


ARTICLE

Adapting language development research paradigms to online testing: Data from preferential looking, word learning and vocabulary assessment in toddlers

Delphine K-L. Nguyen , Nadine Fitzpatrick and Caroline Floccia

School of Psychology, University of Plymouth, UK

Corresponding author: Delphine K-L. NGUYEN; Email: delphine.nguyen@plymouth.ac.uk

(Received 24 October 2022; revised 08 January 2024; accepted 02 February 2024)

Abstract

During the recent pandemic, it became necessary to adapt lab-based studies to online experiments. To investigate the impact of online testing on the quality of data, we focus on three paradigms widely used in infant research: a word recognition task using the Inter-modal Preferential Looking Paradigm, a word learning task using the Switch task, and a language assessment tool (WinG) where children identify a target word amongst a set of picture cards. Our results for synchronous and asynchronous studies provide support for the robustness of online testing. In Experiment 1, robust word recognition was found in 24-month-old toddlers. In Experiment 2, 17-month-old infants consistently learned a new word. Finally, Experiment 3 demonstrated that 19- to 26-month-old children performed well on a language assessment test administered online. Overall, effect sizes or language scores were found to be higher than in lab-based studies. These experiments point to promising possibilities for reaching out to families around the world.

Keywords: language; development; children; online; validation

Introduction

Online research studies have become more popular among developmental researchers since the COVID-19 pandemic (Rhodes et al., 2020; Sheskin et al., 2020). Due to COVID-19 restrictions, studies were not able to be conducted in person but thanks to videoconferencing technologies, many research experiments were run remotely (Blanchard, 2020; Delgado et al., 2021; Mills et al., 2022). There are important potential benefits and promises from using videoconferencing: flexible time and space which benefit both the participant and the researcher, as well as the possibility to widen the scope of participant recruitment, enhancing inclusivity and allowing for a better representation of diversity. However, there might also be some pitfalls in the use of online testing, first and foremost related to the quality of data (due to technological limitations, interruptions, etc). When

© The Author(s), 2024. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

considering infant and toddler language research, where the majority of responses rely on accurate looking time measures, these pitfalls are to be considered carefully. Also, we wondered whether levels of engagement from the participant would be possibly higher (familiar environment, more attentive) or poorer (less controlled setting, less motivation). Indeed, according to an editorial review by Tsuji et al. (2022), online data collection might be more prone to being noisier due to uncontrollable variables such as distractions, lighting conditions, and the quality of recording devices. However, they also reported that it is worth considering that children might feel more at ease in their home environment, which could potentially result in less variability in measurements during online data collection.

Every researcher using an eye tracker in a lab setting has experienced the complexity of minimising a child's head movements and removing external distractions to optimise data quality; therefore, it is potentially challenging to tackle these issues when testing remotely too. Many researchers, including ourselves, worry that remote testing would fail because of the lack of control over motion, parental interference, distraction, equipment, etc. The question we ask in this paper is as follows: given the minimal amount of constraint we can apply to children's movements in a remote situation, and the difficulty to control for external distracting factors, can we still collect data from classic paradigms of early language studies that compare in statistical robustness to what we would obtain in a lab situation? Previous studies aimed to answer that question by testing whether specific paradigms could be adapted to online settings (see review by Tsuji et al., 2022). For instance, Bochynska and Dillon (2021) did not successfully replicate findings from the lab. They conducted two asynchronous online experiments where they adapted the change-detection looking-time paradigm with infants aged 7 months. Their findings indicated that the infants did not show detectable sensitivities to the basic shape information that differentiates between 2D geometric shapes, which contrast with previous lab experiments results. They reported that failure to discriminate between shapes might be due to distraction and infants having difficulties perceiving two distinct events when displayed on small compact screens of personal computers. Indeed, for this paradigm, most lab studies used two separate monitors or large projector screens (Bochynska & Dillon, 2021). On the other hand, Bánki et al.'s (2022) study successfully tested infants (aged 4-6 months) in an eye-tracking task that measures the detection of audio-visual asynchrony. They found a higher quality of webcam-based eye-tracking data collected online and no differences in participant attrition rate and technical issues between the in-lab and online context. In addition, Bacon et al. (2021) found that children's (aged 23 to 26 months) word recognition accuracy on the online synchronous looking-while-listening task was greater than accuracy on the in-lab task. Furthermore, Bulgarelli and Bergelson (2022) investigated, with both in-lab and online experiments, how talking variability (e.g., a new talker of another gender produces the word) during learning could potentially influence children's (aged 7-9 months) ability to learn and recognise words. Using a one-word Switch task paradigm, results collected online and the results collected in the lab were fully similar. The researchers reported a few limitations of testing remotely such as not being able to control the distance to the screen device or the size of the monitor, but concluded that the one-word switch task could be easily adapted for online testing and provide successful results.

This paper adds to this body of knowledge in a number of ways. First, we aim here to demonstrate that effects such as increased looking behaviour modulated by linguistic cues are measurable in children doing the task online and provide benchmarking data between online and lab-based studies, to provide guidance for the design of future studies. We also

Table 1. Overview of the three experiments

| Experiment | Paradigm | Task | Adaptations to in-lab procedure | Children |
|--------------------------------------|--------------------------------------|---|---|--|
| Experiment 1: Word recognition | Intermodal preferential looking task | <ul style="list-style-type: none"> - Replicated from other labs - Online, Gorilla - Asynchronous | <ul style="list-style-type: none"> - Greater number of trials than comparable procedures - Trials are not infant-initiated | <p><i>N</i> = 20</p> <p>24 months</p> |
| Experiment 2: Word learning | Switch task | <ul style="list-style-type: none"> - Replicated from other labs - Online, Zoom - Synchronous | <ul style="list-style-type: none"> - Lower number of trials than comparable procedures - Familiarisation instead of habituation | <p><i>N</i> = 19</p> <p>17 months</p> |
| Experiment 3: Language assessment | WinG test | <ul style="list-style-type: none"> - Replicated from own lab - Synchronous | <ul style="list-style-type: none"> - Similar than in-lab task - Comparison online vs in-person | <p><i>N</i> = 62</p> <p>19-26 months</p> |

explore modifications to accepted in-lab procedures, such as increasing the number of trials and using automatic trial presentation, in place of the standard infant-initiated trial start (see Experiment 1). We chose three paradigms which are widely used in infant research: a word recognition task using Intermodal Preferential Looking (IPL, or look-while-listening procedure), a word learning task using the Switch task, and a language assessment tool relying on children identifying a target word amongst a set of picture cards. For each of these tasks, we conducted an online, simple experiment, whose results we compared to existing data collected face to face by our lab or other labs in the pre-pandemic period. We also explored testing infants online when the experimenter was present (synchronous) or not present (asynchronous) (see Table 1 for an overview of each experiment).

Experiment 1: Word recognition in an intermodal preferential looking task at 24 months

The IPL paradigm is widely used to probe lexical knowledge in the early years, as well as examine infants' sensitivity to various aspects of linguistic details in words (Golinkoff et al., 1987). Our aim was to guide the implementation of an online adaptation of the IPL to collect eye movement data using a participant's webcam in their home context. While this type of asynchronous collection of eye movement data in young children has already been explored using platforms such as Lookit (e.g., C. M. Nelson & Oakes, 2021; Scott & Schulz, 2017) and Labvanced (e.g., Bánki et al., 2022), to our knowledge, no published findings are using the Gorilla Experiment Builder platform (www.gorilla.sc – Anwyl-Irvine et al., 2020b). Most studies testing children using Gorilla have tested older children and collected accuracy and reaction time measures (e.g., Chere & Kirkham, 2021), with tasks requiring, for example, a button press response (e.g., Ross-Sheehy et al., 2021) rather than looking behaviour in infants. This might be because, while Gorilla Experiment Builder can run behavioural studies with the functionality to access a participant's webcam and record looking behaviour, this option is still in Beta. Thus, Experiment 1 tests how well the platform can accommodate an IPL task when testing infants.

Two key aspects of this adaptation were considered. The first was to understand how an online procedure may affect issues of timing in the experiment, due to factors such as internet speed and different device types. The second was to see how much usable data could be collected when children are tested in their home environment and when trials are presented automatically – that is, not infant-led as would be the case in many lab-settings.

A word recognition task was chosen because of its relatively reliable large effect size and replicability when conducted in a lab setting. In a meta-analysis of typically used methods in language development studies, Bergmann et al. (2018) found an average effect size of $d = 1.24$ ($SE = 0.26$) in online word recognition studies ($N = 6$). Thus, choosing this method offered the best chance of developing a proof of concept for an online IPL procedure for paradigms with potentially smaller effect sizes, such as a semantic priming study (e.g., $d = .32$, Jardak & Byers-Heinlein, 2019).

In a typical word recognition task, a participant is played an auditory stimulus which is the label of one of two simultaneously presented visual stimuli. In a lab setting, the participant typically fixates on the named visual stimulus for longer than the unnamed visual stimulus, which is taken as evidence of word recognition. Infants are able to fixate a target referent as young as 6-9 months (Bergelson & Swingley, 2012) in a look-while-listening procedure, with word comprehension and recognition generally observable by 12 months (Vihman et al., 2007). Therefore, by testing at the older age of 24 months we had an optimum chance of replicating the same effect in an online modality. If running the experiment in an online modality was significantly different to an in-lab modality, this might mask the effect of a longer proportion of looking time to the target image.

Method

Pilot study

Using the online experimental platform, Gorilla Experiment Builder (www.gorilla.sc – Anwyl-Irvine et al., 2020b), a small number of participants participated in a pilot study (adults: $N = 2$, infants $N = 4$). As previously mentioned, Gorilla Experiment Builder can access a participant's webcam and record, with their consent, but this feature is in Beta, and has its limitations. One of which is its inability to simultaneously record a participant and the experiment, or precisely what the participant sees on screen and when. While the timing of stimuli presentation and duration can be precisely programmed into the experiment on Gorilla Experiment Builder, when the experiment is run on a participant's device, some variability may exist because of the differences in devices used, internet browsers, and internet connection speeds, though timing accuracy does seem quite stable (Anwyl-Irvine et al., 2020a). Another potential variable aspect of the webcam recording feature is a delay in the command from Gorilla requesting access to a participant's webcam, and the point at which the recording starts. Although this can be up to 500ms according to one of the developers (personal communication, 23rd May, 2021), we found only marginal delays (10-20ms) through piloting. Additionally, a design feature was added to the experimental design (see below) to note which trials began recording before visual stimulus onset, and which did not.

Piloting the experiment on adults and infants was crucial to devise satisfactory solutions to these limitations and to decide how to best minimise variability in executing the experiment online. Email correspondence with parents and viewing the data that were

successfully generated allowed us to make a set of small changes to the paradigm. Differences between the pilot and test are described below in the Procedure section.

Participants

Participants were recruited through the University of Plymouth BabyLab database and Facebook page. Following recommendations for minimum sample sizes for infant studies that are based on a simulation study of the systematic effect of sample size on the results of infant studies ($N = 20\text{--}32$; Oakes, 2017), 20 monolingual British English-learning infants (13 boys, 7 girls) were tested. The target sample size was reached before analyses of the data. The mean age of participants was 24 months 3 days (range 23 months 3 days - 25 months 28 days). Participants were considered ineligible if they spoke more than one language, were born more than six weeks prematurely, or had a diagnosed language or developmental delay. No participants had to be excluded on these bases. For each of our three experiments, parental education was measured on a scale from 1 to 6 (1 = primary education - 6 = postgraduate degree) with the highest value taken from either parent (e.g., Mäkinen et al., 2006; Mossakowski, 2008).

Materials

A total of twenty-four target words (e.g., bed, key) were selected which were familiar, common, highly-imageable nouns known by at least 60% of English monolingual 18-month-olds according to the Oxford Communicative Development Inventory (Hamilton et al., 2000) and the UK CDI (UK-CDI Database, 2016) (see Table 2 for the list and exact percentages). All words were monosyllabic.

Auditory stimuli were recorded individually by a female adult with a neutral south-west British accent. The carrier word “Look!” was also recorded separately. Visual stimuli were colour photographs from the internet, cut out from their background and placed centrally on a light grey background to reduce brightness on the screen. Two versions of each image were created: one for presentation on the left of the screen, and one for the right. Animate objects were positioned to face the centre of the screen.

Target words were organised into word pairs in which there was no semantic or phonological overlap. The twelve pairs formed one block. In each pair, one word acted as the target and the other as a distractor. The distractor words then became the targets in the second block of trials, and these were paired with a different word that had acted as a target in the first block.

