Prior to data inspection, we decided that children should have 9 or more trials to be included in analyses of overall performance (e.g., accuracy and response times for correct responses, henceforth RT); and at least 4 trials to be included in any within-participant analyses (i.e., separating response times for correct and incorrect responses, or separating accuracy and RT by word category or difficulty level). These two criteria of minimum trials were set a priori; inspection of the distribution of trials attempted confirms that these cutoffs are below the first quartile for the whole experiment (24.5 trials, out of a possible maximum of 41); similar to the first quartile for difficulty level (8 trials, out of a possible maximum of 13-14) and lexical category (12.5 trials, out of a possible maximum of 8-22, see Table XX).

All statistical analyses were carried out in R (REF). To study the prediction of word comprehension performance from maternal education, we investigate all the dependent measures from the experimental task that can be calculated at the individual level, namely: total number of trials completed, total number of trials attempted (i.e., in which the child produced a response within 7 seconds), accuracy (number of correct trials divided by total number of trials attempted), and median response times. Additionally, we tallied the total number of words that the parent reported the child to understand, or produced, from the CDI.

Given the large number of dependent measures assessed, we do not interpret our statistics inferentially, but instead use them as descriptors. We report Spearman correlations as the effect size estimate between our outcome measures and the raw education scores, which had a bimodal distribution. For more detailed analyses, we used a median split, with 7 years of post-secondary school education as the cutoff. It should be noted that, in this sample, language exposure and maternal education were not associated: the higher education group contained 8 monolinguals and 10 non-monolinguals, and the lower median group 8 monolinguals and 12 non-monolinguals. The logarithm base 10 was applied to the median RT individual scores, which resulted in them having a normal distribution. We subtracted .5 from proportion correct and centered age to more easily interpret the intercept in regressions, but, for clarity, all illustrations use raw scores. All regressions were fit with the base package (REF), and controlled for age (centered) and a dummy variable coding whether the child was French monolingual or not; for example lm(nb\_trials\_completed ~ medianMaternalEd + age.c + monolingual\_dummy). Analyses investigating within-participant interactions employed mixed models from the lme4 package (REF), declaring the participant as a random factor, the same regressors just mentioned (centered age, language exposure) as well as the factor coding for the median split on maternal education in interaction with the within-subject factor; for instance lmer((proportion\_correct - .5) ~ medianMaternalEd \* difficulty\_level + age.c + monolingual\_dummy + (1|partincipant\_id)). An Analysis of Deviance (Type II Wald χ2) from the package car (REF) was used to determine significance level in these cases.

## Results

We begin by providing a few key global facts that may be deemed sufficient by some readers as a response to our research questions. We found that, by and large, the task was well-received with an average of 31.6 trials being completed. There was some evidence of a side bias, since 18 out of 50 children with at least 9 attempted trials responded on one side more than 60% of the time. This bias was slight, with an average 57% of attempted trials receiving a right-side response, when collapsed across all children. Accuracy and RT had good Cronbach’s split-trial reliability, according to calculations using the alpha function in the psych package (REF). For accuracy, these calculations revealed an average r = .74 and an alpha = .85 (95% confidence interval .77-.93); for RT, they were average r = .78, alpha = .87 (95% confidence interval .8-.94). Bivariate associations between performance and maternal education were non-significant, but by and large not null. The Spearman correlations between maternal years of education and outcome measures were as follows: trials completed **ρ**(49) = .223, *p >* .05; trials attempted **ρ**(49) = .220, *p >* .05; proportion correct **ρ**(45) = .271, *p >* .05; log response times (for correct responses) **ρ**(45) = .051, *p >* .05. ADD CDI

*Detailed analyses of overall performance*

Neither the maximum number of trials nor the number of trials attempted yielded a significant regression model, although the latter was marginal, *F*(3,34) = 2.622, *p =* .066 (multiple R-squared .188). In this model, number of child responses was not predicted by SES or the dummy variable coding language exposure, but it was predicted by age (ß = 10.587, SE = 5.164, *t =* 2.05, p <.05, due to children attempting more trials the older they were).

A significant regression equation was found for accuracy *F*(3,32) = 10.71, *p <* .001 (multiple R-squared .501). The overall proportion of correct trials was significantly higher than chance (ß = 0.240, SE = 0.039, *t =* 6.19, p <.05), increased with age (ß = 0.298, SE = 0.067, *t* = 4.453, p <.001), and was higher for children who were monolingual compared to bilingual/other (ß = 0.106, SE = 0.046, *t =* 2.31, p <.05); again, no effect of the maternal education variable was recorded.

A similar multivariate linear regression model was not significant in an analysis of the median response time for correct trials. Similarly, a mixed model fit declaring the within-subject factor response accuracy (correct, incorrect) in addition to the other regressors (centered age, language exposure, education) revealed only a main effect of response accuracy (χ2(1) = 4.275, *p* < .05), due to longer response times when the child touched the competitor (incorrect) than the target (correct).

*Detailed analyses of interactions with trial types*

Accuracy was affected by difficulty level (χ2(2) = 8.088, *p* < .05; due to somewhat higher scores for moderate and easy trials compared to difficult trials), age (χ2(1) = 19.007, *p* < .001; with significantly better scores for older children), and language exposure (χ2(1) = 4.893, *p* < .05, monolinguals scoring better than bilinguals), but there was no effect of maternal education or interaction between difficulty level and maternal education. Response times were only significantly affected by difficulty level (χ2(2) = 8.871, *p* = .01, due to somewhat shorter RTs for easy and moderate than difficult trials), with age and language exposure being only marginal.