Procedure

Through piloting, the following modifications were made to the experimental design and procedure:

- Participants were restricted to using a laptop or computer. Those without such a device were deemed ineligible. This criterion was set to ensure visual stimulus presentation would be as large and as predictably positioned as possible. Gorilla Experiment Builder’s default positioning of two adjacent images is to space them as far apart, to each edge of a device’s screen as possible.

Table 2. Experiment 1. Percentage of 18-month-olds with knowledge of the stimuli words used in the online IPL task

| Target | % known at 18 months OCDI | % known at 18 months UKCDI |
|--------|---------------------------|----------------------------|
| bed | 85 | 97 |
| bird | 88 | 88 |
| book | 95 | 98 |
| bowl | 58 | 77 |
| box | 48 | 63 |
| bread | 72 | 77 |
| car | 95 | 97 |
| chair | 80 | 95 |
| cheese | 63 | 78 |
| cot | 70 | 68 |
| dog | 98 | 99 |
| duck | 90 | 86 |
| fish | 75 | 81 |
| foot | 70 | 92 |
| frog | 56 | 68 |
| hair | 91 | 86 |
| key | 74 | 81 |
| pig | 77 | 82 |
| plane | 81 | 72 |
| shoe | 99 | 97 |
| spoon | 77 | 76 |
| swing | 64 | 68 |
| train | 66 | 81 |
| tree | 69 | 78 |

- The experiment was programmed to only run on the web browser Google Chrome as there were some upload and display issues with other browsers.
- A calibration phase was added at the start of the experiment to ensure a participant's screen was not working in a 'flipped' mode, and to validate that, when an image was presented on the right only, the child looked to the right.
- A short beep of 100ms was added to coincide with the visual stimulus onset. In the absence of seeing when the pictures appeared on screen in a participant's webcam recording, the beep was a feature to enable the coder to have a reference point when manually coding eye movement offline. Each trial was checked for the presence of the beep during analysis, to ensure that the webcam recording started ahead of the images being presented on screen.

- Trials were divided into two blocks and separated using a short video to maintain attention. As the experiment could not be driven by the child's attention to the screen on every trial, the short video was a way of re-focusing the child if they had lost interest. Piloting showed inattention to be very infrequent.
- A duration of 500ms was added to each trial, resulting in the images remaining on screen for 5500ms (compared to 5000ms in a typical lab-based experiment). This was to compensate for any potential clipping towards the end of the recording.

All of our studies were approved by the University of Plymouth Faculty of Health Ethics Committee. Parents were invited to participate in the study through the Plymouth BabyLab database and through adverts posted to the BabyLab's social media accounts. When a parent expressed interest, further communication moved to email. A participant information sheet was issued and the technical requirements for the online study were reiterated through email communication. A day and time were agreed, on which to complete the study. On the appointed day, an email with instructions for the study was sent to the parent and a unique link to the experiment was activated on the Gorilla Experiment Builder website. By using a unique link, it meant participants could leave the experiment and return to it later, continuing where they left off. The reason behind establishing a day and time to do the online experiment was to ensure a researcher could be available for any questions or support required while participants did the task¹. Parents were instructed to begin the procedure without their child present, to minimise the time a child would need to stay engaged. It was made clear that the parent would be instructed when to prepare their child for the task.

When clicking on the Gorilla Experiment Builder weblink², an overview of the study was displayed, including the eligibility criteria for participation. The next screen was an eligibility questionnaire, to ensure participants were the right age; were not born more than six weeks prematurely; were exposed only to English; and did not have a language or developmental delay. At this point, a participant could be excluded in which case the parent would see an ineligibility screen and be asked to email the Plymouth BabyLab if they believed this to be incorrect, or if they wanted to find out about other studies running that their child might be eligible for.

If eligible, a participant had to consent to the study by completing an online questionnaire which detailed the procedure, the data collected and the right to withdraw. Demographic information was collected in a series of short online questionnaires before the experiment started³.

Following this, participants progressed to a technical eligibility check so they could test their sound and webcam before the experiment, and to grant Gorilla access to webcam recording. A Gorilla pop-up appeared in the web browser asking for consent to access the webcam, at which point a parent could refuse access if they did not agree to their data being accessed in this way. Furthermore, the recording test established the audio and video recording capabilities of a participant's device and it also allowed parents to

¹Parents did occasionally need technical support which often related to needing a new link to be sent. This mainly resulted from not reading the instructions, or pressing a button in error. We modified the email and experimental instructions to try to minimise this. In a couple of cases, parents' browsers blocked the Gorilla pop-up requesting permission to record via the webcam. Since we were online while the parent did the experiment, we were able to talk through various checks to resolve the issue.

²The full procedure can be viewed using this link: <https://app.gorilla.sc/openmaterials/627362>

³See <https://app.gorilla.sc/openmaterials/627362> for the exact questions asked.

autoplay the recording to fully understand the footage that would be recorded of their child when the experiment began. Throughout the procedure, an 'Exit' button was made available in the bottom left-hand corner of the screen in case a participant chose to withdraw from the study⁴. There was explicit mention in the instructional email that a participant should click on this 'Exit' button if they wanted to withdraw and to request, by email, the withdrawal of any data collected on their child up to that point if they desired, without any explanation for their decision.

The experimental procedure began by instructing the parents to place their child on their lap, with their device's webcam focused on their child's eyes. Detailed instructions were provided, using images, so that the parent could see how to prepare their child for the experiment. Rough measurements were provided (e.g., place the device at arm's length, mirroring what other researchers were trialing at the time for online experiments), and opportunities were available for the parent to perform test recordings before they began testing. Based on all this information, the parent deemed when the position of their child was satisfactory, and when they were ready to begin the task. Parents were instructed not to engage with their child when starting the experiment.⁵

The experiment was preceded by four calibration trials in which the word "Look" was followed by the word "biscuit" and an image of a biscuit appeared on the left-hand side of the screen. This process was repeated on the left side with the word "monkey" and a corresponding image. The two words were then repeated with the same images now appearing on the right-hand side of the screen. Neither of the words were used as targets, or distractors on critical trials. The calibration phase established a baseline for the participant's individual looking pattern and validated that the image was presented on the correct side and not in a 'flipped screen' mode.

The parent controlled the start of the word recognition task by clicking on a button. The experiment began with a 5000ms black and white attention-getting video showing simple geometric shapes accompanied by sound. Then, the automatic presentation of trials began and did not stop in their delivery until all trials had been presented, which lasted for about three minutes.

Each experimental trial began with a smiley fixation point in the centre of the screen for 1000ms to focus the child's attention to the middle of the screen. This was replaced by two visual stimuli, positioned on the left and right sides of the screen for 5500ms. In an equivalent lab-based study, a trial would last 5000ms but an additional 500ms was added in case of clipping at the end of the recording. The auditory stimuli began with a beep for 100ms to coincide with the visual stimulus onset, necessary for analysis. This preceded a silence and the carrier 'Look' before the target word onset at 2500ms. Each trial was thus divided into a 2500ms pre-naming and 2500ms post-naming window (see Figure 1).

After 12 trials, the same attention-getting video from the start was played to maintain the child's attention before a second block of 12 trials resumed. The video also separated the two blocks in which visual stimuli acted as targets in one block and distractor pictures in the other. The order of blocks was counterbalanced. The side of the target visual stimulus was counterbalanced. The experiment ended with the same 'reward' video that was played at the start and middle of the experiment.

⁴This was only clicked once by mistake, and not because the participant wanted to exit the experiment.

⁵This could not be controlled due to the remote nature of the testing, which meant that we could not stop the parent looking at the screen during the experiment. However, video recordings of parent and child indicated that the parent did not look at the screen continuously, and sometimes they did not look at all as the parent was sometimes absent.

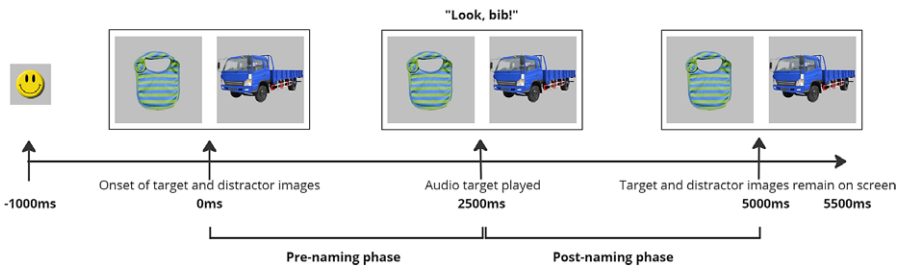


Figure 1. Experiment 1. Trial timeline. Onset of the auditory label of the target picture was always at 2500ms.

To complete the procedure, the parent marked a list of target words as known or unknown to the child, before a final debrief screen, inviting any questions or comments and a chance to mark whether any technical difficulties had been experienced during the tasks.

After completion of the full procedure, a participant's data were downloaded, and the calibration trials were checked to confirm audio and video recording was satisfactory. Questionnaires were reviewed to see if the participant had experienced any technical issues or if further information relating to their responses in the questionnaires was required. A final email was sent, requesting clarification pertaining to comments in the questionnaires (where necessary) and issuing a certificate and £5 Amazon voucher to acknowledge participation. The final email also included a short debrief of the study's aims and application and invited the participant to ask questions if necessary.

Results

Technical Specifications

Devices were restricted to laptops or computers, yet this can still mean a range of screen sizes. Gorilla records the device type used by a participant, including its screen size. The average viewpoint size on screens used was 1432x742 with parents classifying the mean quality of audio as 5 (Very clear, on a scale of 1 to 5. Range: 4-5). Most participants were using the latest operating systems for their devices, and the latest version of Chrome. The full range of technical specifications can be seen in [Table 3](#).

Data Processing and Analysis

Using a bespoke online encoder developed by the UoP School of Psychology technical team, videos of individual trials were uploaded and automatically split into 50ms frames. For each frame, the primary coder, blind to the visual and auditory stimuli presented, assessed the digital videos off-line frame by frame, manually marking the position of the participant's eye position as left, right, away, or indeterminate by using four corresponding keys on the keyboard. This information was saved in .csv format and later downloaded for analysis.

A second, skilled coder manually coded 10 per cent of the full dataset. Inter-rater reliability agreement between coders was 87% and according to Cohen's Kappa calculation, was moderately reliable $\kappa = 0.47$. On further inspection of the discrepancy between the two coders, out of the total 13% disagreement, 6% was specific to whether a gaze was

Table 3. Experiment 1. Overview of device types used in the online IPL study

| Participant | Participant OS | Participant Browser | Participant Monitor Size | Participant Viewport width | Participant Viewport height | Audio Quality (1-5) |
|-------------|----------------|----------------------|--------------------------|----------------------------|-----------------------------|---------------------|
| 0hl65s6s | Windows 10 | Chrome 87.0.4280.88 | 1536x864 | 1536 | 754 | Clear enough- 4 |
| 26o5cpdq | Mac OS 10.14.5 | Chrome 86.0.4240.193 | 1440x900 | 1440 | 821 | Very clear- 5 |
| 4k5u5xs8 | Windows 10 | Chrome 86.0.4240.183 | 1366x768 | 1349 | 625 | Very clear- 5 |
| 4uoxz04w | Windows 10 | Chrome 85.0.4183.121 | 1536x864 | 1438 | 704 | NA |
| ksry8lfl | Mac OS 10.13.6 | Chrome 86.0.4240.80 | 1680x1050 | 1680 | 971 | NA |
| reuryabw | Mac OS 10.14.0 | Chrome 87.0.4280.67 | 1440x900 | 1050 | 752 | Clear enough- 4 |
| 22vg0z4l | Windows 10 | Chrome 86.0.4240.75 | 1536x864 | 1519 | 722 | Clear enough- 4 |
| 5ym93g5p | Windows 10 | Chrome 86.0.4240.75 | 1366x768 | 1366 | 625 | Very clear- 5 |
| cpqtjso9 | Windows 10 | Chrome 86.0.4240.193 | 1366x768 | 1349 | 625 | Very clear- 5 |
| qiamsumn | Windows 10 | Chrome 67.0.3396.99 | 1366x768 | 1349 | 662 | Very clear- 5 |
| uibpbg89 | Windows 7 | Chrome 86.0.4240.111 | 1920x1080 | 1920 | 1009 | Very clear- 5 |
| ye42nool | Windows 10 | Chrome 87.0.4280.66 | 1280x800 | 1280 | 689 | Very clear- 5 |
| lfioaben | Windows 10 | Chrome 86.0.4240.198 | 1280x720 | 1280 | 610 | Very clear- 5 |
| odcoevc8 | Windows 10 | Chrome 86.0.4240.198 | 1920x1080 | 1920 | 969 | Very clear- 5 |
| plrudr83 | Windows 10 | Chrome 86.0.4240.183 | 1368x912 | 1368 | 783 | Very clear- 5 |
| pmwyldgf | Windows 10 | Chrome 86.0.4240.198 | 1680x1050 | 1680 | 939 | Clear enough- 4 |
| xtu5nbo8 | Windows 7 | Chrome 86.0.4240.193 | 1536x864 | 1198 | 630 | Clear enough- 4 |
| heojqujc | Windows 10 | Chrome 86.0.4240.111 | 1366x768 | 1349 | 657 | Very clear- 5 |
| iaahp11j | Mac OS 10.15.7 | Chrome 86.0.4240.193 | 1440x900 | 1200 | 667 | Very clear- 5 |
| s5xh3nt0 | Windows 10 | Chrome 86.0.4240.111 | 1366x768 | 1366 | 625 | Very clear- 5 |

Table 4. Experiment 1. Descriptive data of the whole sample

| <i>Means and Standard Deviations of the children's age and looking times</i> | | |
|--|----------|-----------|
| | <i>M</i> | <i>SD</i> |
| Age (days) | 735.05 | 19.04 |
| Boys' age | 728.58 | 23.31 |
| Girls' age | 737.50 | 19.06 |
| Parental education | 5.48 | 0.60 |
| Looking time pre-naming (PLT) | 0.50 | 0.07 |
| Looking time post-naming (PLT) | 0.62 | 0.07 |

Note. Parental education level is the highest of the two parents' highest educational levels, ranging from 1 to 6.

indeterminate or not, meaning the gaze was still on screen, but unclear where exactly. This might explain the lower-than-expected reliability measure.

Trials were excluded from analysis if a child did not fixate for a minimum of 750ms, somewhere on the screen (left, right or indeterminate) ($n = 0$), or if the child did not know the target word based on a parent's report of their child's word knowledge ($n = 21$ trials, or 4.38% of trials). The latter ensured that an infant was evaluated only on their understanding of known words.

The raw .csv files, generated by coding eye movements using the UoP Encoder, were uploaded in R Studio (v1.4.1717; R Core Team, 2021) for all further analyses⁶. The R tidyverse and dplyr packages (Wickham et al., 2019) were used.

Descriptive Statistics

Table 4 provides the descriptive data for the children's ages and looking times.

When aggregating all participants' looking time by condition, on average, participants spent 82% of the time looking at either the left or right side of the screen, with an additional 16% of the time looking at the screen but at an indeterminate point on the screen (i.e., neither clearly left nor right). This time also accounts for saccades between the left and right sides of the screen. Finally, 2% of looks per participant were looks away from the screen.

Out of a possible 480 trials (a maximum of 24 trials for each of the 20 participants), a total of 459 trials were included for analysis. Reasons for exclusion were entirely due to the target word not being known to the child (21 trials or 4.38% of trials), which was measured by parental report. No trials were excluded due to inattentiveness, measured as <750ms spent looking at the screen per trial. The average number of valid trials per participant was 22.95 ($SD = 1.4$). In summary, 24-month-old infants were very engaged in an online looking task when administered in their home. By way of comparison, in a meta-analysis looking at looking while listening studies, among other methods, Bergmann et al. (2018) used a linear mixed effects model to predict an exclusion rate of 30% of data for this task type, including minimum looking time criteria. In a more recent study, Byers-Heinlein et al. (2021) saw an exclusion rate of 5.07% for equipment failure, parental interference

⁶The analysis code is available on request.

and fussiness, in addition to 23.03% data loss due to infants not attending to objects during the specified window of analysis.

There was no effect of gender on response rate $t(18) = .44, p = .66$.

Proportion of Looking Time to the Target

A participant's looks were aggregated by condition (i.e., target, distractor, away, indeterminate) and the proportion of time spent looking at the target compared to the distractor was calculated for the pre- and post- naming windows.

The pre-naming window of analysis was set at 200ms – 2500ms which allows for an initial 200ms shift in eye gaze (Fernald et al., 1998, 2001) from an attention-getter to one of the pictures, followed by 2300ms of free-looking. The post-naming window was set at 2700ms – 5000ms to allow for initial processing of the onset of the audio, followed by the same amount of free-looking time (equivalent to 46 frames of 50ms per trial, per participant).

The proportion of looking time (PLT) towards the target visual stimulus, relative to the distractor stimulus, was calculated as the dependent variable for the pre-naming and post-naming windows, per trial:

$$\text{Looks to target} / (\text{Looks to target} + \text{Looks to distractor})$$

A two-tailed, paired t-test was run on the PLT in the pre-naming and post-naming windows of analyses. Overall, twenty-four-month-olds looked at the target longer in the post-naming window ($M = 0.62, SD = 0.07$) compared to the pre-naming window ($M = 0.50, SD = 0.07$) (see Figure 2, with the white square indicating the mean). The difference between looking behaviour in these two periods was significant with a very large effect size, $t(19) = 17.22, p < .0001, d = 1.61$. This indicates that participants looked longer at the target picture after it had been named, indexing word recognition.

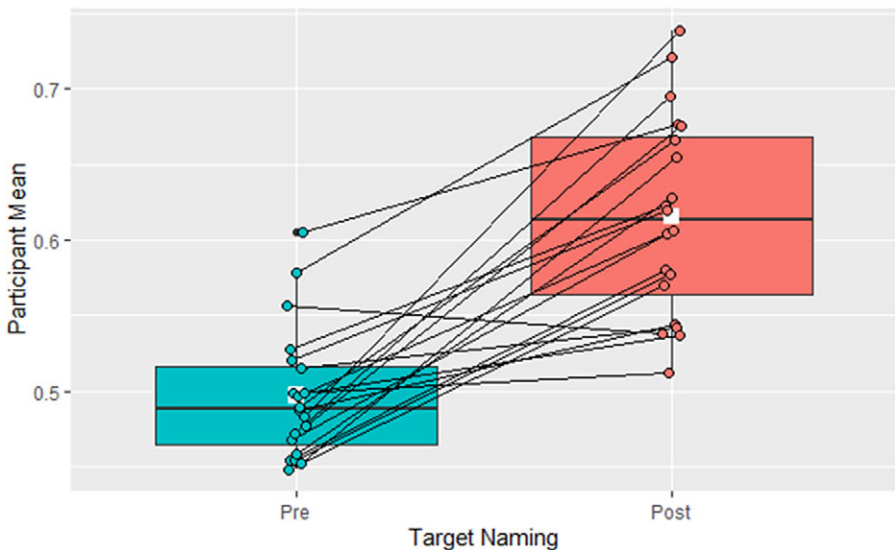


Figure 2. Experiment 1. Proportion of Looking Time Pre- and Post-naming during the online IPL study.

Discussion of Experiment 1

A simple word recognition experiment was run using the online experimental platform Gorilla Experiment Builder (www.gorilla.sc – Anwyl-Irvine et al., 2020b) as a proof of concept to test the feasibility of running online preferential-looking experiments with infants. The results from Experiment 1 indicate that with some modifications to lab-based procedures, an online version of an infant methodology can indeed be run successfully. Experiment 1 adapts the IPL task into an online modality, providing a validation of the general testing paradigm. As far as we are aware, this procedure is one of the first of its kind to be conducted on Gorilla Experiment Builder with young children. This is important as it contributes to the evidence base for testing young children online, using a different online platform than what is currently being used. We found Gorilla Experiment Builder to be a user-friendly platform, requiring no coding experience and that can support an IPL procedure. It is unclear if this platform has the potential to replicate in-lab findings when testing a procedure on infants with a smaller anticipated effect size, such as a semantic priming task, and this is something we have begun to test (Fitzpatrick, 2023) and continue to explore. The full set of materials of this experiment has been made open source for other researchers to use for replication studies.

The results clearly showed that infants aged 24 months looked at a picture on screen longer when the picture was named, compared to a picture that was unnamed. This is an expected outcome which indexes word recognition in children and replicates previous lab-based findings (e.g., Vihman et al., 2007). The novelty lies in the fact that the 24-month-olds were performing the task online, in their own homes and using their own devices. Participants were not overly distracted by their surroundings, nor were there significant issues with differing device types and internet speeds. Compared to lab-based studies, the effect size found in Experiment 1 ($d = 1.61$) is larger in magnitude (e.g., Bergmann et al., 2018, found an average effect size of $d = 1.24$ in a meta-analysis) which is a promising finding for other online studies collecting eye movement data.

Interestingly, participants remained engaged throughout the procedure despite the fact that trials were not infant-led – that is, they ran automatically without pause. This is a very different approach to many lab-based studies in which the start of every trial is initiated by the experimenter when the infant's attention is focused on the computer screen (e.g., Arias-Trejo & Plunkett, 2009; Chow et al., 2017; Floccia et al., 2020; Singh, 2013; Styles & Plunkett, 2009). Automatic presentation of trials was borne out of necessity while using Gorilla Experiment Builder to administer the task online. According to the findings of this study, running the experiment without pause does not seem to have had a negative impact on a child's ability to perform the task. This may be thanks to the features integrated into the design of the experiment such as fixation points and video rewards at the start, middle and end of the procedure.

Participants also remained engaged in the face of a twenty-four-trial experimental design, which is double the number of trials commonly used in infant studies at this age (Arias-Trejo & Plunkett, 2009, 2013; Jarak & Byers-Heinlein, 2019). This is encouraging support for future studies as using this number of trials will help with the power of future studies in the case of potential data loss occurring, as mentioned above (i.e., distraction, technical issues etc).

With regards to this particular study, there was very little attrition or data loss (<5%) compared to some lab-based studies, which can lose up to 30% according to a meta-analysis performed by Bergmann et al. (2018). This might be due to a participant feeling more relaxed in their home environment compared to a lab environment. By informally looking

at the experimental videos, children did not seek out contact as frequently with a parent by turning around, as they do in the lab. Similarly, the child might have felt more at ease on a parent's lap, rather than in an unfamiliar car seat/ booth in a lab. These hypotheses are supported by the data; there was a high proportion of looks on-screen to the left or right (82%) versus off-screen (2%). This amount is likely to be larger considering looks on-screen but to an indeterminate location (16%) may have been looks to the left or right. One explanation might be the manual coding of eye movement which minimised data loss, compared to lab-based studies in which the eye-tracker losing signal leads to data loss.

Taken together, these findings provide encouraging support that other infant paradigms might be suited for adaptation to online testing. Findings from this study indicate that infants can complete twice as many trials as other, comparable word recognition studies specify, while still maintaining attention. Using an increased number of trials will help increase power for testing such hypotheses.

Experiment 2: Word learning in a Switch task at 17 months

Infants can learn word-object associations that can be robustly measured at 12 months (Curtin & Zamuner, 2014). Waxman and Booth's (2001) findings suggested that infants of 14 months can identify novel noun words (e.g., "This one is a *blicket*") and specifically map them to new objects (e.g., carrot, orange). Stager and Werker (1997) developed the Switch task to investigate how infants behave in a situation that requires them to link a new label with a new object. In the Switch task, infants are exposed to a novel word-object pairing where they see a novel object moving back and forth across the screen, while simultaneously hearing a novel word repeatedly. This presentation continues until a predetermined decline in looking is observed in infants. In the following test phase, infants are tested with two types of trials. On the "same" trial, the initial object-word pair stays the same while on the "switch trial", the object is paired with a different word. If infants notice the difference, they should look longer on the "switch" than on the "same" trials (Fennell & Waxman, 2010; Stager & Werker, 1997). A recent meta-analysis has found a low to moderate effect size of Cohen's $d = 0.32$ (141 Switch tasks in infants aged 12 to 20 months; Tsui et al., 2019). Previous research revealed that infants of 14 months learned to associate two distinct sounding words (*lif* and *neem*) to two different objects by looking longer to the "switch" trial. However, infants aged 8 and 12 months fail to associate the different soundings (Werker et al., 1998). We decided to test 17-month-olds following Werker et al.'s (2002) demonstration that infants at this age could apply phonetic detail when learning new words within a short exposure period. We reasoned that it would give us better chances to observe a large effect and an increased power of word learning with phonetically dissimilar words when testing online, especially given that at 17 months, infants are experiencing a boost in vocabulary learning (e.g., Cochet et al., 2011).