The effects of lexical category (χ2(2) = 24.148, *p* < .05), age (χ2(1) = 19.769, *p* < .001), and language exposure (χ2(1) = 7.837, *p* < .001) were all significant, whereas maternal education and the interaction between lexical category and maternal education were not. Response times varied as a function of lexical category (χ2(2) = 45.209, *p* < .01) and language exposure (χ2(2) = 4.853, *p* < .05).

ADDITIONAL ANALYSES USING PROFESSION

## Discussion

We observed a small, but consistently non-significant relationship between maternal education and a broad range of measures, all drawn from a forced-choice lexical recognition task implemented on a tablet. Some of these measures were probably affected by executive functions (e.g., number of trials completed, difference in response time for correct and incorrect trials); others were more specific to word comprehension (overall accuracy, difference in response time for different word types). The largest effect of maternal education on child performance was found for overall accuracy, with a **ρ** of about .27, suggesting that only 7% of variance was explained by this predictor. By comparison, the Spearman correlation between age and the same outcome measure was **ρ** = .585, thus explaining 34% of the variance.

We can be certain that the weak association found between maternal education and child performance is not due to our measures being insensitive, since they pick up individual differences attributable to age (as just mentioned) and prior language exposure. Additionally, Cronbach reliability was very good in the two measures where it could be calculated. Furthermore, it is not the case that our stimuli poorly designed since children’s behavioral measures reflected differential processing effort associated with words varying in lexical category or in their frequency in children’s input.

One possible criticism pertains our proxy for socio-economic status, namely maternal education. Certainly, many other indices are commonly employed to measure socio-economic status, including profession and income, of which we collected only the former. IF WE ADD ANAS, THEN SUMMARIZE THEM HERE. Nonetheless, most previous work on SES effects on child language has used this same proxy of maternal years of education. Its predictive power may have been small in our sample for at least two reasons.

As to the first reason, relates to the way the sample was recruited, namely via community daycares. It is possible that this sampling method reduced the variance among the children, many of whom have spent 3-5 days a week for at least the previous 6 months in a common environment. In other words, daycares could have provided a buffer that provided children with such strong common experiences that prior differences associated with maternal education would be reduced. If that were the case, we could have seen differences across the three daycares, since children in different daycares could have been provided with very different experiences. However, additional analyses (available on the supplementary materials) revealed no such effect, lending little credence to this possibility.

A second explanation relates to the distribution of education within the sample. Although the range of variation was considerable (9 years), the minimum was at 2 years post-secondary school education. Thus, it is possible that additional maternal education beyond this level is associated with only small differences in children’s vocabulary size. While we do not at present have evidence to address this hypothesis for French toddlers, we did collect data sampling from a more extreme distribution in Argentina, as detailed next.

# Experiment 2: Comparison of two daycares in Argentina

* Argentina is an appropriate place to study the question because
  + Stronger variation in education PROVIDE CENSUS OR OECD DATA??
  + In the past, we’ve observed PPVT differences (citation?); others have worked on cognitive skills of poor populations & shown strong impact of unfavorable conditions (Lipina)
  + So we can expect bigger differences

This first experiment compares two daycares, which, in view of their locations, allow us to sample from two very distinct populations. MORE DETAILS

Stimuli

We used a corpus of spontaneous speech (footnote) to select easy, moderately difficult and difficult words. This corpus consists of 468 hours of audio-recorded interactional situations in which 39 four-year-old children of middle and low income social groups participated in their homes (12 hours in each household). Using similar criteria to the CCT, we considered the word to be “easy” if it appeared in the input of 66% of the children, “moderately difficult” if it appeared in the input of between 33% and 66%, and “difficult” if it appeared in less than 33% of the children’s input across both social groups. This allowed us to assume that the opportunity to hear the word was similar for children from both social groups, in an effort to create a culturally non-biased test.

## tests were present in the children’s input in the corpus mentioned above. We selected 31 easy, moderately difficult and difficult words from the CCT which met the criteria explained above and 14 from the Peabody. We found the remaining words needed to complete the pairs by searching directly in the spontaneous corpus, taking into consideration class of word, gender, number, animacy, and picturability to find additional lexical items and create matched pairs.

## As the selection of words we made was based on the input of children who were 4 years old, we assessed the validity of this selection for younger children (XX). We asked 30 parents from both social groups to report toddlers’ knowledge of the words. We excluded from the test those words in which there was no relationship between the level of familiarity that arose from the parents’ reports and the analysis of how the word varied in frequency in the input of the 4-year-old children.

# Experiment 3: Individual variation in Argentina

Why: huge differences in exp. 2 but perhaps this has to due with environment, it could be that one daycare is better than the other, or kids more distracted in one place than other by overhearing the stimuli from the other child. Not likely given the analysis we did in France, but we cannot be certain. So now we test in people’s homes.

Additionally, we did not gather detailed information from the children’s background, so we cannot test what exactly in the socio-economic background most effectively related to vocabulary differences. So now we also gather this info.

Participants

Notes from Celia, better to use info from data sheet: Children are mostly monolingual.

It is true that, the majority of middle and high income families comes from European-old-inmigrant population. Nevertheless, people living in villas de emergencia or in poor neigborhoods in Buenos Aires are in many cases inmigrants from Paraguay, Peru and Bolivia and descendants from indigenous communities.

Also the aprox. 10 babies we are following in Ptte. Derqui are from the Qom community although their parents are mostly Spanish monolingual

Results

* Regressions with maternal education; and our other conceptualizations of SES