Experiment 2 describes an online adaptation of the Switch task with 17-month-olds, using a combination of Zoom and offline coding. The infants were tested using a modified habituation paradigm similar to the design used by Werker et al. (1998) but with only one word-object pairing and not two, as in Fennell and Waxman (2010) and with a different habituation criterion. Specifically, we did not measure a habituation, that would be indexed by a pre-specified decrease in looking times, but we fixed a familiarisation time identical for all participants (see the procedure for more details). The sample size target was 16 participants as in previous Switch tasks experiments (Fennell & Waxman, 2010; Fennell & Werker, 2003). It must be noted that the data reported in Experiment 2 were collected before we read about the study by Bulgarelli and Bergelson (2022) who also

conducted a one-word Switch task but with 18 younger children (7-9 months), and we will address their findings as compared to ours in the Discussion.

Method

Participants

A total of 19 parents with monolingual children (10 boys and 9 girls) aged 17 months, ranging 16 months 4 days to 18 months 10 days, were recruited from the Plymouth Babylab database (with the same inclusion criteria). They were all residents of Plymouth and its surroundings and had signed up to the Babylab to take part in any proposed studies.

Stimuli

The audio stimuli were two nonsense consonants–vowel labels: *neem* and *lef* recorded in infant-directed speech (IDS). IDS is efficient in capturing and keeping the attention of infants (Fernald, 1985). These stimuli highly differ in articulation and a highly dissimilar nonsense consonant–vowel noun, *pok*, was used during the pre- and post-test trials.

An English-speaking female from the South West of England produced several tokens of each syllable in a rise-fall intonation phase, in an infant-directed speech (Fennell & Werker, 2003; Stager & Werker, 1997). Following Fennell and Werker (2003), the final stimuli contained 10 exemplars, each lasting approximately 0.7 sec, including a 1.5-sec silent interval between each exemplar, resulting in audio files of 22.5 sec in duration.

The stimuli were shown as 3D moving objects to highly attract and maintain infants' attention (Baldwin, 1989; L. B. Cohen, 1973; Fennell, 2012). A trophy topper was used for both the pre- and post-tests (see Figure 3a) and a marker toy windmill object was used for the habituation and test trials (see Figure 3b). During the trials, the two objects spun, moved back and forth. The video clips were edited via the Photos laptop Windows application. The Switch task was administered online with the Zoom app using computer/laptop devices and it was recorded through the Zoom app for coding purposes.

Material

Zoom was chosen as the platform of testing for this experiment because unlike many other virtual technologies, it includes advantages that can be used for research purposes. Indeed, according to Archibald et al. (2019), Zoom's capacity to safely record and store

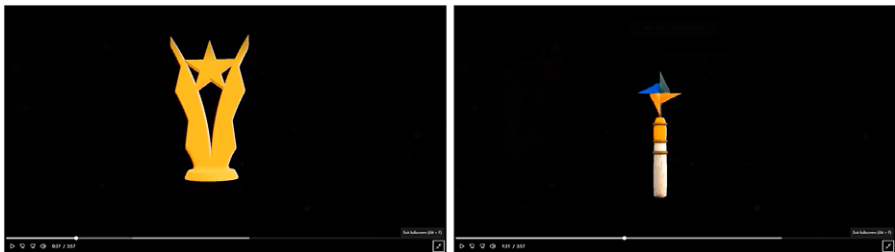


Figure 3. Experiment 2. (a) Trophy topper and (b) Marker toy windmill new objects.

sessions without any other third-party software is one of the advantages of protecting sensitive research data. Also, they reported that the capacity to back-up recordings to online server networks such as “the cloud” or local drives, is an additional security benefit as it allows for recordings to be shared safely for teamwork and real-time encryption (Zoom Video Communications Inc., 2016).

Pilot

Using Zoom, a small number of participants participated in a pilot study ($N = 9$). Piloting the experiment was essential to test the quality of the stimuli (video and sound), the data collection and to check how many habituation trials were needed for this online Switch task. A laptop Lenovo ThinkPad and the Photos app in Windows 10 were used to create, edit the video stimuli, and conduct the experiment. Participants were asked to operate with a computer or laptop to ensure satisfactory visual stimulus presentation.

A limitation of testing online with Zoom was the impossibility of controlling the pace of the trial presentation due to our specific set-up using the auto-advance feature of the Photo app. Therefore, all trials were presented at once without being able to control when to present the next one (as in Experiment 1 using Gorilla). Piloting showed that looking times noticeably declined after 4 trials, therefore the habituation phase was set at 4 trials for further data collection.

Procedure

The parent was sent via email the consent and information forms. At the same time, the parent completed a Communicative Development Inventory (short form of the Oxford CDI, Hamilton et al., 2000). The Oxford CDI is a list of words that are typical in children’s vocabularies. Parents were asked to tick whether their child could understand and/or say the words on the list. Then, the parent and child were invited to participate in the online Switch task.

Contrary to Experiment 1 where the researcher was not in the same virtual space as the child, here the researcher, the parent and the child were connected on Zoom together. The researcher was sitting in front of a laptop, while the child was sitting at home in front of the family electronic device. The session was video recorded. The child was asked to look at the computer’s screen and the parent sat in a chair next to his/her child. It should be noted that the researcher was not visible to the infant during the testing task.

When the child was attentive, the researcher started a 3min30s video to the participant consisting of 8 trials from the Switch task including a short clip of 30s (a talking bunny chasing a flying kite) to test if the participant’s devices’ sound and camera were correctly working. The infants were tested using a modified habituation paradigm, similar to the structure used by Werker et al. (1998) but with only one word-object pairing and not two, as in Fennell and Waxman (2010). Also, it was modified for the trial duration (increased from 14 sec to 22.5 sec). Each trial started with a flashing red light to get the infant’s attention on the screen. On the first trial, infants were presented with a pre-test stimulus: the label *pok* paired with the trophy topper. This pre-test stimulus was re-presented at the end of the experiment, during the post-test phase, and acted as a control of infants’ attention. During the following habituation phase, the infant was shown one word-object pairing (word *neem* and object *toy windmill*). After exactly 4 trials, the habituation phase ended, and was followed by the test phase. One test trial was the “same” trial, in which the

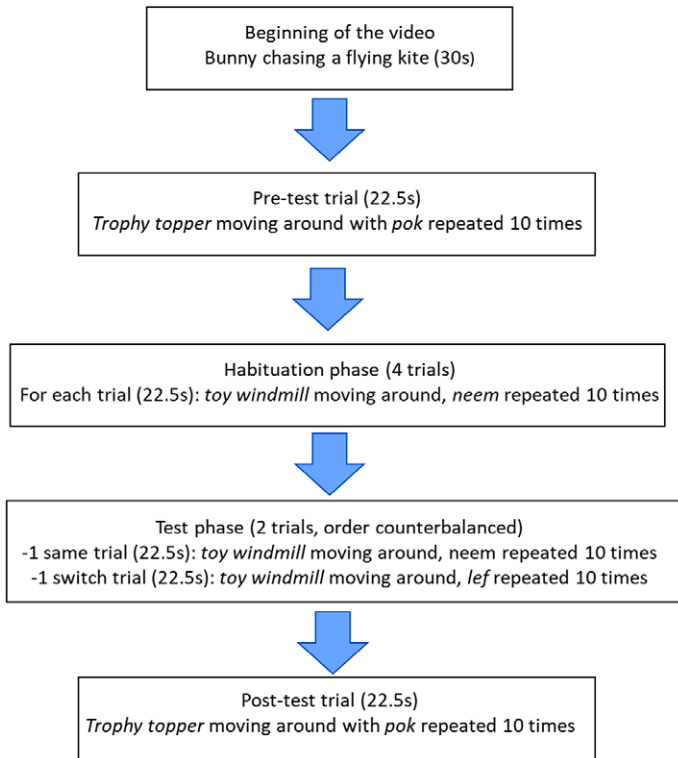


Figure 4. Experiment 2. Diagram of the online Switch task.

word-object pair presented during the habituation phase was shown again to the infant. The other test trial, called the “switch” trial, contained the familiar *toy windmill* object but was paired with a novel word *lef*. The order of presentation of the trials was counterbalanced across participants. If infants had learned the pairing, it was expected that they would notice the switch and look longer during the “switch” trial than during the “same” trial (Fennell & Werker, 2003). In the final post-test trial, the child was presented again with the word *pok* and the *trophy topper*. It was expected that if infants remained engaged throughout the experiment, the looking time during this last trial would be similar to the looking time during the pre-test trial (Fennell & Werker, 2003) (see diagram in Figure 4 for more details).

Coding

Using a frame-by-frame analysis (1 frame = 50 ms), coders scored infants’ looking times. To ensure the reliability of the main experimenter’s coding, a second trained coder scored the looking times of 25% of the participants. Inter-rater reliability agreement between coders was 81.78% and according to a Cohen’s Kappa calculation, was strongly reliable, $\kappa = 0.86$. 18.22% of disagreement between the two coders was due to whether the gaze of the child was still on screen or away, but that was equally distributed across the Switch and

Same trials, which means that it wouldn't have had an impact on the direction of the results.

Results

Table 5 provides the descriptive data for the children's ages, gender (10 boys and 9 girls), parental education, income deprivation scores, CDI scores and looking times.

To ensure that infants did not lose interest throughout the experiment, a paired sample t-test was conducted to compare looking time on the pre-test versus post-test trial. Contrary to what was expected (Fennell, 2012; Werker et al., 2002), children were significantly more engaged at the beginning of the task during the pre-test ($M = 19.67$, $SD = 5.39$) than during the post-test ($M = 14.95$, $SD = 7.80$, $t(18) = 3.85$ $p = .001$).

The main set of analyses addressed infants' performance on the test trials. A paired sample t-test revealed a significant main effect for test trials, with the children looking longer to the switch trial ($M = 17.89$, $SD = 5.52$) than the same trial ($M = 13.37$, $SD = 7.96$), $t(18) = -2.31$, $p = 0.03$, Cohen's $d = 0.53$. There was no main effect of gender and age on looking times. Thus, the 17-month-old infants exposed to the first pairing of word-object did notice the switch in label (see Figure 5).

A Pearson correlation was conducted between vocabulary knowledge as assessed by the CDI (see Table 2 for vocabulary statistics) and the performance on the Switch task as indexed by the "switch" versus "same" difference score in order to determine whether vocabulary size is related to children's Switch task performance (Werker et al., 2002). The correlation was not significant for comprehensive, $r(17) = -.54$, $p = .816$, nor for production scores, $r(17) = -.34$, $p = .883$. Age and gender did not have a significant effect on children's performance either.

Table 5. Experiment 2. Descriptive data of the whole sample

| <i>Means and Standard Deviations of the children's age, gender, CDIs scores and looking times during the different phases of the Switch task trials.</i> | | |
|--|----------|-----------|
| | <i>M</i> | <i>SD</i> |
| Age (days) | 517.21 | 18.74 |
| Boys' age | 514.46 | 18.14 |
| Girls' age | 525 | 18.73 |
| Parental education | 4.48 | 0.93 |
| IDS | 0.15 | 0.2 |
| CDI knows (percentile) | 38.52 | 18.78 |
| CDI says (percentile) | 9.52 | 10.2 |
| Looking time pretest (s) | 19.67 | 5.39 |
| Looking time posttest (s) | 14.95 | 7.8 |
| Looking time same trial (s) | 13.37 | 7.96 |
| Looking time switch trial (s) | 17.89 | 5.52 |
| Difference score (s) | 4.53 | 8.6 |

Note. Difference score is the difference between the looking time to same and switch trials.

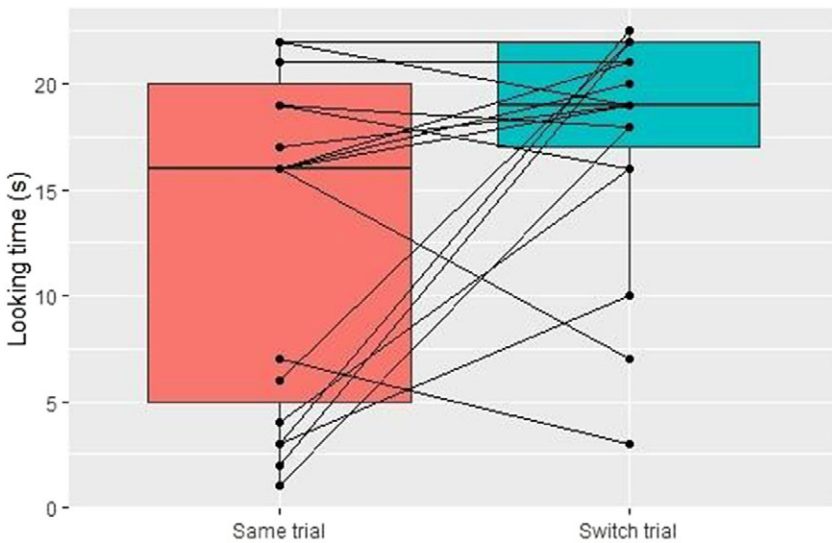


Figure 5. Experiment 2. Mean looking times to same and switch trials for each child.

Discussion of Experiment 2

In this second experiment, 17-month-old infants successfully learned the association between a new object and a new label, as indexed by their longer looking time in switch trials as compared to same trials. Thus, they were able to encode phonetic detail when learning a new word, which is consistent with previous in-lab findings (e.g., Stager & Werker, 1997; Yoshida et al., 2009). Our results are also consistent with Bulgarelli and Bergelson's (2022) which showed that younger infants successfully performed the one-word Switch task on Zoom.

No significant relation was found between vocabulary size and performance on the minimal-pair word-learning task, which is not in line with Werker et al. (2002). They found that at 14 months, both comprehensive and productive vocabulary size correlated with performance on the Switch task, and at 17 months, the correlation was found for comprehension only. However, they did not find an association between vocabulary size and performance success on the Switch task at the age of 20 months. It must be pointed out that many previous studies did not find a consistent relation between vocabulary knowledge as assessed by the CDI and word recognition (Hamilton et al., 2000; Swingley & Aslin, 2000). According to Werker et al. (2002), this would imply that vocabulary knowledge is only predictive of the phonetic detail when children are first building their vocabulary. After the vocabulary reaches some critical threshold, as measured by either comprehension or production, the relation is no longer consistent.

Another unexpected finding is that we did not find a renewed interest in the post-test phase as compared to the pre-test, suggesting that children's interest in the task decreased as the trials went on. It should be noted that Bulgarelli and Bergelson's (2022) Switch task did not have a post-test phase and therefore cannot speak to whether attention recovery was comparable to offline testing. One first reason for our finding is that we used a fixed familiarisation phase, due to technical limitations, contrary to previous researchers who applied a sliding habituation criterion (e.g., Bulgarelli & Bergelson, 2022; Fennell &

Waxman, 2010; Werker et al., 2002). Therefore, some of our participants might have lost interest by the time the test phase ended. Maintaining children's interest and engagement for a prolonged period of time can be a limitation of online methods, at least for the Switch task. Another reason might be that our selection of new objects might have been less interesting than, for example, the objects used by Fennell (2012). Also, the effect size obtained in our study (Cohen's $d = 0.53$), which is smaller than the effect size of 1.04 by Bulgarelli and Bergelson (2022), was noticeably higher than the average effect size of 0.32 computed in the meta-analysis by Tsui et al. (2019), which might potentially suggest a robust online replication of the main finding in the Switch task – that is, that children react to a change of word-object pairing. It should be noted that this interpretation cannot be certain as we cannot know whether our results reveal the robustness of the effect or an over-estimate of the effect size.

Experiment 3: Language assessment task in 19 to 26 months

Developmental language research typically involves the estimation of children's language knowledge, which tends to rely on parental questionnaires like the MacArthur CDI (Fenson et al., 2006). However, there are situations where a face-to-face assessment is needed, to complement or replace a parental questionnaire. In this experiment, a comparison between a parental report of the child's vocabulary knowledge and a vocabulary test directly administered to the child was explored (regardless of the setting). But most importantly, we also asked whether administering a test online would provide equivalent data to running it face-to-face. Most available language tests have been standardised with face-to-face data, with clinical evaluation requiring a face-to-face assessment of a child's language skills. It was an open question as to whether similar scores could be obtained for an online and a face-to-face version of the same standardised test. This is a pragmatic question: could early years professionals, practitioners and researchers trust data obtained in a virtual space? In our third experiment, we collected data with a standardised language assessment test, the WinG test (Cattani et al., 2019) to estimate toddlers' vocabulary knowledge, either online or in the Babylab. It was expected that children's performance on the WinG test would be affected by the environment the test is administered in (home vs Babylab). Our initial hypothesis was that face-to-face children would outperform online children on the WinG test, because it would be more difficult to maintain their attention remotely, and because sound and picture quality might get in the way of a clear communication.

Parents were also asked to fill in the Oxford CDI, which they would do similarly in their own time, whether the session would take place online or in the lab, and therefore the setting (online or face-to-face) was not expected to affect the CDI scores. Additionally, we analysed whether our WinG scores collected were positively correlated with the CDIs scores. Indeed, when the external validity of the WinG was assessed, a subsample of children performed one or more other language assessments including the Oxford CDI. The receptive score of the CDI was significantly positively correlated with the WinG comprehension subtests (noun ($n = 116$) and predicate ($n = 104$) separately). Similarly, the expressive score of the CDI was significantly positively correlated with the production subtests (noun and predicate separately) of the WinG (WinG manual: Cattani Krott et al., 2019).

A sample size of 60 participants was chosen for a study described in another manuscript, which examined the relationship between children's vocabulary knowledge

and parental screen time. Before reaching the sample size target, we did not analyse and compare the results of the online and face-to-face groups.

Method

Participants

Seventy children were tested and the data from 8 children were excluded due to the non-full completion of the WinG test (4 online and 4 face-to-face participants). The final sample included sixty-two healthy monolingual infants (31 boys and 31 girls) aged 19 to 26 months who were recruited from the Plymouth Babylab database with the same inclusion criteria as before. Thirty-two participated in the experiment online due to Covid restrictions at the time of testing and thirty were invited to do it face-to-face in the Babylab when restrictions were lifted. Participants were recruited the same way but were not randomly assigned to participate in the experiment remotely or face-to-face as a result of the COVID-19 pandemic. Forty-six parents completed the CDI (16 from the online group and 30 from the face-to-face group).

Materials

After completing a consent form, parents first filled in a demographic questionnaire to collect information about the family's socioeconomic status (SES). Then, they were invited for their child to do a language test, the WinG test (Cattani et al., 2019), either online, or in the Babylab. At the same time, they were asked to complete the Oxford CDI, prior to the WinG test or during the visit to the Babylab. Parents were also involved in another task related to their usage of screens, with data reported elsewhere (Nguyen, 2024).

For the video chat condition, the WinG was administered online with the Zoom app using computer/laptop devices. The test consists of 44 groups of 3 cards, 4 pre-tests and 40 experimental. Each set of 3 cards contains a comprehension card, a production card and a distractor card. The comprehension task contains 20 noun words and 20 predicate words, the production task also contains 20 noun words and 20 predicate cards. For each of the four components, a standardised score and percentile can be calculated for the number of correct answers that should be reached for each age and each gender. Following the WinG recommendations, only the comprehension tasks for both the noun and predicate were administered with children aged from 19 to 24 months old, whereas for children aged 24 to 26 months, the production task for the noun score was additionally given. The WinG scoring sheet was used to code the child's answers, as included in the WinG manual. For the video chat condition, the WinG test was recorded through the Zoom app, and children's responses were transcribed later. For the face-to-face experiment, the WinG test was recorded on a Canon video camera and responses coded afterwards.

Procedure

The parent was sent via email the consent and information forms. Then, the parent and child were invited to participate in the WinG game test.

For the Zoom session, the WinG cards were set standing against a cardboard box on a table, so that the cards would be visible through the child's screen. The researcher was

sitting in a chair behind the table and a laptop was placed in front of the table, facing the picture cards. The child was in a room at home and sat in front of the electronic device using Zoom and the parent sat in a chair next to their child.

For the face-to-face language test condition, the parent and child were invited to enter the Babylab, in which the WinG cards were set upon a table, with two chairs adjacent to each other on the table (for the child and the experimenter). The parent was sitting beside their child. The camera recorded the session to code offline the responses from the WinG test on the scoring sheet.

The WinG test was administered in line with the instructions from the WinG manual (Cattani et al., 2019), where children were invited to pick up or touch the card corresponding to a target word (in the comprehension task). However, for the WinG test online with Zoom, children could not touch or take the cards. Instead, they were asked to point to their computer's screen at the correct card. The session was video recorded, and the child's answers were scored offline according to their hand gesture and/or eye gaze going to the right, middle or left card. The WinG test started with 2 pre-tests of 3 cards each to give the child practice of what is required for the game. The 3 cards were presented in a random order in a line in front of the child, one comprehension, and 2 distractor cards. The children were first asked to point out or touch the named comprehension card; once they pointed to one of the cards, it did not matter if it was the right one. Then, the comprehension and distractor cards were taken away to move on to the next set of cards (see diagram in Figure 6). This was repeated for the next set of pre-test cards, all 20 experimental noun cards, the 2 sets of pre-test cards for the predicate condition and all 20 experimental predicate cards. Praise was always regularly provided, irrespective of the child's answers.

The WinG test was performed to the best of the child's ability, lasting around 30 minutes. Not all children can stay focused during the entire length of the testing session. Following the WinG manual, when a child began to show signs of boredom or

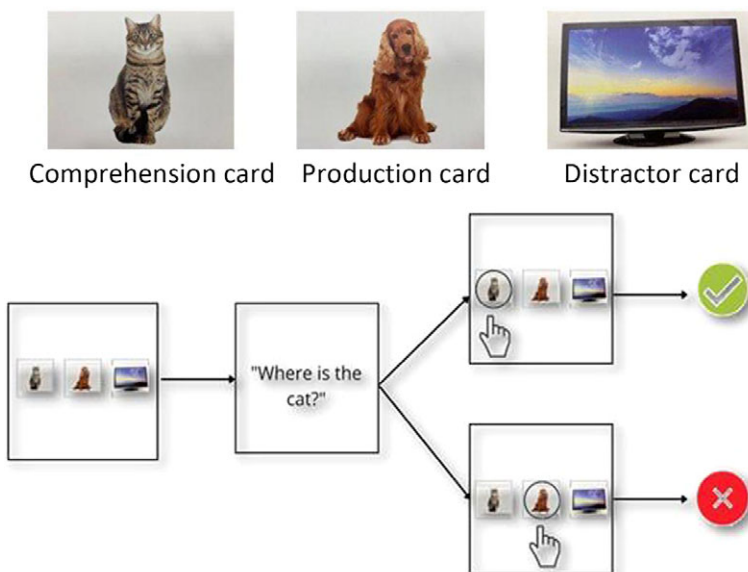


Figure 6. Experiment 3. Diagram of the structure of the WinG.

restlessness, they were offered a short break (e.g., getting a snack or drink). When the child was ready to resume testing, the administrator restarted from the last set of pictures before the break. If necessary, the test was stopped and was resumed another day within one week. The data collected for this study were the children's percentile score for noun comprehension and predicate comprehension as calculated by the standardised scores in the WinG manual. Moreover, the two parents' highest educational levels were used as the SES. The parent's postcode was collected with the demographic questionnaire and was used as a proxy for income, leading to the income deprivation score (IDS). The IDS were obtained from a government website (Ministry of Housing, Communities, and Local Government, 2019). The scores hold significance and correspond to the percentage of the relevant population experiencing that type of deprivation in that area. So, for instance, if an area receives a score of 0.27, it indicates that 27 percent of the population in that area is experiencing income deprivation. The larger the score, the more deprived the area.

It should be noted that the production task data were not reported here because none of the online children did the production task of the WinG test (as they were less than 24 months old and the WinG production part can only be tested on children older than 24 months), so we could have not compared production scores between the online and face-to-face participants.

Results

Table 6 provides the descriptive data for the children's ages, gender (31 boys and 31 girls), parental education, income deprivation scores, CDI scores, and WinG scores.

There was an absence of correlation between parental education and the income deprivation score ($r = -.062$, $N = 77$, $p = .63$). Therefore, only parental education was kept as the SES indicator as it is usually the best predictor of children development (Davis-Kean et al., 2021; Duncan & Magnuson, 2003).

First, participants from the two groups were compared on demographic measures. The online group included 16 boys and 16 girls, and the in-person group had 14 boys and 16 girls. Online participants had similar educational levels ($M = 4.94$, $SD = 0.70$) to the

Table 6. Experiment 3. Descriptive data of the sample

| <i>Means and Standard Deviations of the children's age, gender, CDIs scores and WinG scores.</i> | | |
|--|----------|-----------|
| | <i>M</i> | <i>SD</i> |
| Age (days) | 666.89 | 63.48 |
| Boys' age | 666.32 | 60.64 |
| Girls' age | 667.45 | 67.2 |
| Parental education | 4.82 | 0.82 |
| IDS | 0.1 | 0.06 |
| CDI knows (percentile) | 69.93 | 17.43 |
| CDI says (percentile) | 34.52 | 28.26 |
| WinG nouns (percentile) | 37.02 | 23.67 |
| WinG predicates (percentile) | 40.08 | 21.17 |

in-person parents ($M = 4.70$, $SD = 0.91$; $t(60) = -1.15$, $p = .25$). Children from the online group were about a month younger ($M = 649.22$, $SD = 54.14$) than those in the in-person group ($M = 685.73$, $t(60) = -2.35$, $p = 0.02$).

Then correlations were made between the CDI scores and the WinG scores. No associations were found between the CDI comprehension scores and the WinG comprehension (neither on nouns nor on predicates) scores. Our sample might have not been large enough to detect a relation between the CDI and WinG scores.

Next, independent t-tests were conducted to compare the online and face-to-face children's WinG performances. The results were adjusted by the Bonferroni correction (Abdi, 2007) as the nouns and predicates are both measures of comprehension. Thus, the significance value was divided by 2 and adjusted to 0.025.

It should be noted that standardised WinG scores incorporate age and gender. Online children performed significantly better on the WinG test noun comprehension ($M = 45.47$, $SD = 22.05$) than face-to-face children ($M = 28.00$, $SD = 22.27$); $t(60) = 3.10$, $p = .003$. Similarly, online children did better at the WinG test predicate comprehension ($M = 45.94$, $SD = 21.08$) than the in-person group ($M = 33.83$, $SD = 19.73$), $t(60) = 2.34$, $p = .023$. Figure 7 illustrates the comparison of the online and in-person performances on the noun comprehension.

On CDI comprehension, a regression model forcing age, parents' education, gender and the type of performance (online/in-person) led to a significant model ($R^2 = .32$, $F(4,41) = 4.83$, $p = .003$) with only age as a significant contributor ($\beta = .15$, $t = 4.09$, $p < .001$). On CDI production, the same regression model led to a significant model ($R^2 = .37$, $F(4,41) = 5.90$, $p = .001$) with only age ($\beta = .20$, $t = 3.44$, $p = .001$) and gender ($\beta = 17.52$, $t = 2.50$, $p = .016$) as significant predictors.

Independent t-tests were conducted to compare the online and face-to-face groups on their CDI comprehension and production scores. No corrections for multiple comparisons were made as comprehension and production vocabulary scores are different measures of language. Results indicated that there were no significant differences on the CDI comprehension between online children ($M = 68.06$, $SD = 16.57$) and in-person children ($M = 71.03$, $SD = 18.11$); $t(44) = 1.06$, $p = .57$. Also, there were no significant differences on the CDI production between online participants ($M = 26$, $SD = 20.56$) and face-to-face participants ($M = 39.52$, $SD = 31.18$); $t(44) = -1.60$, $p = .12$. Children who were tested online did not have significantly higher scores on the CDI. It supports the finding that online participants outperformed those who did the language test at the Babylab, but

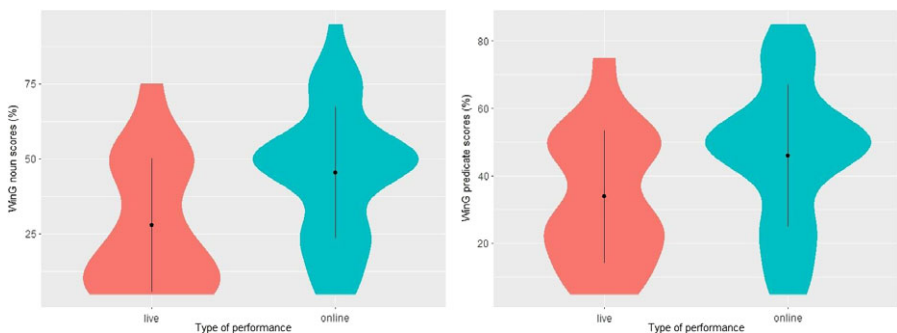


Figure 7. Experiment 3. Comparison of WinG comprehension scores between online participants ($N = 32$) and face-to-face participants ($N = 30$).

only for the WinG test. However, having no significant differences on the CDI scores between the online and face-to-face groups does not establish that there is no difference between the two groups, which we will discuss further in the discussion.

Discussion of Experiment 3

In this last experiment, we investigated the reliability of using a language assessment test, the WinG, online as compared to face to face. We originally expected that the children who did the WinG in the Babylab would outperform the online participants. Indeed, online children were not able to touch or take the cards which could diminish their engagement. In addition, they might have not seen the pictures and heard the words as clearly as in face-to-face interaction. However, the findings are exactly opposite to this hypothesis as online children outperformed the in-person group. Critically the two groups did not differ on CDI scores. A possible explanation for those results is that online children might have been more focused on the task because, first, they were in a familiar environment at home, and second, looking at a computer's screen might be more unusual and compelling. This is in line with what was found in the two previous experiments, where high effect sizes and low attrition rates were observed when testing online. Those results are in line with P. M. Nelson et al. (2021) who compared children's performances between face-to-face and online tasks. They tested children aged 4 to 5 years old on various tasks related to working memory – for example, visual spatial, and numeral competences. On five tasks out of eight, findings did not reveal differences across the two formats that they administered, but on the three other tasks (two related to verbal comprehension and one related to fluid reasoning), online children were found to outperform face-to-face ones.

There could be other explanations for our findings. Participants were recruited the same way and have similar SES but were not randomly assigned to participate in the experiment remotely or face-to-face as a result of the COVID-19 pandemic. Differences in composition of the sample can be one of the reasons why the online and offline group results differ on WinG scores. However, we did not find a difference in CDI scores, so our results are unlikely to be due to the online group having better language skills than the offline group. It might be more likely that participants recruited during the lockdowns might have performed better on the language test because they might have been more at ease with computers due to parents engaging more with them this way at home.

Our findings demonstrate that online data collection might be a feasible option for children's language assessment; however, it also means that norms may not be useful when testing online. Note that our data do not allow us to conclude firmly in this direction: the face-to-face group scored around 30 on the WinG test, and the online group around 50. As expected of standardised scores, 50 is what would be expected from a representative group similar to the population from which standardised scores were derived. It is possible that our face-to-face participants scored particularly low.

One possible explanation for children scoring low in the face-to-face condition was that all adults wore an opaque mask in the aftermath of lockdown. However, it was found by Singh et al. (2021) that opaque masks do not prevent children from recognising spoken words. In their study, 2-year-old toddlers were asked to identify eighteen familiar spoken words (e.g., "Can you see the spoon?") under three distinct conditions: words spoken without any mask, words transmitted through a transparent mask, and words conveyed through an opaque mask. The results indicated that the toddlers could identify familiar

words when presented without a mask or through opaque masks. However, they had difficulties to recognise words when heard through clear masks. Moreover in our study, we ensured at the start of the testing that children understood what was being said with the pre-test cards.

Another possibility would be about the timing of the visits and the general effect of the lockdowns. The online children were tested about 6 months after the start of the initial lockdown in the UK, which means that they had experienced their first years of life in a normal environment. In contrast, the face-to-face group was seen after having experienced lockdowns for at least one year, which represents a proportionally longer time of non-typical experience. Although their language skills might have been comparable (as demonstrated by the CDI scores in the two groups), their level of engagement might have been different enough to explain their lower scores on the WinG test. Davies et al. (2021) showed that children aged 8 to 36 months who did not attend childhood education and care (ECEC) during lockdowns had lower cognitive executive functions skills (cognitive flexibility, inhibitory control and working memory) than those who did.

The important conclusion of our findings here is that children tested online, and who were drawn from the same population as those tested face to face, outperformed the latter. It would have been interesting to replicate these findings, but the data collection opportunity was unique and unrepeatable due to the exceptional lockdowns' circumstances.

General discussion

We adapted three paradigms into online experiments to investigate various ways to estimate looking behaviour in young children. The results from the three experiments provide support for online testing reliability. With some modifications to lab-based procedures, the IPL and Switch tasks successfully collected eye movement data and provided solid replications of established results. In Experiment 1, previous lab-based findings were replicated (e.g., Vihman et al., 2007) and showed word recognition in children. In Experiment 2, infants significantly learned a new word which is consistent with previous in-lab (e.g., Yoshida et al., 2009) and online research (Bulgarelli & Bergelson, 2022) involving the Switch task. Finally, Experiment 3 demonstrated that children can perform well on a language assessment test administered online and that they were strongly engaged and responsive to the task.

The three experiments presented here have highlighted a number of advantages to testing in an online paradigm. Firstly, there can be high levels of engagement for young participants when tested in the home environment (Experiments 1 and 3). Indeed, we found that instead of being distracted by their surroundings, children remained engaged for the duration of the experiment which might be due to children feeling more comfortable and at ease in their home, according to Tsuji et al. (2022). A higher level of engagement in online experiments might also explain why children performed better in our Experiment 3 and in other previous studies (e.g., Bacon et al., 2021; P. M. Nelson et al., 2021).

Another advantage to testing online was the higher than expected effect sizes (Experiments 1 and 2). Comparing online to in-lab testing we found that effect sizes were not only replicated, but were much higher in magnitude. This is promising support for testing online, especially for studies in which small effect sizes are usually expected (e.g., semantic priming studies).

Another interesting finding was that trials which are not infant-led still replicated findings in-lab (Experiments 1 and 2), which generally require participants to attend to the screen before proceeding to the next trial.

A final, but important advantage to testing online was our finding of very little attrition or data loss (e.g., Experiment 1 <5%) compared to some lab-based studies, which can lose up to 30% according to a meta-analysis performed by Bergmann et al. (2018).

Having considered the advantages to testing online, we now turn to specific considerations when testing online. As with all new findings, more replication studies are required before generalising beyond these three paradigms that testing online is suitable for other infant paradigms and other infant populations.

Another point to consider is that children's attention might fade throughout the online session. Indeed, Tsuji et al. (2022) reported that it may be more difficult to maintain children's engagement and interest during online tasks than in the lab. They quoted Chuey et al. (2021) and Shields et al. (2021) who recommended keeping the tasks short and eliciting regular responses from children with synchronous tasks to monitor children's engagement. In our three experiments, we only found evidence for a reduction in attention during one experiment (Experiment 2) which differs, for example, from Experiment 1 in which very few trials were lost, suggesting that children remained engaged throughout this particular experiment. Future testing might explore if these behaviours were specific to the individual experiments.

Additionally, experimental findings in online testing might differ from clinical measures, such as being able to identify language delays (see Experiment 3). Indeed, experimental results obtained through online testing may not align with clinical evaluations, particularly in terms of detecting language delays in children. Standardised children's language tests, which are traditionally validated and considered clinically reliable when administered face-to-face, might not maintain the same level of validity or accuracy when conducted online (Frizelle et al., 2023). Children's language tests are established through meticulous procedures that involve face-to-face interactions between the child and the assessor. These tests are designed to assess various linguistic skills, including comprehension, vocabulary, grammar, and overall language development. During in-person assessments, clinicians can observe not only the child's responses but also their non-verbal cues, engagement, attention span, and other contextual factors that might influence their performance. This comprehensive approach helps in obtaining a holistic understanding of the child's language abilities and facilitates more accurate diagnoses or identification of language delays (see systematic review by Alfano et al., 2022).

Conducting these standardised tests in an online format introduces several potential challenges and limitations that can affect their validity. Online testing lacks the direct, in-person interaction between the experimenter and the child (Frizelle et al., 2023), which restricts the experimenter's ability to observe non-verbal cues, maintain engagement, or adjust the assessment based on the child's behaviour or reactions during the test. Furthermore, the adaptation of standardised tests to an online format might not have yet undergone validation processes as might their traditional face-to-face versions (Manning et al., 2020).

Another limitation is that certain types of paradigms might not be adaptable to an online format depending on the age. Indeed, Lapidow et al. (2021) showed age-related differences in the performance of young children (aged between 2 and 5 years) that are not apparent when conducting studies in person. Their study examined the same developmental task across three different methodologies: in-person, an online synchronous version, and an online asynchronous version. They investigated whether children's

inferences of unobserved populations are influenced by the variability of the observed samples. To examine this, children observed an experimenter randomly selecting balls from two identical containers (Lapidow et al., 2021). One container contained four balls of different colours (varied-sample), while the other container contained four balls of the same colour (uniform-sample). Subsequently, children were asked to determine which container was more likely to hold a ball of a different colour. In the in-person context which consisted of Experiment 1, 72.5% of children significantly chose the correct answer which was the *varied-sample* container and there were no age-related differences on the performance. In the online context, which included Experiments 2 and 3a-b, the researchers failed to replicate the in-person performance with the majority of children responding by chance. However, there was an effect of age with Experiment 2 which might have explained the results suggesting that children's online performance may become more robust with age. Considering that the age and overall population characteristics of the participants were identical in both settings, they conducted another online Experiment 3c with only children aged 4 years old. Similar to the in-person experiment, most of the children (76.2%) significantly chose the varied-sample. Additionally, in the online Experiments 2 and 3a-b, only children older than 3.5 years in the synchronous version and above 4 years in the asynchronous version performed above chance. The findings suggest that children's age significantly influences their performance in an online setting. Notably, older children performed better compared to younger children. These results differ from what would typically be observed in a lab or in-person setting.

An important consideration when testing online is that some platforms collecting eye movement data can involve offline coding of video data which is time-consuming (see C. M. Nelson & Oakes, 2021), though there are platforms, such as LabVanced, which can automatically code the looking behaviour (see Bánki et al., 2022). Despite this fact, performing offline coding on the video data can reduce data loss (see Venker et al., 2020) compared to the automatic calculations performed by in-lab eye-tracking software. Data loss from testing in a lab setting tends to occur when an eye-tracker loses connection, but manually assessing each frame when coding video data offline for online experiments does not present this issue.

Though our findings do not indicate that parents influenced the behaviour of their children during testing, the lack of control over the testing environment and how parents behave during testing should be considered. At a very minimum, clear instructions should be given (with instructional images or videos where possible) to the parents, indicating how they should behave, with an explanation of why this is important. However, further online testing might indicate if this type of instruction is necessary during at-home or in-lab testing, which could inform the instructions we give to parents during in-lab testing.

A final limitation to testing online is that while online testing has the potential to reach broader demographic groups in theory, in order for online testing to work well, basic requirements such as a suitable up-to-date device and a stable internet connection are linked to a financial situation and lifestyle that enable this access. Furthermore, using the same avenues for recruitment (i.e., an institutional database of families) does not extend our reach to test under-represented groups.

Another potential limitation is the use of non-standard exclusion criteria pertaining to less than 6 weeks of prematurity in our experiments, and the absence of information on participants' birth weight. Our participants' samples would have captured "late preterm" births (between 34 weeks up to 37), but not early preterm births. Late preterm births, constituting approximately 8% of all births (Loftin et al., 2010), have been associated with

cognitive outcomes generally poorer than those of full-term infants (see Martínez-Nadal & Bosch, 2021, for a review). However, contextual factors such as preschool attendance and experiencing sensitive parenting have been identified as potential moderators of these outcomes (Shah et al., 2023). This is a very different picture than for extremely preterm or very preterm infants who are at high risk of experiencing cognitive difficulties, including in language development (Foster-Cohen et al., 2007).

Therefore, it is possible that about 8% of our participants were moderately premature, and that out of those, a proportion would have had difficulty with language skills at the time they came to the Babylab. However, they would have been evenly distributed across the three experiments; in addition, what we observed across the three experiments is low attrition rate and robust effect sizes, which suggest that if anything, excluding these children would have led to even more robust effects.

These studies provide encouraging support that other infant paradigms might be suited for adaptation to online testing. For instance, paradigms for measuring children's knowledge of syntax can be applied to online testing such as the elicited production that investigates whether young children have abstract knowledge of a particular structure (e.g., Ambridge, 2011). Paradigms to assess socio-emotional regulation in infants can also be adjusted for online experiments – for example, the Face-to-Face Still-Face paradigm (e.g., Barbosa et al., 2020; Giusti et al., 2018) where the parent and infant engage face-to-face for 2 min (e.g., Play episode). Next, the parent is told to stop engaging and communicating with the child. Instead, they are instructed to maintain eye contact with the child while keeping a still face for 2 min (e.g., Still-Face episode). This paradigm could work online by video live recording the interaction between the caregiver and the infant via a video call application. Indeed, for example, a recent study by McElwain et al. (2022) validated an online procedure that assessed mother and infant behaviour during the Still Face Paradigm (SFP). They compared data collected during in-person lab visits with data collected during remote Zoom visits to establish the validity of the online procedure. For the online procedure, prior to the online session, mothers received an email giving information about the necessary equipment needed for the Zoom visit, such as a bouncy seat or high chair (McElwain et al., 2022). The online session was recorded and during the visit, the experimenter used the Zoom's screen sharing functionality to display slides containing comprehensive instructions for each task. Throughout all the activities, the experimenter disabled the video camera and microphone, with the exception of the Baseline video where he/she remained unmuted. He/she collaborated with the mother to find the most suitable video angle, ensuring that the faces of both the mother and infant were captured effectively (McElwain et al., 2022). When comparing virtual visits to laboratory visits, during the SFP, mothers and infants had similar vocalisations, gaze directions and proportions of facial expressions. Additionally, infants also displayed similar behavioural changes across SFP episodes.

In conclusion, this paper demonstrates that the recent pandemic has inadvertently opened promising avenues of investigation in early language studies, and it is likely that future research will harvest the benefits of the enforced development of online experiments, reaching out to multicultural and multilingual populations around the world.

Author contribution statement. Because the first two authors contributed equally to the manuscript in terms of data and writing-up, first authorship was decided on a coin toss. The last author supervised the studies and contributed to the writing-up.

Competing interest. The authors declare none.

References

- Abdi, H. (2007). The Bonferroni and Sidak corrections for multiple comparisons. In: Salkind NJ (ed) *Encyclopedia of measurement and statistics*. Sage, Thousand Oaks
- Alfano, A. R., Concepcion, I., Espinosa, A., & Menendez, F. (2022). Pediatric language assessments via telehealth: A systematic review. *Journal of Telemedicine and Telecare*, 1357633X221124998. Advance online publication. <https://doi.org/10.1177/1357633X221124998>
- Ambridge, B. (2011). Paradigms for assessing children's knowledge of syntax and morphology. In E. Hoff (Ed.), *Guide to research methods in child language* (pp. 113–132). Hoboken, NJ: Blackwell
- Anwyl-Irvine, A. L., Dalmaijer, E. S., Hodges, N., & Evershed, J. (2020a). Online Timing Accuracy and Precision: A comparison of platforms, browsers, and participant's devices. <https://doi.org/10.31234/osf.io/jfeca>
- Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020b). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, 52(1), 388–407. <https://doi.org/10.3758/s13428-019-01237-x>
- Archibald, M. M., Ambagtsheer, R. C., Casey, M. G., & Lawless, M. (2019). Using Zoom Videoconferencing for Qualitative Data Collection: Perceptions and Experiences of Researchers and Participants. *International Journal of Qualitative Methods*, 18. <https://doi.org/10.1177/1609406919874596>
- Arias-Trejo, N., & Plunkett, K. (2009). Lexical-semantic priming effects during infancy. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364(1536), 3633–3647. <https://doi.org/10.1098/rstb.2009.0146>
- Arias-Trejo, N., & Plunkett, K. (2013). What's in a link: associative and taxonomic priming effects in the infant lexicon. *Cognition*, 128(2), 214–227. <https://doi.org/10.1016/j.cognition.2013.03.008>
- Bacon, D., Weaver, H., & Saffran, J. (2021). A Framework for Online Experimenter-Moderated Looking-Time Studies Assessing Infants' Linguistic Knowledge. *Frontiers in psychology*, 12, 703839. <https://doi.org/10.3389/fpsyg.2021.703839>
- Baldwin, D. A. (1989). Priorities in children's expectations about object labels reference: Form over color. *Child Development*, 60, 1291–1306.
- Bánki, A., de Eccher, M., Falschlehner, L., Hoehl, S., & Markova, G. (2022). Comparing Online Webcam- and Laboratory-Based Eye-Tracking for the Assessment of Infants' Audio-Visual Synchrony Perception. *Frontiers in psychology*, 12, 733933. <https://doi.org/10.3389/fpsyg.2021.733933>
- Barbosa, M., Beeghly, M., Moreira, J., Tronick, E. Z., & Fuertes, M. (2020). Emerging patterns of infant regulatory behavior in the Still- Face paradigm at 3 and 9 months predict mother- infant attachment at 12 months. *Attachment & Human Development*, 1–17.
- Bergelson, E., & Swingle, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109(9), 3253–3258. <https://doi.org/10.1073/pnas.1113380109>
- Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting Replicability in Developmental Research Through Meta-analyses: Insights From Language Acquisition Research [https://doi.org/10.1111/cdev.13079]. *Child Development*, 89(6), 1996–2009. <https://doi.org/10.1111/cdev.13079>
- Blanchard, L. (2020). The effects of COVID-19 on virtual working within online groups. *Group processes and intergroup relations*, 24(2), 290–296, DOI: 10.1177/1368430220983446
- Bochynska, A., & Dillon, M. R. (2021). Bringing Home Baby Euclid: Testing Infants' Basic Shape Discrimination Online. *Frontiers in psychology*, 12, 734592. <https://doi.org/10.3389/fpsyg.2021.734592>
- Bulgarelli, F., & Bergelson, E. (2022). Talker variability shapes early word representations in English-learning 8-month-olds. *Infancy: the official journal of the International Society on Infant Studies*, 27(2), 341–368. <https://doi.org/10.1111/inf.12452>
- Byers-Heinlein, K., Bergmann, C., & Savalei, V. (2021). Six solutions for more reliable infant research. *Infant and Child Development*, e2296. <https://doi.org/10.1002/icd.2296>
- Cattani, A., Krott, A., Dennis, I., & Floccia, C. (2019). *WinG Words in Game Test*: A vocabulary assessment for pre-school children.
- Chere, B., & Kirkham, N. (2021). The Negative Impact of Noise on Adolescents' Executive Function: An Online Study in the Context of Home-Learning During a Pandemic [Original Research]. *Frontiers in Psychology*, 12. <https://doi.org/10.3389/fpsyg.2021.715301>

- Chow, J., Aimola Davies, A., & Plunkett, K. (2017). Spoken-word recognition in 2-year-olds: The tug of war between phonological and semantic activation. *Journal of Memory and Language*, **93**, 104–134. <https://doi.org/10.1016/j.jml.2016.08.004>
- Chuey, A., Asaba, M., Bridgers, S., Carrillo, B., Dietz, G., Garcia, T., Leonard, J. A., Liu, S., Merrick, M., Radwan, S., Stegall, J., Velez, N., Woo, B., Wu, Y., Zhou, X. J., Frank, M. C., & Gweon, H. (2021). Moderated Online Data-Collection for Developmental Research: Methods and Replications. *Frontiers in psychology*, **12**, 734398. <https://doi.org/10.3389/fpsyg.2021.734398>
- Cochet, H., Jover, M., & Vauclair, J. (2011). Hand preference for pointing gestures and bimanual manipulation around the vocabulary spurt period. *Journal of Experimental Child Psychology*, **110**(3), 393–407. <https://doi.org/10.1016/j.jecp.2011.04.009>
- Cohen, L. B. (1973). A two process model of infant visual attention. *Merrill Palmer Quarterly*, **19**(3), 157–180.
- Curtin, S., & Zamuner, T. S. (2014). Understanding the developing sound system: interactions between sounds and words. *Wiley interdisciplinary reviews. Cognitive science*, **5**(5), 589–602. <https://doi.org/10.1002/wcs.1307>
- Davies, C., Hendry, A., Gibson, S. P., Gliga, T., McGillion, M., & Gonzalez-Gomez, N. (2021). Early childhood education and care (ECEC) during COVID-19 boosts growth in language and executive function. *Infant and child development*, **30**(4), e2241.
- Davis-Kean, P. E., Tighe, L. A., & Waters, N. E. (2021). The Role of Parent Educational Attainment in Parenting and Children's Development. *Current Directions in Psychological Science*, **30**(2), 186–192. <https://doi.org/10.1177/0963721421993116>
- Delgado, T., Bhark, S. J., & Donahue, J. (2021). Pandemic Teaching: Creating and teaching cell biology labs online during COVID-19. *Biochemistry and molecular biology education: a bimonthly publication of the International Union of Biochemistry and Molecular Biology*, **49**(1), 32–37. <https://doi.org/10.1002/bmb.21482>
- Duncan, G. J., & Magnuson, K. A. (2003). Off with Hollingshead: Socioeconomic resources, parenting, and child development. In M. H. Bornstein & R. H. Bradley (Eds.), *Socioeconomic Status, Parenting, and Child Development* (pp. 83–106). Lawrence Erlbaum Associates Publishers.
- Fennell, C. T. (2012). Object Familiarity Enhances Infants' Use of Phonetic Detail in Novel Words. *Infancy*, **17**(3):339–353.
- Fennell, C. T., & Waxman, S. R. (2010). What paradox? Referential cues allow for infant use of phonetic detail in word learning. *Child development*, **81**(5), 1376–1383.
- Fennell, C. T., & Werker, J. F. (2003). Early word learners' ability to access phonetic detail in well-known words. *Language and speech*, **46**(Pt 2-3), 245–264. <https://doi.org/10.1177/00238309030460020901>
- Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2006). *MacArthur-Bates Communicative Development Inventories, Second Edition (CDIs)*. APA PsycTests.
- Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior & Development*, **8**(2), 181–195.
- Fernald, A., Pinto, J. P., Swingle, D., Weinberg, A., & McRoberts, G. W. (1998). Rapid gains in speed of verbal processing by infants in the 2nd year. *Psychological Science*, **9**(3), 228–231. <https://doi.org/10.1111/1467-9280.00044>
- Fernald, A., Swingle, D., & Pinto, J. P. (2001). When half a word is enough: Infants can recognize spoken words using partial phonetic information. *Child Development*, **72**(4), 1003–1015. <https://doi.org/10.1111/1467-8624.00331>
- Fitzpatrick, N., & Floccia, C. (2023). What Relationships Exist Between Words In The Lexical-Semantic Systems Of Toddlers? [Doctoral thesis, University of Plymouth] <https://pearl.plymouth.ac.uk/handle/10026.1/20623>
- Floccia, C., Delle Luche, C., Lepadatu, I., Chow, J., Ratnage, P., & Plunkett, K. (2020). Translation equivalent and cross-language semantic priming in bilingual toddlers. *Journal of Memory and Language*, **112**, 104086.
- Foster-Cohen, S., Edgin, J. O., Champion, P. R., & Woodward, L. J. (2007). Early delayed language development in very preterm infants: evidence from the MacArthur-Bates CDI. *Journal of child language*, **34**(3), 655–675.
- Frizelle, P., Buckley, A., Biancone, T., Ceroni, A., Dahly, D., Fletcher, P., Bishop, D. V. M., & McKean, C. (2023). How reliable is assessment of children's sentence comprehension using a self-directed app? A

- comparison of supported versus independent use. *Journal of Child Language*, 1–29. <https://doi.org/10.1017/S0305000923000545>
- Giusti, L., Provenzi, L., & Montiroso, R. (2018). The Face-to-Face Still-Face (FFSF) Paradigm in Clinical Settings: Socio-Emotional Regulation Assessment and Parental Support With Infants With Neurodevelopmental Disabilities. *Frontiers in Psychology*, 9, 789. <https://doi.org/10.3389/fpsyg.2018.00789>
- Golinkoff, R. M., Hirsh-Pasek, K., Cauley, K. M., & Gordon, L. (1987). The eyes have it: Lexical and syntactic comprehension in a new paradigm. *Journal of Child Language*, 14(1), 23–45. <https://doi.org/10.1017/S030500090001271X>
- Hamilton, A. F., Plunkett, K., & Schafer, G. (2000). Infant vocabulary development assessed with a British communicative development inventory. *Journal of Child Language*, 27.
- Jardak, A., & Byers-Heinlein, K. (2019). Labels or Concepts? The Development of Semantic Networks in Bilingual Two-Year-Olds. *Child Development*, 90(2), e212–e229. <https://doi.org/10.1111/cdev.13050>
- Lapidow, E., Tandon, T., Goddu, M., & Walker, C. M. (2021). A Tale of Three Platforms: Investigating Preschoolers' Second-Order Inferences Using In-Person, Zoom, and Lookit Methodologies. *Frontiers in psychology*, 12, 731404. <https://doi.org/10.3389/fpsyg.2021.731404>
- Loftin, R. W., Habli, M., Snyder, C. C., Cormier, C. M., Lewis, D. F., & DeFranco, E. A. (2010). Late preterm birth. *Reviews in obstetrics and gynecology*, 3(1), 10.
- Mäkinen, T., Laaksonen, M., Lahelma, E., & Rahkonen, O. (2006). Associations of childhood circumstances with physical and mental functioning in adulthood. *Soc Sci Med.* ;62(8):1831–9.
- Manning, B. L., Harpole, A., Harriott, E. M., Postolowicz, K., & Norton, E. S. (2020). Taking language samples home: Feasibility, reliability, and validity of child language samples conducted remotely with video chat versus in-person. *Journal of Speech, Language, and Hearing Research*, 63(12), 3982–3990.
- Martínez-Nadal, S., & Bosch, L. (2021). Cognitive and learning outcomes in late preterm infants at school age: A systematic review. *International Journal of Environmental Research and Public Health*, 18(1), 74.
- McElwain, N. L., Hu, Y., Li, X., Fisher, M. C., Baldwin, J. C., & Bodway, J. M. (2022). Zoom, Zoom, Baby! Assessing Mother-Infant Interaction During the Still Face Paradigm and Infant Language Development via a Virtual Visit Procedure. *Frontiers in psychology*, 12, 734492. <https://doi.org/10.3389/fpsyg.2021.734492>
- Mills, F., Bhogal, J. K., Dennis, A., Spoiala, C., Milward, J., Saeed, S., Jones, L. F., Weston, D., & Carter, H. (2022). The effects of messaging on long COVID expectations: An online experiment. *Health Psychology*, 41(11), 853–863. <https://doi.org/10.1037/hea0001230>
- Ministry of Housing, Communities, and Local Government. (2019). *English indices of deprivation 2019*. <https://imd-by-postcode.opendatacommunities.org/imd/2019>
- Mossakowski, K. N. (2008). Dissecting the Influence of Race, Ethnicity, and Socioeconomic Status on Mental Health in Young Adulthood. *Research on Aging*, 30(6), 649–671
- Nelson, C. M., & Oakes, L. M. (2021). “May I Grab Your Attention?": An Investigation Into Infants' Visual Preferences for Handled Objects Using Lookit as an Online Platform for Data Collection [Brief Research Report]. *Frontiers in Psychology*, 12. <https://doi.org/10.3389/fpsyg.2021.733218>
- Nelson, P. M., Scheiber, F., Laughlin, H. M., & Demir-Lira, Ö. E. (2021). Comparing Face-to-Face and Online Data Collection Methods in Preterm and Full-Term Children: An Exploratory Study. *Frontiers in Psychology*, 12, 733192.
- Nguyen, D. (2024). Is there a Relationship between Parents' Screen Usage and Young Children's Development? (Doctoral dissertation, University of Plymouth). <https://pearl.plymouth.ac.uk/handle/10026.1/22045>
- Oakes, L. M. (2017). Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy*, 22(4), 436–469. <https://doi.org/10.1111/inf.12186>
- R Core Team (2021). *R: A language and environment for statistical computing* (Version 1.4.1717) [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., Benitez, J., & Ocampo, J. D. (2020). Advancing Developmental Science via Unmoderated Remote Research with Children. *Journal of Cognition and Development: official journal of the Cognitive Development Society*, 21(4), 477–493. <https://doi.org/10.1080/15248372.2020.1797751>
- Ross-Sheehy, S., Reynolds, E., & Eschman, B. (2021). Unsupervised Online Assessment of Visual Working Memory in 4- to 10-Year-Old Children: Array Size Influences Capacity Estimates and Task

- Performance [Original Research]. *Frontiers in Psychology*, **12**. <https://doi.org/10.3389/fpsyg.2021.692228>
- Scott, K., & Schulz, L. (2017). Lookit (Part 1): A new online platform for developmental research. *Open Mind: Discoveries in Cognitive Science*. Advance Publication. https://doi.org/10.11.62/opmi_a_00002
- Shah, P. E., Poehlmann, J., Weeks, H. M., Spinelli, M., Richards, B., Suh, J., & Kaciroti, N. (2023). Developmental trajectories of late preterm infants and predictors of academic performance. *Pediatric Research*, 1–8.
- Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., Fei-Fei, L., Keil, F. C., Gweon, H., Tenenbaum, J. B., Jara-Ettinger, J., Adolph, K. E., Rhodes, M., Frank, M. C., Mehr, S. A., & Schulz, L. (2020). Online Developmental Science to Foster Innovation, Access, and Impact. *Trends in Cognitive Sciences*, **24**(9), 675–678. <https://doi.org/10.1016/j.tics.2020.06.004>
- Shields, M. M., McGinnis, M. N., & Selmecky, D. (2021). Remote Research Methods: Considerations for Work With Children. *Frontiers in Psychology*, **12**, 703706. <https://doi.org/10.3389/fpsyg.2021.703706>
- Singh, L. (2013). One world, two languages: cross-language semantic priming in bilingual toddlers. *Child Development*, **85**(2), 755–766. <https://doi.org/10.1111/cdev.12133>
- Singh, L., Tan, A., & Quinn, P. C. (2021). Infants recognize words spoken through opaque masks but not through clear masks. *Developmental Science*, **24**(6), e13117.
- Stager, C., & Werker, J. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature* **388**, 381–382.
- Styles, S. J., & Plunkett, K. (2009). How do infants build a semantic system? *Language and Cognition*, **1**(1), 1–24.
- Swingle, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, **76**(2), 147–166.
- Tsui, A. S. M., Byers-Heinlein, K., & Fennell, C. T. (2019). Associative word learning in infancy: A meta-analysis of the switch task. *Developmental Psychology*, **55**(5), 934.
- Tsuji, S., Amso, D., Cusack, R., Kirkham, N., & Oakes, L. M. (2022). Editorial: Empirical Research at a Distance: New Methods for Developmental Science. *Frontiers in psychology*, **13**, 938995. <https://doi.org/10.3389/fpsyg.2022.938995>
- UK-CDI Database. (2016, October 1). UK CDI. Retrieved from <https://www.uk-cdi.ac.uk/>
- Venker, C. E., Pomper, R., Mahr, T., Edwards, J., Saffran, J., & Ellis Weismer, S. (2020). Comparing Automatic Eye Tracking and Manual Gaze Coding Methods in Young Children with Autism Spectrum Disorder. *Autism Research*, **13**: 271–283. <https://doi.org/10.1002/aur.2225>
- Vihman, M. M., Thierry, G., Lum, J., Keren-Portnoy, T., & Martin, P. A. M. (2007). Onset of word form recognition in English, Welsh, and English–Welsh bilingual infants. *Applied Psycholinguistics*, **28**(3), 475–493. <https://doi.org/10.1017/S0142716407070269>.
- Waxman, S. R., & Booth, A. E. (2001). Seeing pink elephants: Fourteen-month-olds' interpretations of novel nouns and adjectives. *Cognitive Psychology*, **43**, 217–242.
- Werker, J. F., Cohen, L. B., Lloyd, V. L., Casasola, M., & Stager, C. L. (1998). Acquisition of word–object associations by 14-month-old infants. *Developmental Psychology*, **34**(6), 1289–1309. <https://doi.org/10.1037/0012-1649.34.6.1289>
- Werker, J. F., Fennell, C. T., Corcoran, K. M., & Stager, C. L. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary. *Infancy*, **3**:1–30
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Golemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V. ... Yutani, H. (2019). “Welcome to the tidyverse.” *Journal of Open Source Software*, **4**(43), 1686. doi:10.21105/joss.01686
- Yoshida, K. A., Fennell, C. T., Swingle, D., & Werker, J. F. (2009). Fourteen-month-old infants learn similar-sounding words. *Developmental Science*, **12**(3), 412–418
- Zoom Video Communications Inc. (2016). Security guide. Zoom Video Communications Inc. Retrieved from <https://d24cgw3%20uvb9a9h.cloudfront.net/static/81625/doc/Zoom-Security-WhitePaper.pdf>

Cite this article: Nguyen, D.K.-L., Fitzpatrick, N., & Floccia, C. (2025). Adapting language development research paradigms to online testing: Data from preferential looking, word learning and vocabulary assessment in toddlers. *Journal of Child Language* **52**, 465–497, <https://doi.org/10.1017/S0305000924000035>